GRAPH EMBEDDING METHODS FOR MULTIPLE-OMICS DATA ANALYSIS

EDITED BY: Chen Qingfeng, Wei Lan, Yi-Ping Phoebe Chen and Wilson Wen Bin Goh

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-600-5
DOI 10.3389/978-2-88971-600-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

GRAPH EMBEDDING METHODS FOR MULTIPLE-OMICS DATA ANALYSIS

Topic Editors:
Chen Qingfeng, Guangxi University, China
Wei Lan, Guangxi University, China
Yi-Ping Phoebe Chen, La Trobe University, Australia
Wilson Wen Bin Goh, Nanyang Technological University, Singapore

Citation: Qingfeng, C., Lan, W., Chen, Y.-P. P., Goh, W. W. B., eds. (2021). Graph Embedding Methods for Multiple-Omics Data Analysis. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-600-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 20 September 2021
doi: 10.3389/fgene.2021.762274

Editorial: Graph Embedding Methods for Multiple-Omics Data Analysis

Wei Lan[1], *Qingfeng Chen*[1]*, *Yi-Ping Phoebe Chen*[2] *and Wilson Wen Bin Goh*[3]

[1] *School of Computer Electronical and Information, Guangxi University, Nanning, China,* [2] *Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC, Australia,* [3] *School of Biological Sciences, Nanyang Technological University, Singapore, Singapore*

Keywords: multi-omics data analysis, graph embedding, computational methods, data integration, disease

Editorial on the Research Topic

Graph Embedding Methods for Multiple-Omics Data Analysis

With the advent of advanced high throughput biotechnologies such as next-generation sequencing and single-cell sequencing, there has been an increasing growth of complex multiple-omics data sets (such as genomics, epigenome, transcriptomics, proteomics, metabolomics, etc.). These data are often heterogeneous, sparse, high dimension and high noise, which provide different levels of insightful information for disease. Integrative analysis of multiple-omics data can help the biomedical researchers to explore biological mechanisms and further assist in designing better diagnostic tools and therapies for the treatment of diseases. However, the development of effective methods for multiple-omics data analysis is very challenging as the complex characteristics of different kinds of data. Recently, machine learning methods especially graph embedding have shown powerful capability in analyzing multiple-omics data. Particularly, they are capable to represent data as low dimensional vectors while the data features are preserved.

To provide a platform bridging graph embedding method and multiple-omics data analysis, we organized we organized a Research Topic on "*Graph Embedding Methods for Multiple-Omics Data Analysis*." This Research Topic presents 19 articles. We expect that these articles promote more advanced studies for multiple-omics data analysis.

Peng L. et al. identified potential antiviral drugs against SARS-CoV-2 by using regularized least squared classifier and bipartite local model. Ninety six virus-drug associations between 11 types of viruses similar to SARS-CoV-2 and 78 small molecular drugs were extracted in this study.

Hou et al. proposed a method to capture potential functions in a microbial co-occurrence network. It integrated topological structures of microbial co-occurrence networks with k-mer compositions of operational taxonomy unit sequences and embedded them into a lower-dimensional continuous latent space.

Pan et al. presented an embedding-based method for predicting the subcellular localization of proteins. The functional and network embeddings from GO terms and protein–protein network were combined as novel representations of protein locations for the construction of the final classification model.

Gu et al. proposed a method incorporating feature engineering and feature selection algorithms to explore the common controlling genes and corresponding pathways among eight different organs' fibrosis. These results were helpful for understanding the molecular mechanisms of fibrosis diseases and finding new therapeutic indications of existing drugs.

Zhang developed a feature selection algorithm for gene expression data classification by using approximate conditional entropy based on fuzzy information granule. The experimental results on large-scale gene datasets show that this algorithm not only greatly reduces the dimension of the gene datasets, but also is superior to the state-of-the-art algorithms in classification accuracy.

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

****Correspondence:***
Qingfeng Chen
qingfeng@gxu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *21 August 2021*
Accepted: *30 August 2021*
Published: *20 September 2021*

Citation:
Lan W, Chen Q, Chen Y-PP and Goh WWB (2021) Editorial: Graph Embedding Methods for Multiple-Omics Data Analysis. Front. Genet. 12:762274. doi: 10.3389/fgene.2021.762274

Yuan and Yang proposed a deep learning method to identify circRNA-RBP interactions by using hybrid double embeddings for representing RNA sequences and a cross-branch attention neural network for classification. The experimental results on benchmark datasets show that their method outperforms the mainstream deep learning-based methods on not only prediction accuracy but also computational efficiency.

He et al. adopted multiple kernel learning (MKL) to integrate somatic mutation to currently molecular data including gene expression, copy number variation (CNV), methylation, and protein expression data for the prediction of breast cancer survival. In addition, the maximum relevance minimum redundancy (mRMR) feature selection method was utilized to select features that present high relevance to survival and low redundancy among themselves for each type of data.

Su et al. designed a multi-level model to improve both the quality and speed of large-scale PPIs prediction. The results showed that their model is promising for large-scale PPI prediction in both accuracy and efficiency, which is beneficial to other large-scale biomedical molecules interactions detection.

Barbiero et al. tested the digital twin model on two simulated clinical case studies combining information at organ, tissue, and cellular level. The results show their approach is able to detect inflammatory cytokines which are known to have effects on blood pressure and have previously been associated with SARS-CoV-2 infection (e.g., CXCR6, XCL1, and others).

Zhang et al. developed a computational method based on the Light Gradient Boosting Machine (LightGBM) to predict potential metabolite–disease associations. It extracted the features from statistical measures, graph theoretical measures. Three case studies confirmed that this method has obvious superiority in predicting metabolite–disease pairs and represents a powerful bioinformatics tool.

Zhao et al. proposed a supervised gene selection method based on permutation and random forest classification. The experimental result on 10 datasets show that the gene selection performance of their method is better than other gene selection methods.

Wang J. et al. presented a computational drug repositioning approach to discover potential drug-disease associations. The experimental results demonstrate that their approach outperforms recent state-of-the-art prediction models. In addition, the case studies further confirm the predictive ability of the proposed method.

Wang, Dai, et al. introduced a pan-cancer classification method to identify a set of genes that can differentiate all tumor types accurately. Extensive experimental results on the public RNA-seq data sets with 33 different tumor types show that this method outperforms the other state-of-the-art classification methods.

Wang, Cao, et al. proposed a computational method to predict and identify the m6A sites on mRNA by utilizing sequence-derived and graph embedding features. The comparison results show that the proposed method achieved the best performance compared with other predictors on four public datasets across three species.

Wang Z. et al. explored the gene expression changes and its potential effects mediated by U11 snRNA in bladder cancer cell. This study show that U11 may be involved in the regulation of gene expression in bladder cancer cells, which may provide a potentially new biomarker for clinical diagnosis and treatment of bladder cancer.

Feng et al. integrated transcriptomic, lipidomic, and metabolomic analyses to identify the differential lipids and metabolites between basal and luminal muscle invasive bladder cancer (MIBC) subtypes. The results suggest that free fatty acids (FFA) and sulfatides (SL), which are closely associated with immune and stromal cell types, have strong capacities to distinguish basal and luminal subtypes of MIBC tumors. Moreover, the results also show that the ratios of glycerophosphocholine (GCP)/imidazoles and nucleosides/imidazoles can accurately identify tumors of basal and luminal MIBC subtypes.

Peng X. et al. presented a method to construct methylation haplotypes for homologous chromosomes in CpG dense regions. The proposed method not only can be applied to methylation analysis, but also can provide a clear explanation for the methylation difference at the resolution of methylation haplotypes.

Liu and Zhang developed a computational model for the detection of copy number variation detection (CNV) of different lengths from whole genome sequencing data. It used a clustering algorithm to divide the read depth segment profile, and assigned an abnormal score to each read depth segment. The experimental results show that the performance of proposed model is better than those of several existing methods.

Zheng and Wu proposed a method for predicting drug-target interactions based on heterogeneous network integration and cascade deep forest. The results show that their model outperforms the previously reported methods on the benchmark datasets.

AUTHOR CONTRIBUTIONS

The article was written by WL and QC. Y-PC and WG have provided guidance to the manuscript preparation and have also reviewed and edited the paper. All authors have approved the final version of the editorial.

FUNDING

This work was partially supported by the National Natural Science Foundation of China (Nos. 62072124, 61963004, and 61972185), the Natural Science Foundation of Guangxi (Nos. 2021GXNSFAA075041 and 2018GXNSFBA281193), the Science and Technology Base and Talent Special Project of Guangxi (No. AD20159044).

ACKNOWLEDGMENTS

The Topic Editors thank all the authors and reviewers of the articles submitted to this Frontiers Research Topic.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 September 2020
doi: 10.3389/fgene.2020.577387

Identifying Effective Antiviral Drugs Against SARS-CoV-2 by Drug Repositioning Through Virus-Drug Association Prediction

Lihong Peng[1]*†, Xiongfei Tian[1]†, Ling Shen[1], Ming Kuang[1], Tianbao Li[2], Geng Tian[2], Jialiang Yang[2]* and Liqian Zhou[1]*

[1] School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2] Geneis (Beijing) Co., Ltd., Beijing, China

A new coronavirus called SARS-CoV-2 is rapidly spreading around the world. Over 16,558,289 infected cases with 656,093 deaths have been reported by July 29th, 2020, and it is urgent to identify effective antiviral treatment. In this study, potential antiviral drugs against SARS-CoV-2 were identified by drug repositioning through Virus-Drug Association (VDA) prediction. 96 VDAs between 11 types of viruses similar to SARS-CoV-2 and 78 small molecular drugs were extracted and a novel VDA identification model (VDA-RLSBN) was developed to find potential VDAs related to SARS-CoV-2. The model integrated the complete genome sequences of the viruses, the chemical structures of drugs, a regularized least squared classifier (RLS), a bipartite local model, and the neighbor association information. Compared with five state-of-the-art association prediction methods, VDA-RLSBN obtained the best AUC of 0.9085 and AUPR of 0.6630. Ribavirin was predicted to be the best small molecular drug, with a higher molecular binding energy of −6.39 kcal/mol with human angiotensin-converting enzyme 2 (ACE2), followed by remdesivir (−7.4 kcal/mol), mycophenolic acid (−5.35 kcal/mol), and chloroquine (−6.29 kcal/mol). Ribavirin, remdesivir, and chloroquine have been under clinical trials or supported by recent works. In addition, for the first time, our results suggested several antiviral drugs, such as FK506, with molecular binding energies of −11.06 and −10.1 kcal/mol with ACE2 and the spike protein, respectively, could be potentially used to prevent SARS-CoV-2 and remains to further validation. Drug repositioning through virus–drug association prediction can effectively find potential antiviral drugs against SARS-CoV-2.

Keywords: SARS-CoV-2, antiviral drugs, drug repositioning, virus-drug association, regularized least square, bipartite local model, neighbor association information

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Qi Zhao,
Liaoning University, China
Weiyang Chen,
Qilu University of Technology, China

****Correspondence:***
Lihong Peng
plhhnu@163.com
Jialiang Yang
yangjl@geneis.cn
Liqian Zhou
zhoulq11@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *29 June 2020*
Accepted: *18 August 2020*
Published: *16 September 2020*

Citation:
Peng L, Tian X, Shen L, Kuang M, Li T, Tian G, Yang J and Zhou L (2020) Identifying Effective Antiviral Drugs Against SARS-CoV-2 by Drug Repositioning Through Virus-Drug Association Prediction. Front. Genet. 11:577387. doi: 10.3389/fgene.2020.577387

INTRODUCTION

Last December 2019, a novel coronavirus called SARS-CoV-2 by the World Health Organization (WHO), first found in Wuhan, China, was rapidly spreading around the world (Kaiser et al., 2020; Sanche et al., 2020). The SARS-CoV-2 outbreak was declared as a global public health emergency by WHO, and a total of 16,558,289 cases have been confirmed with another 656,093 deaths throughout the world by July 29th, 2020 (World Health Organization [WHO], 2020). SARS-CoV-2 caused a severe acute respiratory syndrome named COVID-19, and no special vaccine or antiviral drug against SARS-CoV-2 has been found at present (Lu, 2020; Wang et al., 2020c). Therefore, finding a special antiviral drug as soon as possible is urgent to stop the spread of SARS-CoV-2 (Lu, 2020; Zhang et al., 2020a).

However, designing a new drug to treat COVID-19 in a short time is almost impossible (Zhang et al., 2020a). One of the best strategies is drug repositioning (Chen et al., 2012, 2016; Peng et al., 2017a; Beck et al., 2020). By repositioning already commercialized drugs, the undesired effects can be inferred to find new uses for these drugs. This strategy can thus greatly shorten the time required for an antiviral drug against SARS-CoV-2.

Although little is known about SARS-CoV-2, its complete genome sequence is strongly homologous to SARS-CoV (Huang et al., 2020; Morse et al., 2020). Therefore, in this study, to prioritize available FDA-approved antiviral drugs against SARS-COV-2 for further clinical trials, 11 well-studied viruses similar to SARS-CoV-2 were selected and 96 virus–drug associations (VDAs) with these 11 viruses were integrated. Regularized least squared classifier (RLS), bipartite local model (BLM), and neighbor association information were applied in our new algorithm named VDA-RLSBN to find novel VDAs for new virus (especially for SARS-CoV-2) or new drug. The results showed that ribavirin, remdesivir, and chloroquine may be antiviral drugs against SARS-CoV-2.

Molecular docking techniques investigate the behavior of small molecular drugs in the binding site of a target protein. As more target protein structures are confirmed experimentally, molecular docking approaches are widely applied to drug design (Zhang et al., 2020b). AutoDock (Goodsell et al., 1996; Ruyck et al., 2016) is an available software applied to identify the bound conformations of a small molecular drug to a macromolecular target. The AutoDock affinity scoring function is applied to rank the candidate poses based on the sum of the van der Waals and electrostatic energies. We conducted molecular docking between the predicted top 10 antiviral drugs against SARS-CoV-2 and two target proteins including the spike protein of SARS-CoV-2 and human angiotensin-converting enzyme 2 (ACE2) molecule (Wang et al., 2020a). The molecular binding energies between the above three drugs and ACE2 are ribavirin with −6.39 kcal/mol, remdesivir with −7.4 kcal/mol, and chloroquine with −6.29 kcal/mol. These three small molecules have been under clinical trial or supported by recent publications. In addition, we found that FK506 shows higher molecular binding energies of −10.1 kcal/mol and −11.06 kcal/mol with these two targets, which suggest that FK506 may be applied to stop COVID-19 although there is no report about its association with SARS-CoV-2.

MATERIALS AND METHODS

Dataset

Aiming at identifying potential VDAs related to SARS-CoV-2, 96 known VDAs between 11 viruses similar to SARS-CoV-2 and 78 small molecular drugs were selected from the DrugBank (Wishart et al., 2018), NCBI (Sayers et al., 2020), and PubMed (Canese and Sarah, 2013) databases. The element y_{ij}^{ori} in the VDA matrix $Y^{ori} \in \Re^{\mathrm{n}\times\mathrm{m}}$ was represented as

$$y_{ij}^{ori} = \begin{cases} 1 \text{ if the } i\text{th virus associates with the } j\text{th drug} \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{1}$$

These similar viruses included SARS-CoV (Ding et al., 2004), MERS-CoV (Groot et al., 2013), human immunodeficiency virus type 1 (Wei et al., 1995) and type 2 (Guyader et al., 1987) (HIV-1 and HIV-2), chronic hepatitis C virus (HCV) (Jacobson et al., 2011), influenza A viruses [A-H1N1 (Kumar et al., 2009), A-H5N1 (Subbarao et al., 1998), A-H7N9 (Gao et al., 2013)], Hendra virus (Bonaparte et al., 2005), human cytomegalovirus (Cobbs et al., 2002), and respiratory syncytial virus (Hall, 2001). Complete genome sequences of these 11 viruses and SARS-CoV-2 were downloaded from the NCBI database, and virus similarity matrix $S_v \in \Re^{\mathrm{n}\times\mathrm{n}}$ was computed based on MAFFT, a multiple-sequence alignment software. Chemical structures of drugs were downloaded from the DrugBank database, and drug similarity matrix $S_d \in \Re^{\mathrm{m}\times\mathrm{m}}$ was obtained by RDKit, an open-source cheminformatics tool. The details are shown in **Table 1**.

Methods

Problem Formalization

Bleakley and Yamanishi (2009) represented a drug–target interaction network as a bipartite graph and developed a BLM-based method to predict possible drug–target interactions. The proposed method first inferred targets of a given FDA-approved drug and drugs targeting a known protein and then combined these two independent predictions. The results demonstrated the excellent performance of BLM. Similar to the drug–target interaction network, the VDA network can also be taken as a bipartite graph. Results in this study are thus presented to evaluate the prediction performance in each of the following four cases for a given putative virus–drug pair:

- The virus with at least one known drug and the drug with at least one known virus.
- The virus with at least one known drug and the drug without any known virus (new drug).
- The virus without any known drug (new virus) and the drug with at least one known virus.
- New virus and new drug.

Based on these four cases, we represent a VDA network as a bipartite graph and thus the predicted VDA matrix $Y_{n\times m}^{\mathrm{pre}}$ can be

TABLE 1 | Statistics of viruses and drugs.

Virus	No. of drugs	Virus	No. of drugs
SARS-CoV	15	Hendra virus	1
MERS-CoV	9	HIV-1	35
A-H1N1	4	HIV-2	3
A-H5N9	2	HCV	15
A-H7N9	4	Respiratory syncytial virus	2
Human cytomegalovirus	6	SARS-CoV-2	0

denoted as Eq. (2):

$$Y_{n\times m}^{\text{pre}} = \begin{bmatrix} (Y_1)_{n_{cv}\times m_{cv}} & (Y_2)_{n_{cv}\times \bar{m}} \\ (Y_3)_{\bar{n}\times m_{cv}} & (Y_4)_{\bar{n}\times \bar{m}} \end{bmatrix} \tag{2}$$

where $\bar{n} = n - n_{cv}$ is the number of new viruses (for example, SARS-CoV-2), and $\bar{m} = m - m_{cv}$ is the number of new drugs. Y_1 represents VDAs from n_{cv} existing viruses and m_{cv} existing drugs, Y_2 represents VDAs from n_{cv} existing viruses and $\bar{m}$ new drugs, Y_3 denotes VDAs from $\bar{n}$ new viruses and m_{cv} existing drugs, and Y_4 denotes VDAs from $\bar{n}$ new viruses and $\bar{m}$ new drugs. Our aims are to identify potential VDAs in the subnetwork Y_1 as well as in Y_2, Y_3, and Y_4. **Figure 1** shows the flowchart of VDA-RLSBN.

Regularized Least Square

To infer possible VDA candidates, we develop an RLS-based VDA identification model (VDA-RLS) to compute the association profile $\hat{y}$ for each virus–drug pair:

$$\hat{y} = K(K + \sigma I)^{-1} y \tag{3}$$

where K represents the kernel matrix, y denotes the original association profile, and σ is a regularization parameter.

To compute VDA matrix Y_1 from n_{cv} existing viruses and m_{cv} existing drugs, we consider the ensemble of independent virus-based prediction and drug-based prediction with RLS. The solution of Y_1 can be thus divided down into the following four steps:

Step 1 For a given virus v_i with at least one known association, its new association profile $\hat{y}_i^v$ can be computed from its original association profile y_i^v and the kernel matrix K_v based on RLS classifier:

$$\hat{y}_i^v = K_v(K_v + \sigma I_v)^{-1} y_i^v \tag{4}$$

where $K_v = (S_v + S_v^T)/2$, and y_i^v represents the ith row of Y^{ori}. We can compute virus-based VDA matrix Y^v by Eq. (4).

Step 2 For a given drug d_j with at least one known association, its new association profile $\hat{y}_j^d$ can be computed from its original association profile y_j^d and the kernel matrix K_d based on RLS classifier:

$$\hat{y}_j^d = K_d(K_d + \sigma I_d)^{-1} y_j^d \tag{5}$$

where $K_d = (S_d + S_d^T)/2$, and y_j^d represents the jth column of Y^{ori}. We can compute drug-based VDA matrix Y^d by Eq. (5).

Step 3 Integrate Y^v with the element y_{ij}^v and Y^d with the element y_{ij}^d to compute the predicted VDA matrix Y^{RLS} based on RLS:

$$y_{ij}^{\text{RLS}} = \max\left(y_{ij}^v,\ y_{ij}^d\right) \tag{6}$$

Step 4 Obtain Y_1 by Eq. (7):

$$Y_1 = Y^{ori} + Y^{RLS} \tag{7}$$

Regularized Least Square With Neighbor Association Information

We can identify novel VDAs between existing viruses and existing drugs, or known/new viruses and new/existing drugs based on RLS and BLM. However, VDA-RLS was not able to predict associations between new viruses and new drugs. To solve this problem, we developed a VDA prediction model (VDA-RLSBN) by integrating neighbor association information into the RLS model.

Based on the "guilt-by-association" method, similar viruses/drugs tend to associate with similar drugs/viruses, so the association profile of an unknown virus could be possibly found by its neighbors' association information. Viruses highly similar to a new virus can be considered as its neighbors. Since the new virus has no associated drugs (i.e., its current association profile is a vector with all the elements of 0), complete genome sequence similarity of viruses is applied to define its neighbors.

For a new virus v_i, its association weight with a drug d_j can be computed by its neighbors' associations with d_j and its association profile $a_i^v(j)$ is defined as Eq. (8):

$$a_i^v(j) = \sum_{k=1}^{n_{cv}} S_{ik}^v y_{kj}^{ori} \tag{8}$$

where S_{ik}^v is the complete genome sequence similarity between two viruses v_i and v_j. $a_i^v(j) > 0$ when the jth associated drug d_j exists, i.e., $y_{kj}^{\text{ori}} > 0$ for at least one k and $a_i^v(j) = 0$ when the jth associated drug d_j is new, i.e., $y_{kj}^{\text{ori}} = 0$ for all k. $a_i^v(j)$ is normalized to make its value in the range of [0, 1] by Eq. (9):

$$a_i^v(j) = \frac{\left(a_i^v(j) - \min_k a_i^v(k)\right)}{\left(\max_k a_i^v(k) - \min_k a_i^v(k)\right)} \tag{9}$$

Also, an independent virus-based association profile y_i^v for a virus–drug pair can be represented as Eq. (10):

$$\hat{y}_i^v = K_v(K_v + \sigma I_v)^{-1} a_i^v \tag{10}$$

Similarly, for a new drug d_j, its association profile y_j^d for the same virus–drug pair can be represented as Eq. (10):

$$\hat{y}_j^d = K_d(K_d + \sigma I_d)^{-1} a_j^d \tag{11}$$

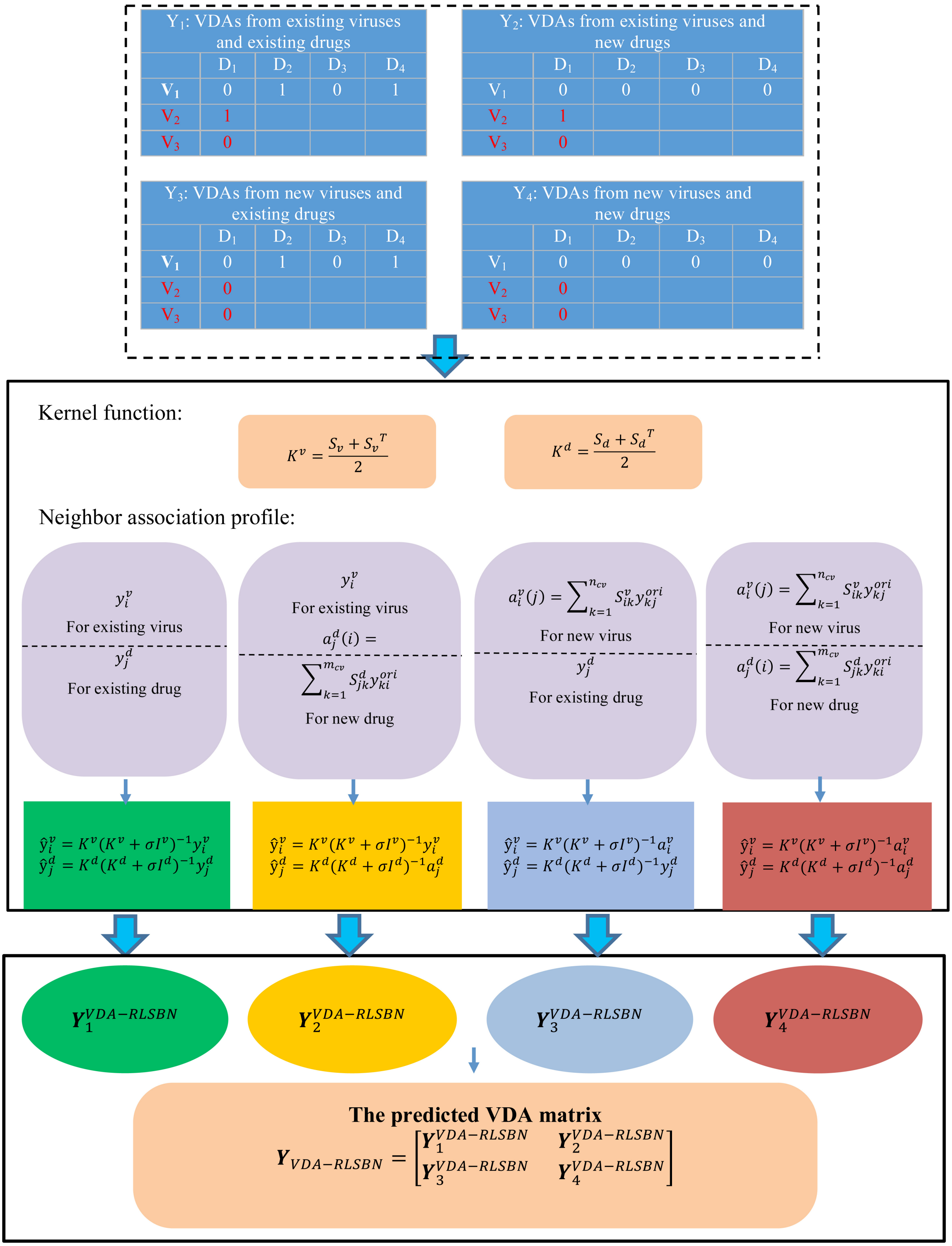
where a_j^d denotes the neighbor association profile of d_j .

The final VDA network can be represented as

$$Y_{\text{VDA-RLSBN}} = \begin{bmatrix} Y_1^{\text{VDA-RLSBN}} & Y_2^{\text{VDA-RLSBN}} \\ Y_3^{\text{VDA-RLSBN}} & Y_4^{\text{VDA-RLSBN}} \end{bmatrix} \tag{12}$$

where $Y_1^{\text{VDA-RLSBN}}$ can be computed by Eqs (4–7); $Y_2^{\text{VDA-RLSBN}}$ can be computed by Eqs (4), (11), and (6); $Y_3^{\text{VDA-RLSBN}}$ can be obtained by Eqs (10), (5), and (6); and $Y_4^{\text{VDA-RLSBN}}$ can be obtained by Eqs (10), (11), and (6). Specially, the VDA matrix related to SARS-CoV-2 can be obtained from $Y_3^{\text{VDA- RLSBN}}$.

Finally, we used AutoDock to analyze the druggability of the predicted top 10 chemical agents and their binding activities with two target proteins including the SARS-CoV-2 spike protein and ACE2.



Y_1: VDAs from existing viruses and existing drugs	D_1	D_2	D_3	D_4
V_1	0	1	0	1
V_2	1			
V_3	0			

Y_2: VDAs from existing viruses and new drugs	D_1	D_2	D_3	D_4
V_1	0	0	0	0
V_2	1			
V_3	0			

Y_3: VDAs from new viruses and existing drugs	D_1	D_2	D_3	D_4
V_1	0	1	0	1
V_2	0			
V_3	0			

Y_4: VDAs from new viruses and new drugs	D_1	D_2	D_3	D_4
V_1	0	0	0	0
V_2	0			
V_3	0			

Kernel function:
$K^v = \frac{S_v + S_v{}^T}{2}$
$K^d = \frac{S_d + S_d{}^T}{2}$
Neighbor association profile:
y_i^v
For existing virus
y_j^d
For existing drug
y_i^v
For existing virus
$a_j^d(i) = \sum_{k=1}^{m_{cv}} S_{jk}^d y_{ki}^{ori}$
For new drug
$a_i^v(j) = \sum_{k=1}^{n_{cv}} S_{ik}^v y_{kj}^{ori}$
For new virus
y_j^d
For existing drug
$a_i^v(j) = \sum_{k=1}^{n_{cv}} S_{ik}^v y_{kj}^{ori}$
For new virus
$a_j^d(i) = \sum_{k=1}^{m_{cv}} S_{jk}^d y_{ki}^{ori}$
For new drug
$\hat{y}_i^v = K^v(K^v + \sigma I^v)^{-1} y_i^v$ $\hat{y}_j^d = K^d(K^d + \sigma I^d)^{-1} y_j^d$
$\hat{y}_i^v = K^v(K^v + \sigma I^v)^{-1} y_i^v$ $\hat{y}_j^d = K^d(K^d + \sigma I^d)^{-1} a_j^d$
$\hat{y}_i^v = K^v(K^v + \sigma I^v)^{-1} a_i^v$ $\hat{y}_j^d = K^d(K^d + \sigma I^d)^{-1} y_j^d$
$\hat{y}_i^v = K^v(K^v + \sigma I^v)^{-1} a_i^v$ $\hat{y}_j^d = K^d(K^d + \sigma I^d)^{-1} a_j^d$
$\boldsymbol{Y}_1^{VDA-RLSBN}$
$\boldsymbol{Y}_2^{VDA-RLSBN}$
$\boldsymbol{Y}_3^{VDA-RLSBN}$
$\boldsymbol{Y}_4^{VDA-RLSBN}$
The predicted VDA matrix
$\boldsymbol{Y}_{VDA-RLSBN} = \begin{bmatrix} \boldsymbol{Y}_1^{VDA-RLSBN} & \boldsymbol{Y}_2^{VDA-RLSBN} \\ \boldsymbol{Y}_3^{VDA-RLSBN} & \boldsymbol{Y}_4^{VDA-RLSBN} \end{bmatrix}$

FIGURE 1 | Flowchart of VDA-RLSBN.

RESULTS

Evaluation Metrics and Experimental Settings

In this section, we performed extensive experiments to evaluate our proposed VDA-RLSBN method. We compared VDA-RLSBN with five state-of-the-art machine learning-based models, including LRLSHMDA (Wang et al., 2017), SMiR-NBI (Li et al., 2016), CMF (Zheng et al., 2013), NetLapRLS (Xia et al., 2010), and WNN-GIP (Laarhoven and Marchiori, 2013). The experiments were performed on a MAC with 2.4 GHz Inter Core i5, 8 GB 2133 MHz LPDDR3 of the RAM and OS Catalina 10.15.4 operating system.

Sensitivity, specificity, accuracy, AUC, and AUPR are widely applied to evaluate various machine learning-based models. In this study, we used these five metrics to measure the performance of five state-of-the-art models and VDA-RLSBN. Accuracy denotes the ratio of correctly predicted VDAs to all VDAs. Sensitivity denotes the ratio of correctly predicted positive VDAs to all positive VDAs. Specificity is the ratio of correctly predicted negative VDAs to all negative VDAs. AUC is the area under the ROC curve. The ROC curve can be plotted by a true positive rate [TPR, i.e., Eq. (13)] and a false-positive rate [FPR, i.e., Eq. (14)].

$$\mathrm{TPR} = TP/\left(TP + FN\right) \tag{13}$$

$$\mathrm{FPR} = FP/(FP + TN) \tag{14}$$

where TPR represents the ratio of correctly predicted positive VDAs to all positive VDAs and FPR represents the ratio of mistakenly predicted positive VDAs to all negative VDAs.

AUPR is the area under the PR curve. The PR curve can be plotted by precision and recall. Precision represents the ratio of correctly predicted positive VDAs to all predicted positive VDAs, and recall represents the ratio of correctly predicted positive VDAs to all positive VDAs.

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{16}$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. Generally, larger AUC/AUPR value denotes better performances.

We used five-fold cross validation to train our proposed VDA-RLSBN method. In each round, 80% of VDAs in the known VDA network was used as a training set and the remaining 20% of VDAs was the test set. The experiments were performed 100 times, and the final performance was on average over 100 times. In each round, a virus/drug is new if all of its associated drugs/viruses are selected as a test set.

For the parameters in five comparative methods and VDA-RLSBN, we conducted grid search to determine their optimal values. In VDA-RLSBN, we set the parameter σ in the range of $[0,\ 0.1,\ 0.2,\ \ldots,\ 1]$ and found that VDA-RLSBN obtained the best performance when σ is set as 0.4. In LRLSHMDA, we set the parameter lw in the range of $[0,\ 0.1,\ 0.2,\ \ldots,\ 1]$ and found that LRLSHMDA obtained better accuracy when lw is set as 0.1. In CMF, we set the parameters λ_l, λ_d, and λ_t in the range of $[2^{-2},\ \ldots,\ 2^{1}]$, $[2^{-3},\ \ldots,\ 2^{5}]$, and $[2^{-3},\ \ldots,\ 2^{5}]$, respectively. We found that CMF obtained better performance when $\lambda_l = 1$, $\lambda_d = 0.25$, and $\lambda_t = 0.125$. In NetLapRLS, we set four parameters γ_d, γ_t, β_d, and β_t in the range of $[1e^{-6}, \ldots, 1e^{2}]$ and found that NetLapRLS performed better when these four parameters were set as $1e - 6$. In WNN-GIP, we set five parameters T, α_d, α_t, σ, and γ in the range of $[0,\ 0.1,\ \ldots,\ 1.0]$ and found that WNN-GIP obtained the optimal performance when $T = 0.7$, $\alpha_d = 0.6$, $\alpha_t = 0.6$, $\sigma = 1$, and $\gamma = 0.5$. All parameters in these six models were set as the corresponding values where the corresponding method obtained the optimal performance.

Comparison With Five State-of-the-Art Methods

The performance of our proposed VDA-RLSBN and these five machine learning-based models is shown in **Table 2**. The best performance in each row is shown in bold in **Table 2**. LRLSHMDA (Wang et al., 2017), NetLapRLS (Xia et al., 2010), WNN-GIP (Laarhoven and Marchiori, 2013), and VDA-RLSBN are RLS-based methods. LRLSHMDA (Wang et al., 2017) used Laplacian RLS to tackle microbe–disease association prediction, NetLapRLS (Xia et al., 2010) extended the standard Laplacian RLS incorporating drug–target network, and WNN-GIP (Laarhoven and Marchiori, 2013) integrated a simple weighted nearest neighbor method and Gaussian kernels into RLS. SMiR-NBI (Li et al., 2016) constructed a heterogeneous network connecting genes, drugs, and miRNAs and then combined a network-based inference algorithm to characterize the responses of anticancer drugs. CMF (Zheng et al., 2013) was a collaborative matrix factorization-based drug–target interaction prediction method.

The results showed that VDA-RLSBN outperformed LRLSHMDA, SMiR-NBI, CMF, and WNN-GIP in terms of five evaluation metrics. Although the specificity value of VDA-RLSBN is slightly lower compared to NetLapRLS, its AUC and AUPR are significantly higher than NetLapRLS. Since AUC and AUPR are more important evaluation metrics compared to other three measurements, VDA-RLSBN, with the highest AUC and AUPR, is considered to be better in finding potential VDAs of novel viruses.

TABLE 2 | The performance of VDA-RLSBN with other five methods.

Methods	Accuracy	Sensitivity	Specificity	AUC	AUPR
LRLSHMDA	0.5841	0.6702	0.5823	0.8303	0.1778
SMiR-NBI	0.2080	0.8437	0.1935	0.5721	0.4912
CMF	0.8980	0.8971	0.9916	0.7500	0.4210
NetLapRLS	0.8974	0.8974	**0.9992**	0.6758	0.1777
WNN-GIP	0.8786	0.8961	0.9072	0.8491	0.5356
VDA-RLSBN	**0.9298**	**0.9279**	0.9841	**0.9085**	**0.6630**

Among six VDA prediction methods, LRLSHMDA, NetLapRLS, WNN-GIP, and VDA-RLSBN are RLS-based methods. VDA-RLSBN obtained better performance than the other three methods. Although other RLS-based prediction methods have good performance, they cannot predict the relationship between new drug candidates and new candidate targets. If a virus/drug has no known drug/virus, it is a new virus/drug. Since there are many new viruses/drugs, our proposed VDA-RLSBN approach learned labeled information from neighbors and used the information to train the model and make predictions. So VDA-RLSBN obtained better performance compared to other RLS-based methods. The results suggest that RLS combining neighbor association information can better identify new VDAs.

Case Study

The prediction performance of the proposed VDA-RLSBN method was confirmed in the last section. As a means to finding potential antiviral drugs against SARS-CoV-2, small molecular drugs were ranked based on the association scores with SARS-CoV-2 and the top 10 drugs with the highest scores were listed in **Table 3**. Among the predicted top 10 VDAs, 4 VDAs are reported by related literature, that is, 40% small molecular drugs are confirmed to be possible antiviral drugs against SARS-CoV-2.

Ribavirin is inferred to be the best small molecular drug against SARS-CoV-2. It is a broad-spectrum antiviral drug that can inhibit the replication of respiratory syncytial virus (Laarhoven and Marchiori, 2013). For example, it has been applied to prevent respiratory syncytial virus infection in lung transplant recipients (Hayden and Whitley, 2020) and specially used to treat SARS-CoV and MERS-CoV (Permpalung et al., 2019). Similar to SARS-CoV and MERS-CoV, SARS-CoV-2 is a respiratory syndrome betacoronavirus and may cause serious respiratory diseases. A few studies (Li and De, 2020; Wang et al., 2020b) have reported that ribavirin may take an inhibitory effect on SARS-CoV-2. More importantly, remdesivir and chloroquine are inferred to be other effective antiviral drugs. Wang et al. (2020b) presented that remdesivir and chloroquine can effectively inhibit SARS-CoV-2 and they have been used in the clinical stage. These results suggest that ribavirin, remdesivir, and chloroquine may be applied to the treatment of COVID-19.

TABLE 3 | The predicted top 10 drugs associated with SARS-CoV-2.

Rank	Drug	Confirmed
1	Ribavirin	doi: 10.1038/d41573-020-00016-0 PMID:32034637
2	Remdesivir	PMID:32036774, 32035533, 32035018, 31971553, 32022370, 31996494, 32020029 doi: 10.1101/2020.01.28.922922 doi: 10.26434/chemrxiv.11831101.v1
3	Mycophenolic acid	Unconfirmed
4	Chloroquine	PMID:32020029
5	Phenothiazine	Unconfirmed
6	Mizoribine	doi: 10.20944/preprints202002.0061.v1
7	FK506	Unconfirmed
8	Pentoxifylline	Unconfirmed
9	6-Azauridine	Unconfirmed
10	Protein phosphatase 1	Unconfirmed

Molecular Docking

We conducted molecular docking between the predicted top 10 small molecules and the SARS-CoV-2 spike protein/ACE2. The chemical structures of these small molecular drugs were downloaded from the DrugBank database. The structure of the virus spike protein was obtained based on homologous modeling from Zhang Lab (2020). The structure of ACE2 can be downloaded from the RCSB Protein Data Bank (Helen et al., 2000) (ID:6MJ0). AutoDock used the genetic algorithm as a search algorithm and selected the entire protein as a grid box.

The molecular binding energies between the predicted top 10 small molecules and these two target proteins are described in **Table 4**. The results show that the predicted top 10 drugs have higher molecular binding activities with the spike protein and/or ACE2. For example, ribavirin, which is predicted to be the most possible drug against SARS-CoV-2, has a higher molecular binding energy of −6.39 kcal/mol with ACE2. In addition, remdesivir, mycophenolic acid, and chloroquine are predicted to have higher association scores with SARS-CoV-2. These three small molecular drugs showed higher binding energies of −7.4, −5.35, and −6.29 kcal/mol with ACE2, respectively. More importantly, ribavirin, remdesivir, and chloroquine have been used for the treatment of SARS, which has about 79% sequence identity with SARS-CoV-2. So the potential use of these three small molecules as a treatment for COVID-19 may be under investigation. Interestingly, FK506 is an immunesuppressive drug and mainly used to decrease the activity of the immune system after organ transplant. The molecular docking results show that

TABLE 4 | The molecular binding energies between the predicted top 10 antiviral drugs and two target proteins.

Target protein	Drug	Binding energy
The spike protein	Ribavirin	−5.29
	Remdesivir	−5.22
	Mycophenolic acid	−3.6
	Chloroquine	−5.03
	Phenothiazine	−5.44
	Mizoribine	−6.07
	FK506	−10.1
	Pentoxifylline	−8.59
	6-Azauridine	−7.72
	Protein phosphatase 1	−8.46
ACE2	Ribavirin	−6.39
	Remdesivir	−7.4
	Mycophenolic acid	−5.35
	Chloroquine	−6.29
	Phenothiazine	−8.12
	Mizoribine	−7.62
	FK506	−11.06
	Pentoxifylline	−5.98
	6-Azauridine	−10.74
	Protein phosphatase 1	−9.13

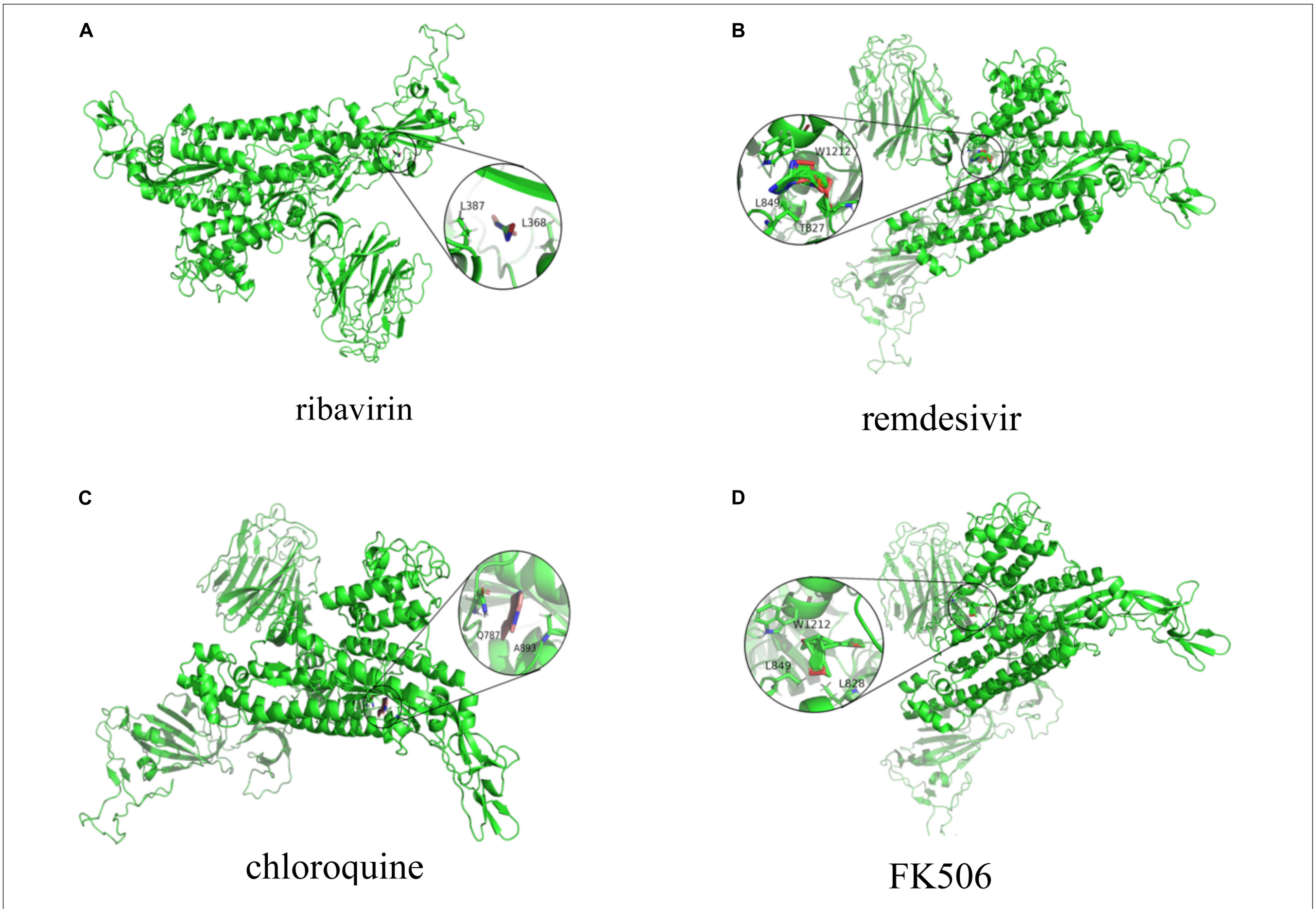


FIGURE 2 | Molecular docking between **(A)** ribavirin, **(B)** remdesivir, **(C)** chloroquine, and **(D)** FK506 and the spike protein.

FK506 has a strong molecular binding energy of −11.06 and −10.1 kcal/mol with ACE2 and the spike protein, respectively, although it has a slightly lower rank in the predicted drugs against SARS-CoV-2 by VDA-RLSBN.

Figures 2, **3** represent the docking results about four small molecules including ribavirin, remdesivir, chloroquine, and FK506 and two target proteins. The subfigure in each circle denotes the residues at the binding site of the SARS-CoV-2 spike protein/ACE2 and their corresponding orientations. For example, the amino acids L387, L368, P565, and V209 were inferred to be the key residues for ribavirin binding to the SARS-CoV-2 spike protein/ACE2 while L828, L849, W1212, N163, and N194 were inferred as the key residues for FK506 binding to the SARS-CoV-2 spike protein/ACE2.

DISCUSSION

With the spreading of SARS-CoV-2 around the world, the incidence rate is rapidly increasing, and lack of effective treatment options made it a public health threat. Therefore, various strategies are being exploited. Drug repositioning, aiming to offer a potentially valuable opportunity to find new clues of treatment for existing FDA-approved drugs, provides a far more rapid option to the clinic than novel drug design.

In the proposed VDA-RLSBN method, we predicted VDA candidates based on RLS and BLM. However, SARS-CoV-2 is a new coronavirus and has no associated drugs verified by biomedical experiments. We cannot find potential VDAs related to the virus by RLS and BLM. Therefore, we used association information of other RNA viruses similar to SARS-CoV-2 and similarities between SARS-CoV-2 and these viruses. The originality of our proposed method remains, predicting possible antiviral drugs against SARS-CoV-2 by drug repositioning through virus–drug association identification. More importantly, we integrated neighbor association information to RLS to find associated chemical agents for the new virus. The experimental results showed the merits of the VDA-RLSBN model. Higher AUC and AUPR indicated that the predicted antiviral drugs against SARS-CoV-2 are likely to be effective for preventing the rapid transmission of COVID-19.

VDA-RLSBN can obtain superior performance regardless of AUC, AUPR, accuracy, or sensitivity. This observation may be attributed to the following two features. First, VDA-RLSBN divides new VDA prediction into four cases based on BLM, a state-of-the-art method applied in various association prediction

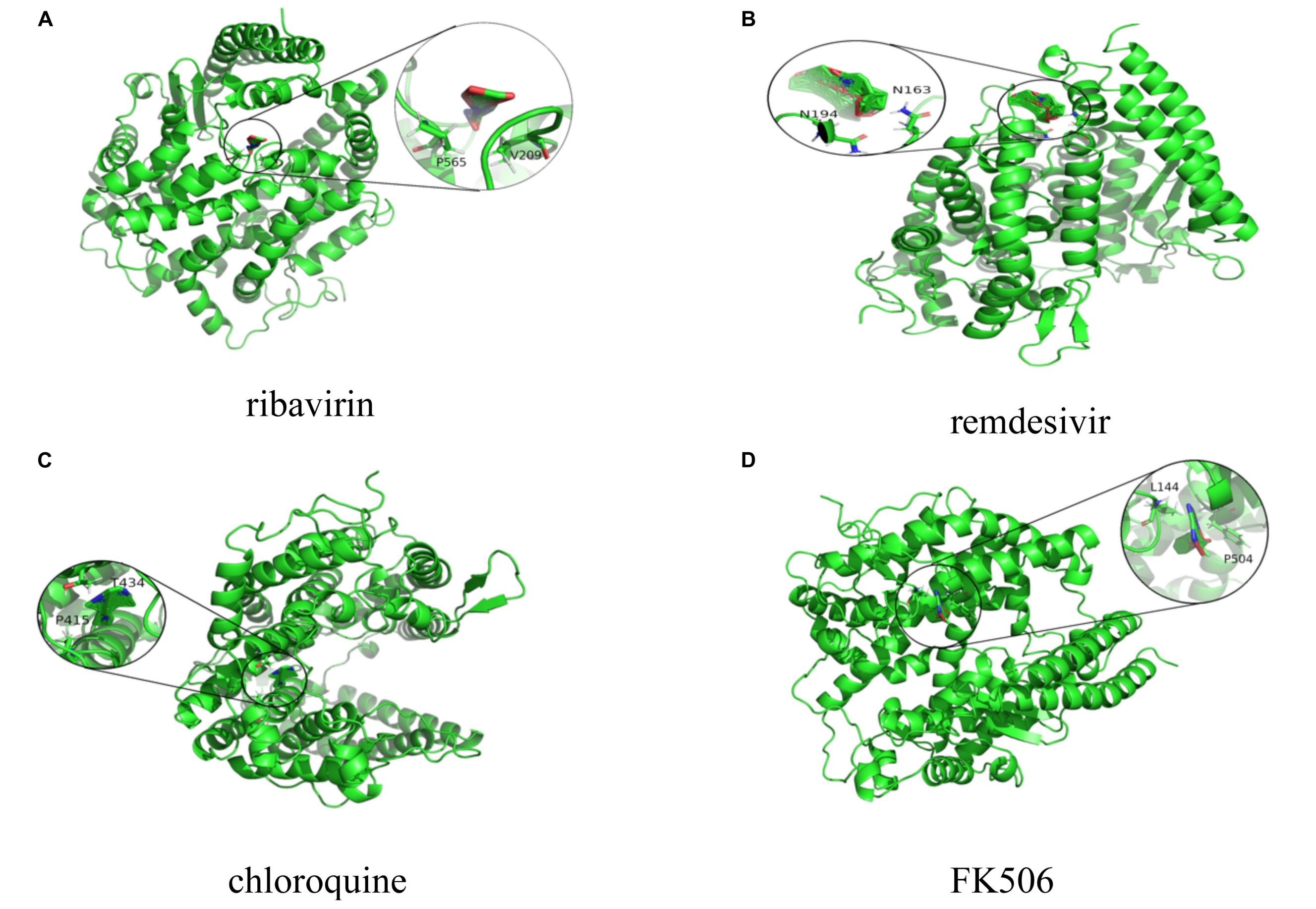


FIGURE 3 | Molecular docking between **(A)** ribavirin, **(B)** remdesivir, **(C)** chloroquine, and **(D)** FK506 and ACE2.

areas. More importantly, neighbor association information can help to identify possible antiviral drugs against new viruses (for example, SARS-CoV-2).

The proposed VDA-RLSBN approach is also helpful in designing and interpreting pharmacological experiments. The method can be further applied to select potential antiviral drugs against other new viruses, for example, infectious bronchitis virus.

CONCLUSION

In this study, we considered the clues of treatment from SARS-CoV, MERS-CoV, and other diseases caused by single-strand RNA viruses and developed a VDA prediction method based on RLS, BLM, and neighbor association information. VDA-RLSBN inferred commercially available small molecular drugs that could be applied to experimental therapy options against SARS-CoV-2. We conducted molecular docking between the predicted four chemical compounds including ribavirin, remdesivir, chloroquine, and FK506 and two target proteins including the spike protein and ACE2. The results show that ribavirin, remdesivir, and chloroquine have better molecular binding activities with ACE2 and may be the best small molecular drugs against SARS-CoV-2. In addition, we found that several antiviral drugs, such as FK506, could be used to combat COVID-19. Nevertheless, the 4 predicted drugs ranked 1, 2, 4, and 6 have been supported by recent works. We hope that our predicted small molecules may be helpful in the prevention of the transmission of SARS-CoV-2.

In the future, we will develop ensemble frameworks (Hu et al., 2018; Peng et al., 2020) and positive-unlabeled learning methods (Lan et al., 2016a; Peng et al., 2017b) to further improve the prediction performance. More importantly, we will enlarge the existing dataset. We will also integrate various biological data including long noncoding RNA (Lan et al., 2017; Zhao et al., 2018; Liu et al., 2020) and disease symptom information (Lan et al., 2016b).

CODE AVAILABILITY

Source code is freely downloadable at: https://github.com/plhhnu/VDA-RLSBN/.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

LP and XT contributed equally to this work. LP, XT, JY, and LZ designed the VDA-RLSBN method. XT and MK ran VDA-RLSBN. XT wrote the original manuscript. LP, TL, and JY revised the original draft. LS conducted molecular docking for the predicted results. LP, GT, JY, and LZ discussed the proposed method and gave further research. All authors read and approved the final manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant No. 61803151) and Natural Science Foundation of Hunan Province (Grant Nos. 2018JJ3570 and 2018JJ2461).

ACKNOWLEDGMENTS

We are thankful for help from Guangyi Liu and Longjie Liao from Hunan University of Technology; Ruyi Dong, Lebin Liang, Qinqin Lu, and Jidong Lang from Geneis (Beijing) Co., Ltd.; and Junlin Xu from Hunan University. We would like to thank all authors of the cited references.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.577387/full#supplementary-material

REFERENCES

Beck, B. R., Shin, B., Choi, Y., Park, S., and Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* 18, 784–790. doi: 10.1016/j.csbj.2020.03.025

Bleakley, K., and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 25, 2397–2403. doi: 10.1093/bioinformatics/btp433

Bonaparte, M. I., Dimitrov, A. S., Bossart, K. N., Crameri, G., Mungall, B. A., Bishop, K. A., et al. (2005). Ephrin-B2 ligand is a functional receptor for Hendra virus and Nipah virus. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10652–10657. doi: 10.1073/pnas.0504887102

Canese, K., and Sarah, W. (2013). "PubMed: the bibliographic database," *The NCBI Handbook*, 2nd edition. United States: National Center for Biotechnology Information.

Chen, X., Liu, M., and Yan, G. (2012). Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* 8, 1970–1978. doi: 10.1039/c2mb00002d

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2016). Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform.* 17, 696–712. doi: 10.1093/bib/bbv066

Cobbs, C. S., Harkins, L., Samanta, M., Gillespie, G. Y., Bharara, S., King, P. H., et al. (2002). Human cytomegalovirus infection and expression in human malignant glioma. *Cancer Res.* 62, 3347–3350.

Ding, Y., He, L., Zhang, Q., Huang, Z., Che, X., Hou, J., et al. (2004). Organ distribution of severe acute respiratory syndrome (SARS) associated coronavirus (SARS-CoV) in SARS patients: implications for pathogenesis and virus transmission pathways. *J. Pathol.* 203, 622–630. doi: 10.1002/path.1560

Gao, H. N., Lu, H., Cao, B., Du, B., Shang, H., Gan, J., et al. (2013). Clinical findings in 111 cases of influenza A (H7N9) virus infection. *N. Engl. J. Med.* 368, 2277–2285.

Goodsell, D. S., Morris, G. M., and Olson, A. J. (1996). Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* 9, 1–5. doi: 10.1002/(sici)1099-1352(199601)9:1<1::aid-jmr241>3.0.co;2-6

Groot, R. J. D., Baker, S. C., Baric, R. S., Brown, C. S., Drosten, C., Enjuanes, L., et al. (2013). Commentary: Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. *J. Virol.* 87, 7790–7792.

Guyader, M., Emerman, M., Sonigo, P., Clavel, F., Montagnier, L., and Alizon, M. (1987). Genome organization and transactivation of the human immunodeficiency virus type 2. *Nature* 326, 662–669. doi: 10.1038/326662a0

Hall, C. B. (2001). Respiratory syncytial virus and parainfluenza virus. *N. Engl. J. Med.* 344, 1917–1928. doi: 10.1056/nejm200106213442507

Hayden, F. G., and Whitley, R. J. (2020). Respiratory syncytial virus anti-virals: problems and progress. *J. Infect. Dis.* 2020:jiaa029. doi: 10.1093/infdis/jiaa029

Helen, B. M., John, W., Feng, Z., Gary, G., Bhat, T. N., Helge, W., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-Ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan China. *Lancet* 395, 497–506.

Jacobson, I. M., McHutchison, J. G., Dusheiko, G., Di Bisceglie, A. M. D., Reddy, K. R., Bzowej, N. H., et al. (2011). Telaprevir for previously untreated chronic hepatitis C virus infection. *N. Engl. J. Med.* 364, 2405–2416.

Kaiser, U. B., Mirmira, R. G., and Stewart, P. M. (2020). Our response to COVID-19 as endocrinologists and diabetologists. *J. Clin. Endocrinol. Metab.* 105:dgaa148.

Kumar, A., Zarychanski, R., Pinto, R., Cook, D., Marshall, J., and Lacroix, J. (2009). Critically ill patients with 2009 influenza A (H1N1) infection in Canada. *JAMA* 302, 1872–1879. doi: 10.1001/jama.2009.1496

Laarhoven, T. V., and Marchiori, E. (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* 8:e66952. doi: 10.1371/journal.pone.0066952

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F., Pan, Y., et al. (2017). LDAP: a web server for lncRNA-disease association prediction. *Bioinformatics* 33, 458–460.

Lan, W., Wang, J., Li, M., Liu, J., Li, Y., Wu, F. X., et al. (2016a). Predicting drug–target interaction using positive-unlabeled learning. *Neurocomputing* 206, 50–57. doi: 10.1016/j.neucom.2016.03.080

Lan, W., Wang, J., Li, M., Liu, J., Wu, F., and Pan, Y. (2016b). Predicting MicroRNA-disease associations based on improved MicroRNA and disease similarities. *IEEE/ACM Trans. Computat. Biol. Bioinform.* 15, 1774–1782. doi: 10.1109/tcbb.2016.2586190

Li, G., and De, C. E. (2020). Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nat. Rev. Drug Discov.* 19, 149–150. doi: 10.1038/d41573-020-00016-0

Li, J., Lei, K., Wu, Z., Li, W., Liu, G., Liu, J., et al. (2016). Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* 7, 45584–45596. doi: 10.18632/oncotarget.10052

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., and Zhao, Q. (2020). Predicting lncRNA-miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowl. Based Syst.* 191:105261. doi: 10.1016/j.knosys.2019.105261

Lu, H. (2020). Drug treatment options for the 2019-new coronavirus (2019-nCoV). *Biosci. Trends* 14, 69–71. doi: 10.5582/bst.2020.01020

Morse, J., Lalonde, T., Xu, S., and Liu, W. (2020). Learning from the past: possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV. *Chembiochem* 21, 730–738. doi: 10.1002/cbic.202000047

Peng, L., Liao, B., Zhu, W., Li, Z., and Li, K. (2017a). Predicting drug–target interactions with multi-information fusion. *IEEE J. Biomed. Health Inform.* 21, 561–572. doi: 10.1109/jbhi.2015.2513200

Peng, L., Zhu, W., Liao, B., Duan, Y., Chen, M., Chen, Y., et al. (2017b). Screening drug-target interactions with positive-unlabeled learning. *Sci. Rep.* 7, 1–17.

Peng, L., Zhou, L., Chen, X., and Piao, X. (2020). A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. *Front. Bioeng. Biotechnol.* 8:40. doi: 10.3389/fbioe.2020.00040

Permpalung, N., Thaniyavarn, T., Saullo, J. L., Arif, S., Miller, R. A., Reynolds, J. M., et al. (2019). Oral and inhaled ribavirin treatment for respiratory syncytial virus infection in lung transplant recipients. *Transplantation* 104, 1280–1286. doi: 10.1097/TP.0000000000002985

Ruyck, J. D., Brysbaert, G., Blossey, R., and Lensink, M. F. (2016). Molecular docking as a popular tool in drug design, an in silico travel. *Adv. Appl. Bioinform. Chem.* 9, 1–11. doi: 10.2147/aabc.s105289

Sanche, S., Lin, Y., Xu, C., Severson, E. R., Hengartner, N. W., and Ke, R. (2020). The novel coronavirus, 2019-nCoV, is highly contagious and more infectious than initially estimated. *medRxiv* [Preprint], doi: 10.1101/2020.02.07.20021154

Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., et al. (2020). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 48, D9–D16.

Subbarao, K., Klimov, A., Katz, J., Regnery, H., Lim, W., Hall, H., et al. (1998). Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* 279, 393–396. doi: 10.1126/science.279.5349.393

Wang, C., Wang, S., Li, D., Zhao, X., Han, S., Wang, T., et al. (2020a). Lectin-like intestinal defensin inhibits 2019-nCoV Spike binding to ACE2. *bioRxiv* [Preprint], doi: 10.1101/2020.03.29.013490

Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., et al. (2020b). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* 30, 269–271. doi: 10.1038/s41422-020-0282-0

Wang, W., Tang, J., and Wei, F. (2020c). Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. *J. Med. Virol.* 92, 441–447. doi: 10.1002/jmv.25689

Wang, F., Huang, Z., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017). LRLSHMDA: laplacian regularized least squares for human microbe–disease association prediction. *Sci. Rep.* 7:7601.

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., et al. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 373, 117–122.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082.

World Health Organization [WHO] (2020). *Coronavirus Disease 2019 (COVID-19): Situation Report-190.* Available online at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200729-covid-19-sitrep-191.pdf?sfvrsn=2c327e9e_2 (June 29th, 2020 (accessed June 30, 2020).

Xia, Z., Wu, L. Y., Zhou, X., and Wong, S. T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* 4(Suppl. 2):S6. doi: 10.1186/1752-0509-4-S2-S6

Zhang, H., Saravanan, K. M., Yang, Y., Hossain, M. T., Li, J., Ren, X., et al. (2020a). Deep learning based drug screening for novel coronavirus 2019-nCoV. *Interdiscip. Sci.* 1, 1–9.

Zhang, Z., Li, X., Zhang, W., Shi, Z., Zheng, Z., and Wang, T. (2020b). Clinical features and treatment of 2019-nCoV pneumonia patients in Wuhan: report of a couple cases. *Virol. Sin.* 35, 330–336. doi: 10.1007/s12250-020-00203-208

Zhang Lab (2020). Available online at: https://zhanglab.ccmb.med.umich.edu/C-I-TASSER/2019-nCov (accessed May 29, 2020).

Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., and Liu, H. (2018). The bipartite network projection recommended algorithm for predicting long noncoding RNA–protein interactions. *Mol. Ther. Nucleic Acid* 13, 464–471. doi: 10.1016/j.omtn.2018.09.020

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Kyoto.

Conflict of Interest: GT, TL, and JY were employed by the company Geneis (Beijing) Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 January 2021
doi: 10.3389/fgene.2020.627396

Identification of Common Genes and Pathways in Eight Fibrosis Diseases

Chang Gu[1,2†], Xin Shi[3†], Xuening Dang[4,5†], Jiafei Chen[2], Chunji Chen[1], Yumei Chen[6], Xufeng Pan[1*] and Tao Huang[7*]*

[1] Department of Thoracic Surgery, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China, [2] Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China, [3] Department of Cardiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China, [4] Department of Colorectal and Anal Surgery, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [5] Shanghai Colorectal Cancer Research Center, Shanghai, China, [6] Department of Nuclear Medicine, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [7] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

Acute and chronic inflammation often leads to fibrosis, which is also the common and final pathological outcome of chronic inflammatory diseases. To explore the common genes and pathogenic pathways among different fibrotic diseases, we collected all the reported genes of the eight fibrotic diseases: eye fibrosis, heart fibrosis, hepatic fibrosis, intestinal fibrosis, lung fibrosis, pancreas fibrosis, renal fibrosis, and skin fibrosis. We calculated the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment scores of all fibrotic disease genes. Each gene was encoded using KEGG and GO enrichment scores, which reflected how much a gene can affect this function. For each fibrotic disease, by comparing the KEGG and GO enrichment scores between reported disease genes and other genes using the Monte Carlo feature selection (MCFS) method, the key KEGG and GO features were identified. We compared the gene overlaps among eight fibrotic diseases and connective tissue growth factor (CTGF) was finally identified as the common key molecule. The key KEGG and GO features of the eight fibrotic diseases were all screened by MCFS method. Moreover, we interestingly found overlaps of pathways between renal fibrosis and skin fibrosis, such as GO:1901890-positive regulation of cell junction assembly, as well as common regulatory genes, such as CTGF, which is the key molecule regulating fibrogenesis. We hope to offer a new insight into the cellular and molecular mechanisms underlying fibrosis and therefore help leading to the development of new drugs, which specifically delay or even improve the symptoms of fibrosis.

Keywords: fibrotic diseases, genes, pathways, Monte Carlo feature selection, CTGF

OPEN ACCESS

Edited by:
Chen Qingfeng,
Guangxi University, China

Reviewed by:
Xuechao Wan,
Northwestern University, Philippines
Yungang Sun,
Nanjing Chest Hospital, China

****Correspondence:***
Tao Huang
tohuangtao@126.com
Xufeng Pan
panxfshch@163.com
Yumei Chen
chym0215@hotmail.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 November 2020*
Accepted: *15 December 2020*
Published: *15 January 2021*

Citation:
Gu C, Shi X, Dang X, Chen J, Chen C, Chen Y, Pan X and Huang T (2021) Identification of Common Genes and Pathways in Eight Fibrosis Diseases. Front. Genet. 11:627396. doi: 10.3389/fgene.2020.627396

INTRODUCTION

Acute and chronic inflammation often leads to fibrosis, which is also the common and final pathological outcome of chronic inflammatory diseases (Rockey et al., 2015). Fibrosis is defined as overaccumulation of fibrous connective tissue in and around the tissues with inflammation or damage, triggering irreversible scar formation. The clinical manifestations are renal disease, idiopathic pulmonary fibrosis (IPF), heart failure, end-stage liver diseases, and so on (Bataller and Brenner, 2005). Besides, fibrosis can also be observed in many chronic autoimmune diseases, such as rheumatoid arthritis, scleroderma, myelofibrosis, and Crohn disease. But the common characteristics of these fibrosis diseases were still unknown.

Fibrosis can affect chronic graft rejection, tumor invasion and metastasis, and the pathogenesis of many progressive myopathies. With regard to chronic graft rejection, fibrosis is one of the most common symptoms in chronic graft rejection. For example, liver transplantation in children has a 20-year survival of more than 80% at present, but the long-term results of these grafts still remain uncertain. Biopsies after liver transplantation show idiopathic post-transplant hepatitis and graft fibrosis occur even in children with good graft function (Kelly et al., 2016). As for tumor invasion and metastasis, carcinoma-associated fibroblasts are able to enhance tumor cells migration and invasion via activating the process of specific pathways. For example, as lung cancer maintains the leading cause of cancer-related deaths, IPF has been demonstrated that it increases the risk of lung cancer development by 7–20%, and there are multiple common molecular processes that associated IPF with lung cancer, such as epithelial–mesenchymal transition (EMT), endoplasmic reticulum stress, and abnormal expression of growth factors (Gu et al., 2018, 2020a,b; Ballester et al., 2019; Jiao and Yang, 2020). In the tissue of myopathies, there is prominent endomysial fibrosis, but little or no inflammation.

The fact that fibrotic changes are commonly observed in different diseases of diverse organ systems suggests common pathogenic pathways (Rockey et al., 2015). The wound healing in the fibrotic tissue is regulated by complex processes within different cells, and therefore some specific molecular pathways are activated. For example, in IPF, the fibrosis starts from the lung periphery to the lung center, finally causing respiratory failure. The underlying mechanisms of IPF were proven that elevated mechanical tension activates a transforming growth factor β (TGF-β) signaling loop in alveolar stem cells (AT2).

In this study, we proposed a new computational method incorporating feature engineering and feature selection algorithms to explore the common controlling genes and corresponding pathways among eight different organs' fibrosis. The key genes and pathways were revealed, and the cross-talks between diseases were investigated. These results were helpful for understanding the molecular mechanisms of fibrosis diseases and finding new therapeutic indications of existing drugs, i.e., drug repositioning.

MATERIALS AND METHODS

The Reported Genes of the Eight Fibrotic Diseases

All the genes of the related eight fibrotic diseases (eye fibrosis, heart fibrosis, hepatic fibrosis, intestinal fibrosis, lung fibrosis, pancreas fibrosis, renal fibrosis, and skin fibrosis) extracted from published researches are listed in **Supplementary Table 1**. In **Supplementary Table 1**, "1" refers to the genes associated with the specific fibrotic diseases, whereas "0" means the genes have no relationship with the specific fibrotic diseases. We compared the reported genes of the eight fibrotic diseases using R package SuperExactTest,[1] which has the function of identification of sets of objects with shared features, which is a common operation in all disciplines. Analysis of intersections among multiple sets is fundamental for in-depth understanding of their complex relationships. This package implements a theoretical framework for efficient computation of statistical distributions of multiset intersections based on combinatorial theory and provides multiple scalable techniques for visualizing the intersection statistics (Wang et al., 2015). There were 954 genes that were associated with at least one of the eight fibrotic diseases. In each fibrotic disease, the numbers of reported genes are listed in **Table 1**.

Encoding the Fibrotic Disease Genes With KEGG and GO Features

We calculated the KEGG (Kyoto Encyclopedia of Genes and Genomes) and GO (Gene Ontology) enrichment scores of all fibrotic disease genes. For each specific fibrotic disease, the reported genes of this disease were considered as positive samples, and the other genes were considered as negative samples. The KEGG and GO enrichment scores (Shi et al., 2018; Gu et al., 2020c) were used as features to encode genes and characterize their functions.

The KEGG and GO enrichment scores were the functional profiles of a gene. To be more specific, we enriched the neighbors of genes in STRING network (version 11.0[2]) (Szklarczyk et al., 2018) on to KEGG pathway and GO terms. Given a gene g, let $S(g)$ be a gene set consisting of genes that have functional associations with gene g in STRING network (Szklarczyk et al., 2018). Given a gene g and a GO term GO_j, the GO enrichment score was defined as the $-\log_{10}$ of the hypergeometric test P-value (Chen et al., 2016) of the gene set $S(g)$ and the GO term GO_j, which can be computed as follows:

$$S_{GO}\left(l, GO_j\right) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}\right) \tag{1}$$

where N was the total number of human genes in STRING database, M and n were the number of genes annotated to GO_j and the number of genes in $S(g)$, respectively, and m was the number of genes in $S(g)$ that were annotated to GO_j .

Similarly, the KEGG enrichment scores can be calculated by replacing the GO terms with KEGG pathways. The higher enrichment score meant this gene can affect this biological function. In total, there were 22,130 features (324 KEGG enrichment scores and 21,806 GO enrichment scores). The GO (2019-Apr24) annotations were downloaded from ftp://ftp.geneontology.org/, and the KEGG (Release 91.0) annotations

[1] https://CRAN.R-project.org/package=SuperExactTest

[2] https://string-db.org/

TABLE 1 | The number of reported genes in the eight fibrotic diseases.

Index	Disease	No. of reported genes
1	Eye fibrosis	39
2	Heart fibrosis	207
3	Hepatic fibrosis	173
4	Intestinal fibrosis	150
5	Lung fibrosis	185
6	Pancreas fibrosis	43
7	Renal fibrosis	160
8	Skin fibrosis	125

were extracted from https://www.kegg.jp/ using R/Bioconductor package KEGGREST[3] on July 1, 2019.

Identifying the Key KEGG and GO Features for Each Fibrotic Disease

The Monte Carlo feature selection (MCFS) method (Draminski et al., 2008) was applied to rank all the KEGG and GO features based on their importance in classification. It has been widely used and showed great power in identify robust key features for complex biological problems (Pan et al., 2018, 2020; Chen et al., 2020; Li et al., 2020a; Ren et al., 2020). As a supervised feature selection method, the MCFS method was based on tree classifiers. It constructed a series of tree classifiers on a series of subsets randomly selected from the whole dataset. By considering how much a feature contributed in these tree classifiers, the importance of this feature was calculated. By comparing with its importance calculated on permuted datasets, its significance can be calculated. As it ensembled a series of trees, the results were robust and trustworthy (Pan et al., 2019a,b,c, 2020; Li et al., 2020b).

For each fibrotic disease, the KEGG and GO enrichment features of the positive samples (the reported genes of this disease) and the negative samples (the other genes) were compared, and the relative importance (RI) of each feature was evaluated using MCFS algorithm. The significant KEGG and GO features were selected and analyzed. Software dmlab downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html was used to apply the MCFS algorithm, and the default parameters were used.

RESULTS

The Overlapped Genes of the Eight Fibrotic Diseases

We compared the reported genes of the eight fibrotic diseases using R package SuperExactTest (see text footnote 1) (Wang et al., 2015). The results are shown in **Supplementary Table 2**. In **Supplementary Table 2**, degree 1 represents the original gene lists of the eight fibrotic diseases, degree 2 means the gene overlaps between any two groups, and degree 3 shows the gene overlaps among any three groups. By that analogy, degree 8 means the gene overlaps among all the eight groups. The data visualization is illustrated in **Figure 1**. The numbers of overlapped genes are listed over the histogram, and the darkness of the color represents how significant the overlap was. The connective tissue growth factor (CTGF) was finally identified as the common key molecule in the process of fibrosis.

The Key KEGG and GO Features of the Eight Fibrotic Diseases

The key KEGG and GO features of the eight fibrotic diseases were screened by MCFS method. As shown in **Supplementary Table 3**, it means that if a gene could influence a specific function, it may cause a certain fibrotic disease.

As for eye fibrosis, the top three GO terms are GO:0033693 neurofilament bundle assembly, GO:1904530 negative regulation of actin filament binding, and GO:0031113 regulation of microtubule polymerization, respectively. GO:0033693 is associated with neurofilament bundle assembly, which means the assembly of neurofilaments into bundles, in which the filaments are longitudinally oriented, with numerous cross-bridges between them. GO:1904530 is related to negative regulation of actin filament binding, which means reducing physiological activities of actin filament binding. GO:0031113 is connected with the normal physiological activities of microtubule polymerization. Corneal fibrosis is the major type of eye fibrosis. Vimentin, a major structural type III intermediate filament, is a required component of keratocyte activation and differentiation corneal fibrosis, which often accelerates the process of fibrosis (Das et al., 2014).

As for heart fibrosis, the top three GO terms are GO:0032971 regulation of muscle filament sliding, GO:0070296 sarcoplasmic reticulum calcium ion transport, and GO:1990584 troponin complex, respectively. GO:0032971 is in connection with the process that regulates the frequency, rate, or extent of muscle filament sliding. GO:0070296 determines the movement of calcium ions, and GO:1990584 is associated with the cardiac troponin complex and influences muscle contraction. Therefore, muscle filament sliding and calcium ions have been proven to play important roles in the process of hypertrophic cardiomyopathy and heart fibrosis (Huang et al., 2014).

As for hepatic fibrosis, the top three GO terms are GO:0047747 cholate-CoA ligase activity, GO:0008508 bile acid:sodium symporter activity, and GO:0051264 mono-olein transacylation activity, respectively. GO:0047747 affects the activity of cholate-CoA ligase, which catalyzes some reactions in liver. GO:0008508 is related with bile acid and sodium ion transport. GO:0051264 is connected with mono-olein metabolism. Serum bile acids and total cholesterol (TC) are closely related to liver cirrhosis; the potential diagnostic value of total bile acid-to-cholesterol ratio (TBA/TC) for liver fibrosis has been proven (Yan et al., 2020).

As for intestinal fibrosis, the top three GO terms are GO:0032500 muramyl dipeptide binding, GO:0032498 detection of muramyl dipeptide, and GO:0045076 regulation of interleukin 2 (IL-2) biosynthetic process, respectively. GO:0032500 is related with muramyl dipeptide binding, whereas GO:0032498 is associated with detection of muramyl dipeptide. GO:0045076

[3] https://bioconductor.org/packages/KEGGREST/

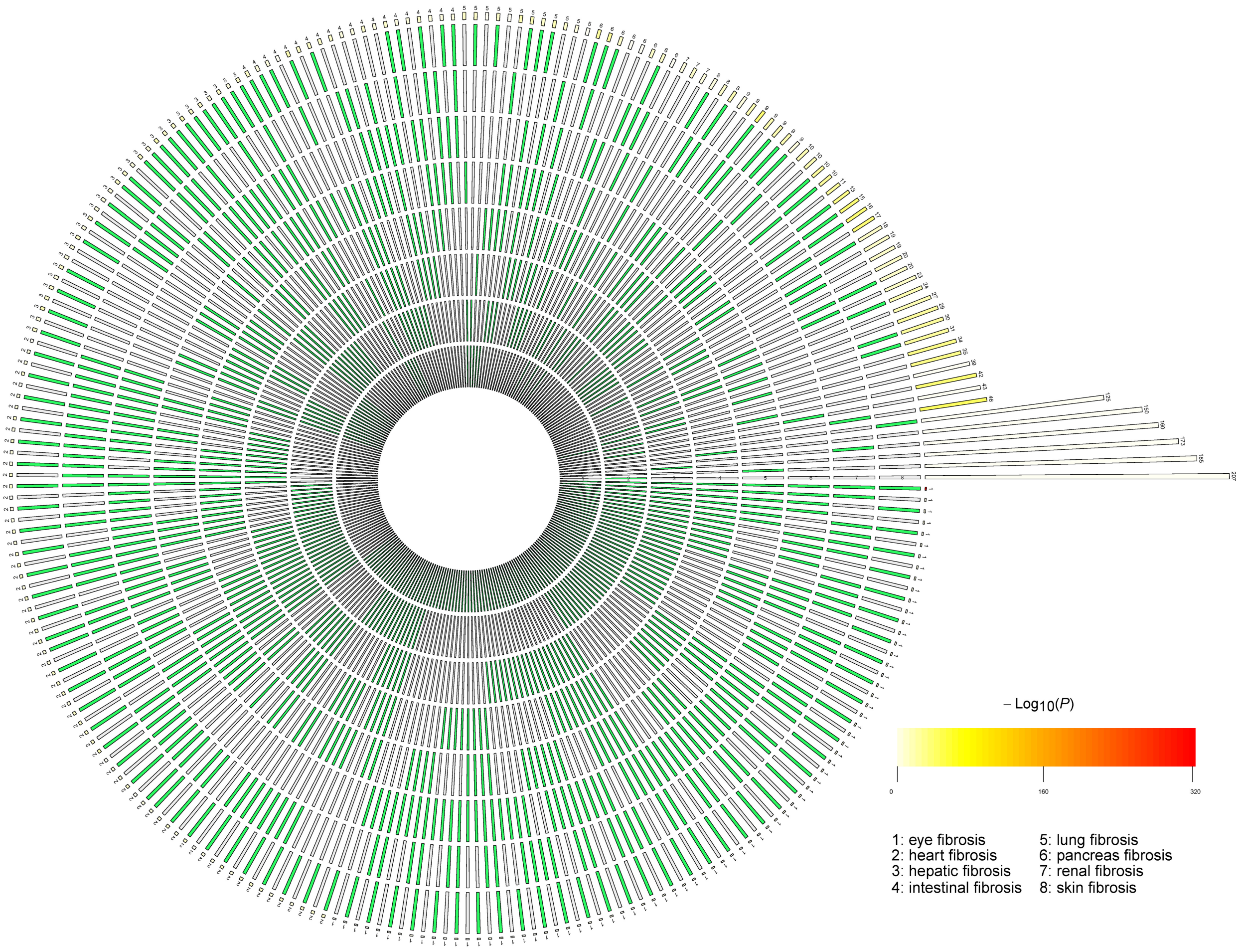


FIGURE 1 | The number of overlapped genes among the eight fibrotic diseases. A circular plot illustrating all possible intersections and the corresponding statistics. The eight circles from inside to outside represent the eight fibrotic diseases (1, eye fibrosis; 2, heart fibrosis; 3, hepatic fibrosis; 4, intestinal fibrosis; 5, lung fibrosis; 6, pancreas fibrosis; 7, renal fibrosis; and 8, skin fibrosis), respectively. The height of the bars in the outer layer is proportional to the intersection sizes, as indicated by the numbers on the top of the bars. The color intensity of the bars represents the *P*-value significance of the intersections.

regulates the process of IL-2 in fibrosis, which has also been proven in patients with cirrhosis and ascitic fluid (Juanola et al., 2016).

As for lung fibrosis, the top three GO terms are GO:0070950 regulation of neutrophil mediated killing of bacterium, GO:0070951 regulation of neutrophil mediated killing of Gram-negative bacterium, and GO:0004957 prostaglandin E receptor activity, respectively. GO:0070950 is related with regulation of neutrophil mediated killing of bacterium. GO:0070951 participates in regulation of neutrophil-mediated killing of Gram-negative bacterium. GO:0004957 means fibrogenesis via prostaglandin E receptor activity. It has been reported that neutrophil-mediated Gram-negative bacterial killing was connected with the cystic fibrosis (CF) lung (Vega-Carrascal et al., 2014).

As for pancreas fibrosis, the top three GO terms are GO:2000878-positive regulation of oligopeptide transport, GO:2000880-positive regulation of dipeptide transport, and GO:2001150-positive regulation of dipeptide transmembrane transport, respectively. All of the three are related to peptide transport. GO:2000878 is associated with positive regulation of oligopeptide transport, whereas GO:2000880 with positive regulation of dipeptide transport. GO:2001150 is related to positive regulation of dipeptide transmembrane transport. CF in the pancreas is characterized by an abnormality in cAMP-regulated chloride transport, which supports the findings of the predicted GO terms (Marino et al., 1991).

As for renal fibrosis, the top three GO terms are GO:0072015 glomerular visceral epithelial cell development, GO:0036057 slit diaphragm, and GO:0005362 low-affinity glucose:sodium symporter activity, respectively. GO:0072015 affects glomerular visceral epithelial cell development and therefore influences its formation to the mature structure. GO:0036057 associated a specialized cell–cell junction, which affects glomerular filtration. GO:0005362 is related to the transfer function of a solute. Renal fibrosis is often caused by renal glomerular sclerosis and interstitial fibrosis. Therefore, glomerular visceral epithelial cell development and formation, glomerular filtration, and transfer function act as the internal causes of renal fibrosis (Qi et al., 2020).

As for skin fibrosis, the top three GO terms are GO:0005600 collagen type XIII trimer, GO:0030936 transmembrane collagen trimer, and GO:0030316 osteoclast differentiation, respectively. GO:0005600 plays a role by collagen type XIII trimer, whereas GO:0030936 via transmembrane collagen trimer. Collagen trimer contributes to derangements in extracellular matrix (ECM) remodeling and leads to fibrosis (Madahar et al., 2018).

The Cross-Talks Between Different Fibrotic Diseases

From the key KEGG and GO features of all the eight fibrotic diseases, we interestingly found overlaps of pathways within some specific fibrotic diseases. For example, renal fibrosis and skin fibrosis jointly influence GO:1901890-positive regulation of cell junction assembly. Some researchers have demonstrated that in renal fibrosis, MG132 successfully sustained cytoskeletal assembly and tight junction, preventing EMT process via RhoA-dependent TGF-β1 pathway, whereas in systemic sclerosis, endothelial junction–associated protein plays vital importance to the pathogenicity (Kanno et al., 2017).

To explore the cross-talk between renal fibrosis and skin fibrosis, we mapped the genes of renal fibrosis, the genes of skin fibrosis, and the genes of GO:1901890-positive regulation of cell junction assembly, which was the common GO feature between renal fibrosis and skin fibrosis, onto STRING network (**Figure 2**). In **Figure 2**, genes in red refer to the overlaps between renal fibrosis and skin fibrosis, whereas the specific genes in renal fibrosis, skin fibrosis, and GO:1901890 are shown in light yellow, light blue, and pink circles, respectively. As illustrated in **Figure 2**, the overlapped genes between renal fibrosis and skin fibrosis included CCL2, SIRT1, KLF5, PPARG, AKT1, SHH, NOTCH, SMAD7, TGFB1, CTNNB1, MMP2, CTGF, FN1, ITGB1, PLAUR, MMP14, NOX4, and COL1A1.

DISCUSSION

Fibrosis is a pathological characteristic of most chronic inflammatory diseases, and many deep learning methods have been developed to study human diseases (Wynn and Ramalingam, 2012; Chen Q. et al., 2019; Cheng and Ghany, 2020; Feng et al., 2020; Lan et al., 2020; Zhao et al., 2020). In recent years, fibrosis is recognized as a main reason of the occurrence of adverse events in many chronic inflammatory diseases. However, the underlying mechanisms in different organs are various and the generality among diverse fibrotic diseases still need to be uncovered. In this study, we applied a new computational method incorporating several machine learning algorithms to explore the common controlling genes and their corresponding pathways among eight different organs' fibrosis.

Common Genes

In our study, CTGF was identified as the common regulatory gene in the eight kinds of fibrotic diseases by MCFS method. It has been around 30 years since the discovery of CTGF from human umbilical vein endothelial cells. In previous researches, CTGF plays an important role in diverse diseases, including cancers, neurodegenerative diseases, systemic sclerosis, kidney diseases, pancreatic diseases, and so on, which means CTGF expresses generally. Mao et al. (2019) demonstrated that megakaryocytic leukemia 1 (MKL1) mediates TGF-β–induced CTGF transcription to promote renal fibrosis. CTGF knockdown dampened TGF-β–induced profibrogenic response in renal tubular epithelial cells. In cardiac fibrosis, Tan et al. (2019) developed an the lamin gene (LMNA) dilated cardiomyopathy (DCM) mouse model and found silencing of cardiac LMNA-induced DCM with associated cardiac fibrosis and inflammation and further uncovered that Yy1 suppresses DCM and cardiac fibrosis through regulation of bmp7 and CTGF. Besides, another study also proved that in patients with rheumatic heart disease, high CTGF expression was related to enlarged left atrial diameter, atrial fibrosis, and atrial anatomical remodeling (Chen J.Q. et al., 2019). In lung fibrosis, disintegrin and metalloproteinase 17, and CTGF were found to play critical roles in fibrotic procedures and contribute to lung fibrosis (Chen et al., 2018).

With regard to the gene overlaps of pathways within some specific fibrotic diseases, we have identified some common pathways and genes within renal fibrosis and skin fibrosis. For example, in chronic renal allograft injury resulting in progressive interstitial fibrosis, early urinary CCL2 is an independent predictor for the subsequent development of interstitial fibrosis and tubular atrophy at 24 months (Ho et al., 2010). Similarly, in systemic sclerosis (skin fibrosis), the levels of circulating CCL2, CCL3, and CCL5 chemokines were significantly higher in patients with systemic sclerosis than in controls.

Common Pathways

Fibrosis and resultant organ failure result in approximately one-third of deaths worldwide (Zeisberg and Kalluri, 2013). Now that fibrosis is common and has harmful effects in almost all organs, it is a potential therapeutic target. As for predicted pathways, we have demonstrated some new pathways associated with the specific fibrotic diseases. In intestinal fibrosis, the GO term, GO:0045076, regulates the process of IL-2 in fibrosis. In patients with cirrhosis and ascitic fluid, Juanola et al. (2016) identified how the role of regulatory T cells played for compensating the inflammatory environment in cirrhosis when norfloxacin was applied, and they found norfloxacin immunomodulatory effect on IL-2 and interferon γ reduction. In lung fibrosis, GO:0070951 participates in regulation of neutrophil-mediated killing of Gram-negative bacterium. It has been reported that neutrophil-mediated Gram-negative bacterial killing was connected with the CF lung. The underlying mechanism was that galectin-9 (Gal-9) signaling through the T-cell Ig and mucin domain-containing molecule (TIM) and neutrophil TIM-3/Gal-9 signaling is perturbed in the CF airways due to proteolytic degradation of the receptor (Vega-Carrascal et al., 2014). GO:0004957 means fibrogenesis via prostaglandin E receptor activity. As Sieber et al. (2018) demonstrated, pathological features of pulmonary

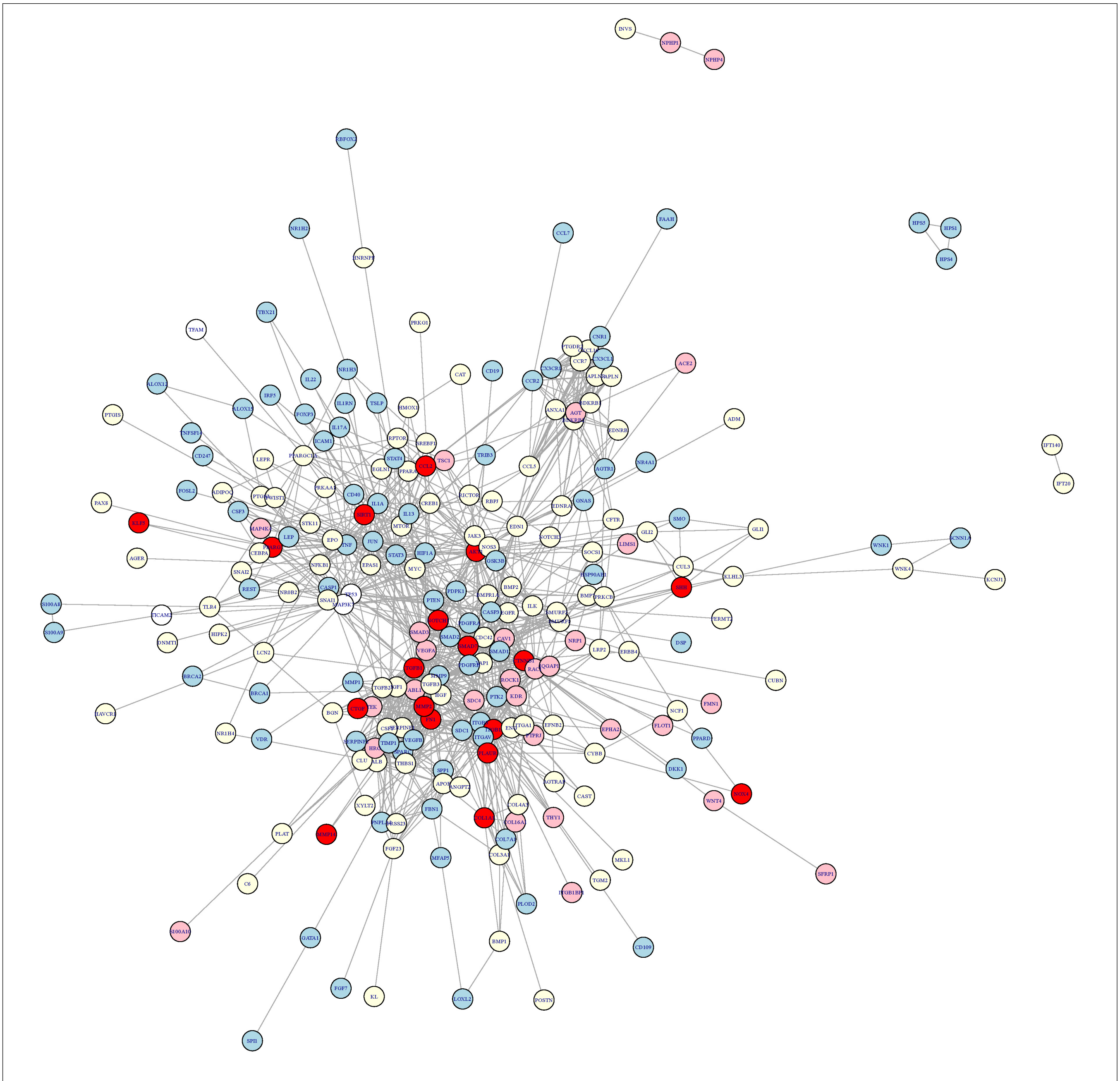


FIGURE 2 | The cross-talk network between renal fibrosis and skin fibrosis. The genes in red refer to the overlaps between renal fibrosis and skin fibrosis, whereas the specific genes in renal fibrosis, skin fibrosis, and GO:1901890 are shown in light yellow, light blue, and pink circles, respectively. The overlapped genes between renal fibrosis and skin fibrosis included CCL2, SIRT1, KLF5, PPARG, AKT1, SHH, NOTCH, SMAD7, TGFB1, CTNNB1, MMP2, CTGF, FN1, ITGB1, PLAUR, MMP14, NOX4, and COL1A1.

fibrosis include accumulation of myofibroblasts and increased ECM deposition in lung tissue; they developed a new assay with therapeutic potential in pulmonary fibrosis that acts via EP2 and EP4 receptors. In heart and renal fibrosis, angiotensin-converting enzyme inhibitors and angiotensin-receptor blockers that ameliorate cardiac and renal damage and fibrosis through many pathways such as TGF-β and SMAD pathways (Lambers Heerspink et al., 2013). In liver fibrosis, as hepatocytes process the ability of regeneration, intervention is needed for patients with hepatic fibrosis. For example, colchicine has been proven to prevent hepatic fibrosis via suppressing collagen secretion (Rockey, 2013). As the common pathways and genes were identified by our new computational method, old drugs for a specific fibrosis may be effective for another organ fibrosis.

CONCLUSION

In conclusion, we identified that CTGF is acted as the key molecule regulating the processes of fibrogenesis and some common pathways within different fibrotic diseases via a new computational method. We hope to offer a new insight into the cellular and molecular mechanisms underlying fibrosis and therefore help lead to the development of new drugs that specifically delay or even improve the symptoms of fibrosis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

TH, XP, and YC: conception and design and administrative support. CG, XS, and XD: collection and assembly of the data and data analysis and interpretation. All authors wrote the manuscript and approved the submitted version.

FUNDING

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), National Key R&D Program of China (2017YFC1201200), the National Natural Science Foundation of China (81900280 and 31701151), Shanghai Sailing Program (19YF1431600), National Key R&D Program of China (2018YFC0910403), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), and The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.627396/full#supplementary-material

Supplementary Table 1 | The reported genes of the eight fibrotic diseases.

Supplementary Table 2 | The overlapped genes of the eight fibrotic diseases.

Supplementary Table 3 | The key KEGG and GO features of the eight fibrotic diseases.

REFERENCES

Ballester, B., Milara, J., and Cortijo, J. (2019). Idiopathic pulmonary fibrosis and lung cancer: mechanisms and molecular targets. *Int. J. Mol. Sci.* 20:593. doi: 10.3390/ijms20030593

Bataller, R., and Brenner, D. A. (2005). Liver fibrosis. *J. Clin. Invest.* 115, 209–218. doi: 10.1172/JCI24282

Chen, H. Y., Lin, C. H., and Chen, B. C. (2018). ADAM17/EGFR-dependent ERK activation mediates thrombin-induced CTGF expression in human lung fibroblasts. *Exp. Cell Res.* 370, 39–45. doi: 10.1016/j.yexcr.2018.06.008

Chen, J. Q., Guo, Y. S., Chen, Q., Cheng, X. L., Xiang, G. J., Chen, M. Y., et al. (2019). TGFbeta1 and HGF regulate CTGF expression in human atrial fibroblasts and are involved in atrial remodelling in patients with rheumatic heart disease. *J. Cell Mol. Med.* 23, 3032–3039. doi: 10.1111/jcmm.14165

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y. P., et al. (2019). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2936476 Online ahead of print.

Chen, L., Pan, X., Guo, W., Gan, Z., Zhang, Y.-H., Niu, Z., et al. (2020). Investigating the gene expression profiles of cells in seven embryonic stages with machine learning algorithms. *Genomics* 112, 2524–2534. doi: 10.1016/j.ygeno.2020.02.004

Chen, L., Zhang, Y. H., Zheng, M., Huang, T., and Cai, Y. D. (2016). Identification of compound-protein interactions through the analysis of gene ontology. KEGG enrichment for proteins and molecular fragments of compounds. *Mol. Genet. Genom.* 291, 2065–2079. doi: 10.1007/s00438-016-1240-x

Cheng, X., and Ghany, M. G. (2020). *Hepatitis C virus—from Discovery to the Nobel Prize*. Cambridge, MA: Cell Press.

Das, S. K., Gupta, I., Cho, Y. K., Zhang, X., Uehara, H., Muddana, S. K., et al. (2014). Vimentin knockdown decreases corneal opacity. *Invest. Ophthalmol. Vis. Sci.* 55, 4030–4040. doi: 10.1167/iovs.13-13494

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Feng, M., Pan, Y., Kong, R., and Shu, S. (2020). Therapy of primary liver cancer. *Innovation* 1:100032. doi: 10.1016/j.xinn.2020.100032

Gu, C., Huang, Z., Chen, X., Liu, C., Rocco, G., Zhao, S., et al. (2020a). TEAD4 promotes tumor development in patients with lung adenocarcinoma via ERK signaling pathway. *Biochim Biophys. Acta Mol. Basis Dis.* 1866:165921. doi: 10.1016/j.bbadis.2020.165921

Gu, C., Shi, X., Dai, C., Shen, F., Rocco, G., Chen, J., et al. (2020b). RNA m6A modification in cancers: molecular mechanisms and potential clinical applications. *Innovation.* 1:100066. doi: 10.1016/j.xinn.2020.100066

Gu, C., Shi, X., Huang, Z., Chen, J., Yang, J., Shi, J., et al. (2020c). A comprehensive study of construction and analysis of competitive endogenous RNA networks in lung adenocarcinoma. *Biochim Biophys. Acta Proteins Proteom.* 1868:140444. doi: 10.1016/j.bbapap.2020.140444

Gu, C., Pan, X., Chen, Y., Yang, J., Zhao, H., and Shi, J. (2018). Short-term and mid-term survival in bronchial sleeve resection by robotic system versus thoracotomy for centrally located lung cancer. *Eur. J. Cardiothorac. Surg.* 53, 648–655. doi: 10.1093/ejcts/ezx355

Ho, J., Rush, D. N., Gibson, I. W., Karpinski, M., Storsley, L., Bestland, J., et al. (2010). Early urinary CCL2 is associated with the later development of interstitial fibrosis and tubular atrophy in renal allografts. *Transplantation* 90, 394–400. doi: 10.1097/TP.0b013e3181e6424d

Huang, W., Liang, J., Kazmierczak, K., Muthu, P., Duggal, D., Farman, G. P., et al. (2014). Hypertrophic cardiomyopathy associated Lys104Glu mutation in the myosin regulatory light chain causes diastolic disturbance in mice. *J. Mol. Cell Cardiol.* 74, 318–329. doi: 10.1016/j.yjmcc.2014.06.011

Jiao, D., and Yang, S. (2020). Overcoming resistance to drugs targeting KRASG12C mutation. *Innovation* 1:100035. doi: 10.1016/j.xinn.2020.100035

Juanola, O., Gomez-Hurtado, I., Zapater, P., Moratalla, A., Caparros, E., Pinero, P., et al. (2016). Selective intestinal decontamination with norfloxacin enhances a regulatory T cell-mediated inflammatory control mechanism in cirrhosis. *Liver Int.* 36, 1811–1820. doi: 10.1111/liv.13172

Kanno, Y., Shu, E., Kanoh, H., Matsuda, A., and Seishima, M. (2017). alpha2AP regulates vascular alteration by inhibiting VEGF signaling in systemic sclerosis: the roles of alpha2AP in vascular dysfunction in systemic sclerosis. *Arthritis Res. Ther.* 19:22. doi: 10.1186/s13075-017-1227-y

Kelly, D., Verkade, H. J., Rajanayagam, J., McKiernan, P., Mazariegos, G., and Hubscher, S. (2016). Late graft hepatitis and fibrosis in pediatric liver allograft recipients: current concepts and future developments. *Liver Transpl.* 22, 1593–1602. doi: 10.1002/lt.24616

Lambers Heerspink, H. J., de Borst, M. H., Bakker, S. J., and Navis, G. J. (2013). Improving the efficacy of RAAS blockade in patients with chronic kidney disease. *Nat. Rev. Nephrol.* 9, 112–121. doi: 10.1038/nrneph.2012.281

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). LDICDL: LncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.3034910 Online ahead of print.

Li, J., Lu, L., Zhang, Y.-H., Xu, Y., Liu, M., Feng, K., et al. (2020a). Identification of leukemia stem cell expression signatures through monte carlo feature selection strategy and support vector machine. *Cancer Gene Therapy* 27, 56–69. doi: 10.1038/s41417-019-0105-y

Li, J., Xu, Q., Wu, M., Huang, T., and Wang, Y. (2020b). Pan-Cancer classification based on self-normalizing neural networks and feature selection. *Front. Bioeng. Biotechnol.* 8:766. doi: 10.3389/fbioe.2020.00766

Madahar, P., Duprez, D. A., Podolanczuk, A. J., Bernstein, E. J., Kawut, S. M., Raghu, G., et al. (2018). Collagen biomarkers and subclinical interstitial lung disease: the multi-ethnic study of atherosclerosis. *Respir. Med.* 140, 108–114. doi: 10.1016/j.rmed.2018.06.001

Mao, L., Liu, L., Zhang, T., Wu, X., Zhang, T., and Xu, Y. (2019). MKL1 mediates TGF-beta-induced CTGF transcription to promote renal fibrosis. *J. Cell Physiol.* 235, 4790–4803. doi: 10.1002/jcp.29356

Marino, C. R., Matovcik, L. M., Gorelick, F. S., and Cohn, J. A. (1991). Localization of the cystic fibrosis transmembrane conductance regulator in pancreas. *J. Clin. Invest.* 88, 712–716. doi: 10.1172/JCI115358

Pan, X., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y. H., Kong, X. Y., et al. (2019a). Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms. *Int. J. Mol. Sci.* 20:2185. doi: 10.3390/ijms20092185

Pan, X., Hu, X., Zhang, Y.-H., Chen, L., Zhu, L., Wan, S., et al. (2019b). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genom.* 294, 95–110. doi: 10.1007/s00438-018-1488-1484

Pan, X., Zeng, T., Yuan, F., Zhang, Y. H., Chen, L., Zhu, L., et al. (2019c). Screening of methylation signature and gene functions associated with the subtypes of isocitrate dehydrogenase-mutation gliomas. *Front. Bioeng. Biotechnol.* 7:339. doi: 10.3389/fbioe.2019.00339

Pan, X., Hu, X., Zhang, Y. H., Feng, K., Wang, S. P., Chen, L., et al. (2018). Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes (Basel)* 9:208. doi: 10.3390/genes9040208

Pan, X., Zeng, T., Zhang, Y. H., Chen, L., Feng, K., Huang, T., et al. (2020). Investigation and prediction of human interactome based on quantitative features. *Front. Bioeng. Biotechnol.* 8:730. doi: 10.3389/fbioe.2020.00730

Qi, S. S., Zheng, H. X., Jiang, H., Yuan, L. P., and Dong, L. C. (2020). Protective effects of chromium picolinate against diabetic-induced renal dysfunction and renal fibrosis in streptozotocin-induced diabetic rats. *Biomolecules* 10:398. doi: 10.3390/biom10030398

Ren, X., Wang, S., and Huang, T. (2020). Decipher the connections between proteins and phenotypes. *Biochim Biophys. Acta Proteins Proteom* 1868:140503. doi: 10.1016/j.bbapap.2020.140503

Rockey, D. C. (2013). Translating an understanding of the pathogenesis of hepatic fibrosis to novel therapies. *Clin. Gastroenterol. Hepatol.* 11, 224–31.e1-5. doi: 10.1016/j.cgh.2013.01.005

Rockey, D. C., Bell, P. D., and Hill, J. A. (2015). Fibrosis–A common pathway to organ injury and failure. *N. Engl. J. Med.* 373:96. doi: 10.1056/NEJMc1504848

Shi, X., Huang, T., Wang, J., Liang, Y., Gu, C., Xu, Y., et al. (2018). Next-generation sequencing identifies novel genes with rare variants in total anomalous pulmonary venous connection. *EBioMedicine* 38, 217–227. doi: 10.1016/j.ebiom.2018.11.008

Sieber, P., Schafer, A., Lieberherr, R., Le Goff, F., Stritt, M., Welford, R. W. D., et al. (2018). Novel high-throughput myofibroblast assays identify agonists with therapeutic potential in pulmonary fibrosis that act via EP2 and EP4 receptors. *PLoS One* 13:e0207872. doi: 10.1371/journal.pone.0207872

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2018). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tan, C. Y., Wong, J. X., Chan, P. S., Tan, H., Liao, D., Chen, W., et al. (2019). Yin Yang 1 suppresses dilated cardiomyopathy and cardiac fibrosis through regulation of Bmp7 and Ctgf. *Circ. Res.* 125, 834–846. doi: 10.1161/CIRCRESAHA.119.314794

Vega-Carrascal, I., Bergin, D. A., McElvaney, O. J., McCarthy, C., Banville, N., Pohl, K., et al. (2014). Galectin-9 signaling through TIM-3 is involved in neutrophil-mediated Gram-negative bacterial killing: an effect abrogated within the cystic fibrosis lung. *J. Immunol.* 192, 2418–2431. doi: 10.4049/jimmunol.1300711

Wang, M., Zhao, Y., and Zhang, B. (2015). Efficient test and visualization of multi-set intersections. *Sci. Rep.* 5:16923. doi: 10.1038/srep16923

Wynn, T. A., and Ramalingam, T. R. (2012). Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat. Med.* 18, 1028–1040. doi: 10.1038/nm.2807

Yan, L. T., Wang, L. L., Yao, J., Yang, Y. T., Mao, X. R., Yue, W., et al. (2020). Total bile acid-to-cholesterol ratio as a novel noninvasive marker for significant liver fibrosis and cirrhosis in patients with non-cholestatic chronic hepatitis B virus infection. *Medicine (Baltimore)* 99:e19248. doi: 10.1097/MD.0000000000019248

Zeisberg, M., and Kalluri, R. (2013). Cellular mechanisms of tissue fibrosis. 1. common and organ-specific mechanisms associated with tissue fibrosis. *Am. J. Physiol. Cell Physiol.* 304, C216–C225. doi: 10.1152/ajpcell.00328.2012

Zhao, K., Liu, A., and Xia, Y. (2020). Insights into hepatitis B Virus DNA integration-55 years after virus discovery. *Innovation* 1:100034. doi: 10.1016/j.xinn.2020.100034

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

METHODS
published: 18 January 2021
doi: 10.3389/fgene.2020.632901

Integrating Somatic Mutations for Breast Cancer Survival Prediction Using Machine Learning Methods

Zongzhen He[1], *Junying Zhang*[1]*, *Xiguo Yuan*[1] *and Yuanyuan Zhang*[2]

[1] *School of Computer Science and Technology, Xidian University, Xi'an, China,* [2] *School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China*

Breast cancer is the most common malignancy in women, and because it has a high mortality rate, it is urgent to develop computational methods to increase the accuracy of breast cancer survival predictive models. Although multi-omics data such as gene expression have been extensively used in recent studies, the accurate prognosis of breast cancer remains a challenge. Somatic mutations are another important and promising data source for studying cancer development, and its effect on the prognosis of breast cancer remains to be further explored. Meanwhile, these omics datasets are high-dimensional and redundant. Therefore, we adopted multiple kernel learning (MKL) to efficiently integrate somatic mutation to currently molecular data including gene expression, copy number variation (CNV), methylation, and protein expression data for the prediction of breast cancer survival. Before integration, the maximum relevance minimum redundancy (mRMR) feature selection method was utilized to select features that present high relevance to survival and low redundancy among themselves for each type of data. The experimental results demonstrated that the proposed method achieved the most optimal performance and there was a remarkable improvement in the prediction performance when somatic mutations were included, indicating that somatic mutations are critical for improving breast cancer survival predictions. Moreover, mRMR was superior to other feature selection methods used in previous studies. Furthermore, MKL outperformed the other traditional classifiers in multi-omics data integration. Our analysis indicated that through employing promising omics data such as somatic mutations and harnessing the power of proper feature selection methods and effective integration frameworks, the breast cancer survival predictive accuracy can be further increased, thereby providing a more optimal clinical diagnosis and more effective treatment for breast cancer patients.

Keywords: breast cancer, multi-omics, survival prediction, somatic mutation, mRMR, MKL

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Ergude Bao,
Beijing Jiaotong University, China
Zhenhua Yu,
Ningxia University, China

***Correspondence:**
Junying Zhang
jyzhang@mail.xidian.edu.cn

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *24 November 2020*
Accepted: *30 December 2020*
Published: *18 January 2021*

Citation:
He Z, Zhang J, Yuan X and Zhang Y (2021) Integrating Somatic Mutations for Breast Cancer Survival Prediction Using Machine Learning Methods. Front. Genet. 11:632901. doi: 10.3389/fgene.2020.632901

INTRODUCTION

Breast cancer is the most common malignant tumor in women. Although there are millions of breast cancer survivors in the United States, breast cancer is the main cause of cancer-related deaths worldwide because of its high mortality rate (Ferlay et al., 2010). Thus, it is urgent to design highly accurate methods to predict the survival of breast cancer patients. Accordingly, effective survival predictors could finally contribute to the reduction of the overall mortality of breast cancer and could further improve the life quality and increase the lifespan of breast cancer patients.

Recently, the Cox regression model (Yuan et al., 2014; Xu et al., 2016) and traditional machine learning classification methods, such as support vector machine (SVM) (Xu et al., 2013), Bayes classifier (Gevaert et al., 2006), and random forest (RF) (Nguyen et al., 2013), have been widely deployed to identify breast cancer prognostic biomarkers. Multiple survival prediction models have been mainly developed based on gene expression data. The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, 2013; Brennan et al., 2014) provides multiple types of molecular data such as gene expression (Exp), copy number variation (CNV), methylation (Methy), protein expression (Protein), and somatic mutation (SM) data for various cancers, including breast cancer. Moreover, the advancement of machine learning technologies enables various data types to be combined within a model (Chen et al., 2019; Lan et al., 2020), which may increase the accuracy of predictive models.

One of the biggest challenges in breast cancer research involves the effective combination of heterogeneous data sources into survival prediction models, making the selection of a proper integration method essential. In previous studies (Seoane et al., 2014; Zhang et al., 2016; Sun et al., 2018; Zhang A. et al., 2019; Zhang Y. et al., 2019), multiple kernel learning (MKL) (Lanckriet et al., 2004; Rakotomamonjy et al., 2008; Kloft et al., 2011) was successfully used to integrate different types of data into a universal model to distinguish short-term and long-term cancers survivors. MKL uses different kernels for different types of data, and then trains the weight of each kernel to select the best combination of kernel functions for classification. These studies have demonstrated that models that were obtained using integrated data improved the performance of survival prediction compared to models that used only one single data type.

A previous study (Sun et al., 2018) showed that MKL outperformed Cox-based regression models for breast cancer survival prediction. However, omics data, such as Exp, CNV, and methylation data, are usually extremely high-dimensional and redundant (Dey et al., 1990). In the previous study (Sun et al., 2018), information gain ratio (IGR) was utilized to select survival relevant features from multi-omics data, but the redundancy of dataset features was not considered. Despite the promising performance of the above MKL-based studies for breast cancer prognosis, somatic mutations are rarely considered for breast cancer survival prediction due to their complexity and heterogeneity in serious disease. Therefore, there is still much room to increase the accuracy of breast cancer survival models by incorporating somatic mutations into the MKL model.

Currently, somatic mutations are strongly correlated with the clinical symptoms of breast cancer (Griffith et al., 2018), and they have been successfully adopted for the classification of primary cancer sites (Chen et al., 2015) and identification of survival-related cancer subtypes (Hofree et al., 2013; He et al., 2017; Ronen et al., 2018; Arslanturk et al., 2020). Somatic mutations are sparse but common mutations of that offer less accuracy in the prediction of cancer survival (Zhang et al., 2018; Ye et al., 2019). Previous studies (Haricharan et al., 2014; Griffith et al., 2018; Zhang et al., 2018; Ye et al., 2019) have reported that mutations enriched in specific pathways have shown potential for breast cancer survival prediction. The authors of a previous study (Griffith et al., 2018) stated that uncommon recurrent somatic mutations should be further explored to explain breast cancer survival outcomes. In the present study, the effect of somatic mutations on the integrated prognosis of breast cancer is explored.

In the present study, we applied the state-of-the-art MKL method in the integration of somatic mutation datasets with previously used omics data, including Exp, CNV, Methy, and Protein, to train and test an integrated breast cancer survival prediction model. The maximum relevance minimum redundancy (mRMR) algorithm (Ding and Peng, 2005; Radovic et al., 2017) was used to alleviate the redundancy of the data, by simultaneously selecting highly predictive but non-redundant features from each type of molecular data. Then, selected features from multiple data type were integrated into the MKL classification.

In order to gauge the performance of our method, first, the newly introduced method was compared with different single data types and integrated datasets to verify the effectiveness of somatic mutations, and the results indicated that there was a remarkable improvement in the prediction performance when somatic mutations were included. Different feature selection algorithms were then studied, and the experimental results demonstrated that mRMR was the most optional among them. Furthermore, the MKL classification method was compared with other traditional classifiers, and the experimental results proved the superiority of MKL in data integration. Finally, the newly introduced model was validated in an independent validation dataset and achieved a promising high accuracy in survival prediction. According to the results, the most optimal performance was achieved by our method, which demonstrated the feasibility of integrating somatic mutations in the prognostic models and the usefulness of mRMR and MKL in breast cancer prognosis.

The reminder of this article is organized as follows. A workflow of our proposed method and related methods are described. Next, comparative studies were carried out to evaluate the performance of the proposed methods and their comparison methods, as well as to analyze the most informative features discovered by our model. Then, we applied our model on the validation dataset. Finally, the proposed method is discussed, and it is expected to undergo the improvement in future studies.

MATERIALS AND METHODS

Workflow of the Proposed Method

The workflow chart of the proposed method is shown in **Figure 1**. Preprocessing of the input dataset initially occurred, during which entire datasets were randomly divided into a learning dataset (80% of the entire dataset) and validation dataset (20%). Then, three main steps were carried out to realize the prediction of breast cancer prognosis.

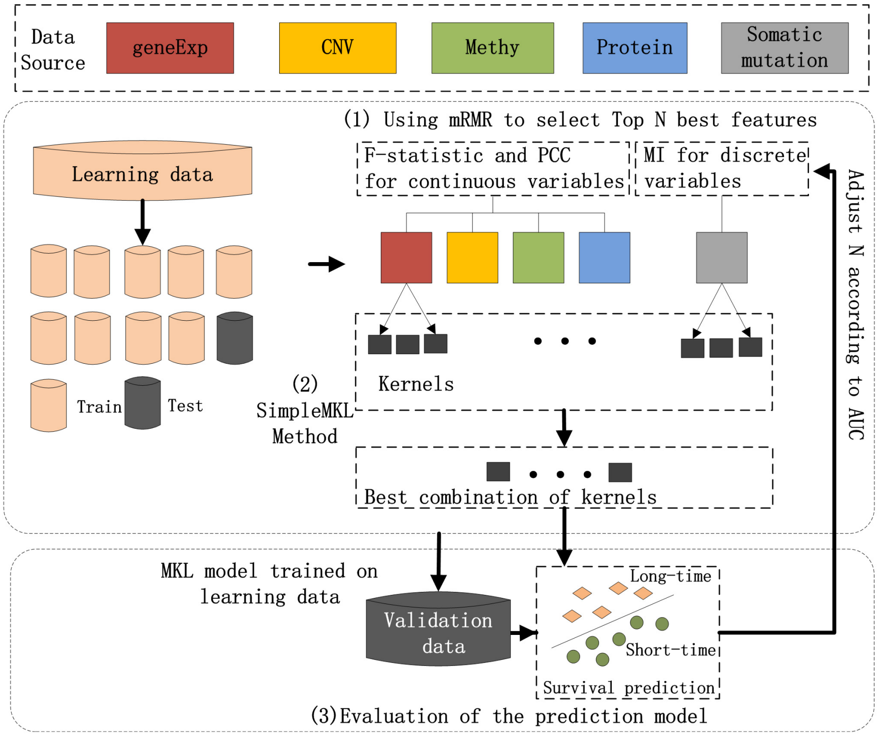

Data Source
geneExp
CNV
Methy
Protein
Somatic mutation
(1) Using mRMR to select Top N best features
F-statistic and PCC for continuous variables
MI for discrete variables
Learning data
Train
Test
(2) SimpleMKL Method
Kernels
Best combination of kernels
Adjust N according to AUC
MKL model trained on learning data
Validation data
Long-time
Short-time
Survival prediction
(3)Evaluation of the prediction model

FIGURE 1 | Workflow of the hybrid combination of the MKL model with the mRMR feature selection method to integrate five types of molecular data for the prognosis of breast cancer. (1) The most N-informative features were separately selected using the mRMR method for each type of data in the learning dataset; (2) SimpleMKL with 10-fold cross-validation was deployed on the learning dataset for breast cancer prognosis to train an optimal model; and (3) the prediction model on learning dataset and the validation dataset were evaluated.

The three main steps include: (1) The most N-informative features were separately selected using the mRMR method for each type of data in the learning dataset; (2) SimpleMKL with 10-fold cross-validation was deployed on the learning dataset for breast cancer prognosis to train an optimal model; and (3) the prediction model on learning dataset and the validation dataset were evaluated for their ability to learn data. A detailed description of each of the steps is listed below.

Data Input and Preprocessing

The Cancer Genome Atlas provides multiple types of biomolecular data. High-level molecular data for breast cancer were retrieved from TCGA, including gene expression, gene CNV, gene methylation, protein expression, and somatic mutation along with clinical features from the University of California Santa Cruz (UCSC) cancer browser website[1] (Mary et al., 2014). The downloaded dataset consisted of five types of data, including different numbers of samples, and the original data matrixes were structured with rows denoting patient samples and columns denoting features. A total of 139 true normal, seven metastatic, and 13 male patients' samples were removed, and regarding somatic mutations, samples with less than 10 mutations were removed (Hofree et al., 2013; He et al., 2017). We finally obtained 488 primary breast tumors together with survival time, and all samples of them included all of the five aforementioned genomic data types. The details of our dataset are illustrated in **Table 1**. The median age at diagnosis was 57.37, and the median survival time was 42.43 months, which is in agreement with the previous research (Sun et al., 2018).

We followed the protocol from our previously published studies (He et al., 2017, 2019), and we first removed the genes with missing values in more than 10% of samples for gene expression, CNV, gene methylation, protein expression, and somatic mutations. After that, flat variables that had the same values in more than 80% of the samples (non-informative) were discarded except in the case of somatic mutations (Yuan et al., 2014; He et al., 2019). According to the previous study (He et al., 2019), the RNA-Seq gene expression level 3 transcription was log2 transformed and RSEM-normalized (Li and Dewey, 2011). Regarding the CNV features, we directly utilized the gene-level

TABLE 1 | The detailed information in our breast cancer dataset.

Properties	Number
Total population of primary cancer	488
Long-term survivors	119
Short-term survivors	369
Mean age at diagnosis (years old)	57.37
Median survival (months)	42.43

[1] https://xenabrowser.net/datapages/

copy number values that were estimated using the GISTIC2 method (Mermel et al., 2011; Yuan et al., 2017; Yuan et al., 2019, 2020a,b). For gene methylation and protein expression, we directly used the original data with z-score normalization. For somatic mutation, we also directly utilized the original binary data, and in addition, genes that were mutated in more than one sample were reserved for further analysis. The gene expression, CNV, gene methylation, and somatic mutations contained 18,000, 25,000, 22,000, and 14,000 features, respectively, after data filtering, and the properties of these datasets are shown in **Table 2**.

In the present study, the survival prediction for breast cancer was defined as a binary classification problem with a threshold of 5 years as conducted in previous studies (Seoane et al., 2014; Zhang et al., 2016; Sun et al., 2018; Zhang A. et al., 2019). Of the total, 369 out of the final 488 patients with survival shorter than 5 years were considered as short-term survivors, and 119 patients with survival longer than 5 years were considered as long-term survivors. Moreover, the long-term patients were labeled as 1, while short-term patients were labeled as 0. After the initial data preprocessing, the entire dataset was randomly divided into the learning dataset (80%) and validation dataset (20%). For each type of data, we initially conducted the following feature selection on the learning dataset containing 390 breast cancer patients, and trained and tested the integrated MKL model on it to obtain the optimal parameters. Then, we applied the optimal model on the validation dataset that included 98 patients.

mRMR Feature Selection

Five different types of genomic data were used in the present study, as described above, and the number of variables for most types of genomic data exceeded 10,000 after feature preprocessing. However, this large number of features may cause poor performance due to dimensionality and high redundancy (Jain and Zongker, 1997; Jie et al., 2015). Therefore, according to our previous study (He et al., 2019), mRMR was adopted in the present study to select the most useful features for the prognostic model.

The mRMR is a feature selection method that aims to select a subset of features that are highly related to the output classes and have low redundancy between them (Radovic et al., 2017). In the present work, mRMR was deployed to select features from five types of molecular data that are the most highly relevant with respect to survival and the least correlated among themselves. Then, the most relevant features for each molecular dataset were combined to form a candidate feature set to be used for classification. A feature of one type of genomic dataset for the *i*th variable with N individuals is denoted as $v_i \in R^M, i = 1, \ldots, M$, and the survival prediction labels with N individuals as $l \in R$. For label l, mRMR aims to search a feature subset S with k features$\{v_i\}$, which collectively have the maximal relevance (Max-Relevance) $\mathrm{Rel}(S, l)$on the target label l and the minimal redundancy (Min-Redundancy) $\mathrm{Red}(S)$.

TABLE 2 | The properties of five types of genomic data for our breast cancer prediction.

Data types	Feature number
Gene expression	18624
CNV	24774
Gene methylation	21136
Protein expression	170
Somatic mutations	13602

The F-statistic (F) was used to calculate the relevance between feature variables with binary survival terms and the Pearson correlation coefficient (PCC) was used to measure the redundancy for the continuous feature variables of the gene expression, CNV, gene methylation, and protein datasets. Max-Relevance is defined in Eq. 1, where relevance $\mathrm{Rel}(S, l)$ is calculated using the mean value of all F-statistic values F of the individual variables v_i with the label l. In parallel, the Min-RedundancyRed(S) constraint was adopted to select irrelevant features, and is shown as Eq. 2.

$$\max \mathrm{Rel}(S, l), \quad \mathrm{Rel} = \frac{1}{|S|} \sum_{v_i \in S} F(v_i; l), \tag{1}$$

$$\min \mathrm{Red}(S), \quad \mathrm{Red} = \frac{1}{|S|^2} \sum_{v_i, v_j \in S} PCC(v_i; v_j) \tag{2}$$

For binary discrete feature variables of somatic mutation data, the mutual information (MI) was used to calculate both the relevance between feature variables and survival terms, and the redundancy between mutations. Max-Relevance is used to select features satisfying Eq. 3, where relevance $\mathrm{Rel}(S, l)$ is obtained by the mean value of all MI values of individual variable v_i with label l. The Min-Redundancy constraint $\mathrm{Red}(S)$ is used to select irrelevant features, and is shown as Eq. 4.

$$\max \mathrm{Rel}(S, l), \quad \mathrm{Rel} = \frac{1}{|S|} \sum_{v_i \in S} MI(v_i; l), \tag{3}$$

$$\min \mathrm{Red}(S), \quad \mathrm{Red} = \frac{1}{|S|^2} \sum_{v_i, v_j \in S} MI(v_i; v_j) \tag{4}$$

Finally, as shown in Eq. 5, the operator $\phi(\mathrm{Rel}, \mathrm{Red})$ was deployed to simultaneously optimize the two constraints "Max-Relevance" and "Min-Redundancy" based on the MI quotient (MIQ) criterion (Radovic et al., 2017; He et al., 2019) to obtain the best feature subsets, as shown in Eq. 5:

$$\max_{v_k} \phi(\mathrm{Rel}, \mathrm{Red}), \quad \phi = \mathrm{Rel}/\mathrm{Red} \tag{5}$$

The area under the curve (AUC) value is used as a metric to evaluate the performance and the most optimal number of the most relevant and non-redundant features k for each data type was determined by comparing the AUC valued for the models. After the mRMR features were selected for each type of genomic data, the most informative features were combined and used as the input feature set for the classification problems.

Multiple Kernel Learning

In our study, we aimed to integrate multiple types of genomics data, with a focus on somatic mutations. Although the fusion of multiple types of data into one model is one of the most widely used methods for classification, this is not feasible due to the fact that different types of molecular data present different feature representations (Khademi and Nedialkov, 2016). MKL has become a natural method to enhance the interpretability of models and to address the data integration problem. The optimal function can be obtained by constructing a linear weighted combination of predefined M kernels. The optimal combination of kernels is given as Eqs 6 and 7:

$$K(x_i, x_j) = \sum_{\mathrm{m}=1}^{M} d_m K_m(x_i, x_j), \tag{6}$$

$$s.t.\ d_m \geq 0,\ \text{and} \sum_{m=1}^{M} d_m = 1, \tag{7}$$

where d_{m} denotes the weight of the mth different kernel $K_m\ (x_i, x_j)$.

Some methods based on MKL have been proposed and many of them outperformed uni-MKL (Rakotomamonjy et al., 2008; Gönen and Alpaydin, 2011; Kloft et al., 2011). However, most of the weights d_m of the kernels were 0 and thus non-contributory to the MKL model (Ikonomov et al., 2013). In the present work, SimpleMKL (Zhang et al., 2016), which is based on a weighted L2-norm regularization and is more powerful than other methods (Yan et al., 2009), was adopted as our classification model. It employs dual kernels in the of classic kernel optimization problem, which can be presented as Eq. 8:

$$f(x) = \sum_{i=1}^{l} \alpha_i^* K(x_j, x_i) + b^* \tag{8}$$

The decision function is given as:

$$\begin{array}{ll} \min\limits_{f,b,\varepsilon} \frac{1}{2} f_H^2 + C \sum_i \varepsilon_i & \\ s.t. \quad y_i(f(x_i) + b) \geq 1 - \varepsilon_i\ \forall_i & , \\ \varepsilon_i \geq 0 \qquad\qquad\qquad \forall_i & \end{array} \tag{9}$$

where $||\mathrm{f}||_H$ denotes a kernel in Hilbert space related to a kernel K_m. The overall kernel can be divided into different kernels, and we replace $||\mathrm{f}||_H$ with $\sum_m ||\ \mathrm{f}_m\ ||_{HM}$ to obtain:

$$\begin{array}{ll} \min\limits_{f_m,b,\varepsilon,d} \frac{1}{2} \sum\limits_m ||f_m||_{HM}^2 + C \sum_i \varepsilon_i & \\ s.t. \quad y_i \sum\limits_m f_m(x_i) + y_i b \geq 1 - \varepsilon_i\ \forall_i & \\ \varepsilon_i \geq 0 \qquad\qquad\qquad \forall_i & , \\ \sum\limits_m d_m = 1,\ d_m \geq 0 \qquad \forall_m & \end{array} \tag{10}$$

Optimization matter is performed using the convex optimization mathematical algorithm (Rakotomamonjy et al., 2008). Using multiple kernels increases the decision the power of the decision function and also increases the prediction performance compared to using one single kernel. In the present study, SimpleMKL was deployed to integrate five different types of molecular data including gene expression, CNV, gene methylation, protein expression, and somatic mutation.

Considering the number of data types used in our study, five different kernels were independently built and further integrated into a generic model. Each kernel corresponds to each individual data type (gene expression, CNV, gene methylation, protein expression, and somatic mutation). The "Poly" (Eq. 11) polynomial base kernel with a search range of degrees of freedom d\{1 2 3\} (Seoane et al., 2014) and the "Gaussian" (Eq. 12) kernel with a search range of the parameter δ \{0.25 0.5 1 2 5 7 10 12 15 17 20\} (Zhang et al., 2016; Sun et al., 2018) were used as kernel types.

$$K(x_i, x_j) = (x_i^T x_j + 1)^d, \tag{11}$$

$$K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\delta^2}) \tag{12}$$

In summary, the SimpleMKL directly addressed a multiple kernel SVM optimization problem and greatly reduced computation costs when compared to the use of learning kernel combinations from individual kernels.

Evaluation

The dataset used in our study was randomly divided into learning and validating sets in order to assess the performance of the proposed method. For the learning set, we used mRMR to select the most optimal features and to determine the model through 10-fold cross-validation experiments. Then, the pre-trained MKL model and its optimal parameters were used to predict the validation set. Because the validation dataset was not used in the cross-validation process, the model derived from the learning dataset was tested on an independent validation dataset.

To assess the performance of our model, AUC, the most widespread evaluation metric for classification problems, was used to assess the performance of the proposed model. AUC is defined as the area under the receiver operating characteristic (ROC) curve, and it is used to quantify the overall performance of a classification model. Specifically, AUC = 1 denotes perfect performance, and 0.5 denotes random guessing. Pre (precision, Eq. 13), Sn (sensitivity, Eq. 14), Sp (Specificity, Eq. 15), and Acc (Accuracy, Eq. 16) were also employed in addition to AUC as classification performance metrics for breast cancer prognosis. The definitions of those metrics are provided below:

$$Pre = \frac{TP}{TP + FP}, \tag{13}$$

$$Sn = \frac{TP}{TP + FN}, \tag{14}$$

$$Sp = \frac{TN}{TN + FP}, \tag{15}$$

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{16}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

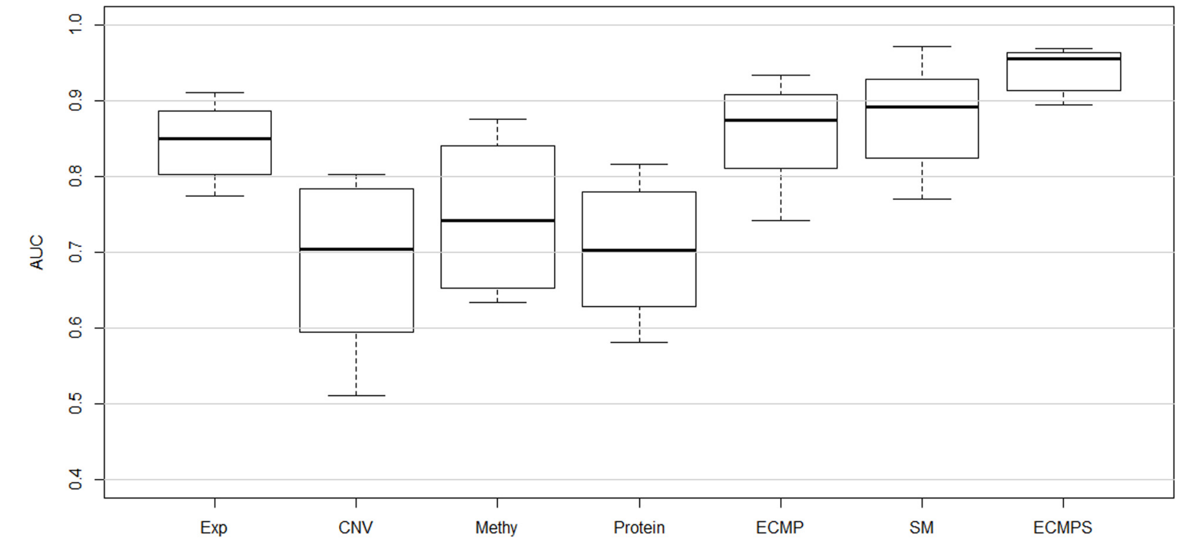


FIGURE 2 | Performance of classifying long-term and short-term survivors from a breast cancer dataset using different types of data based on the proposed hybrid combination of mRMR feature selection and MKL classification methods.

RESULTS

Comparison Studies on Learning Datasets

The proposed method was compared with other methods in three different applications: (1) comparison of the results of the models with different datasets based on the same method; (2) comparison of the results of different feature selection methods under the same datasets; and (3) comparison of the integration results of classification methods, under the same integrated datasets. AUC was used as an evaluation metric when comparing different methods and 10-fold cross-validation was applied for all methods.

Comparison of ECMPS and Other Data Types

Seven different MKL-based models were built using five single types of molecular data [gene expression (Exp), CNV, gene methylation (Methy), protein expression (Protein), and somatic mutations (SM)] and two integrated datasets with and without somatic mutation data in order to evaluate the role of somatic mutations in breast cancer survival prediction. The dataset integrating gene expression, CNV, gene methylation, and protein expression is abbreviated as "ECMP," and the dataset integrating all five molecular datasets including somatic mutations is denoted as "ECMPS."

The corresponding mean of the AUC value of 10-fold cross-validation (CVmean_AUC) for each of the seven models, using the mRMR feature selection and the MKL classification method, was calculated to compare the predictive performance of breast cancer survival models. The results are displayed in **Figure 2**, with the mean values of the boxplots corresponding to the red line in **Figure 3**. As shown in **Figure 3**, the ECMPS model consistently exhibited significantly more optimal performances than all the other models for all three feature selection methods. The two integrated models present obvious improvements compared to the single data type model results, suggesting that integrated models are more optimal than single data type ones, which is consistent with previous studies (Zhang et al., 2016; Sun et al., 2018).

In **Figure 2**, the mean value of the AUC for the multi-data ECMP model without somatic data is 0.8854, and the corresponding value for the ECMPS model increased to 0.9421 when incorporating somatic mutation. In addition, among the single data type models, the AUC of the somatic mutation model was higher than that of the model using the other four single data types and ECMP. Thus, our experimental results indicated that the somatic mutation data is able to increase the accuracy of the survival prediction for breast cancer patients.

The Pre, Sn, Sp, and Acc values for each dataset model were calculated in addition to the AUC based on the proposed method, and the results are presented in **Figure 4A**. **Figure 4A** shows that the integrative models combining different types of data, including somatic mutations, overcome the models using single data types for classification. The experimental results indicated that the proposed integrated model can successfully predict the survival time for breast cancer patients and somatic mutations can improve predictive accuracy.

Comparison of mRMR With Different Feature Selection Methods

We used mRMR to select the variables for each of the five types of molecular data. Then, the features with the largest relevance to the survival and lowest redundancy among themselves were selected, and they were combined as integrated features using the MKL classification model. The most optimal number of selected non-redundant features *k* for each molecular data type was determined by comparing the AUC values in the prediction results. According to the number of features reported in the

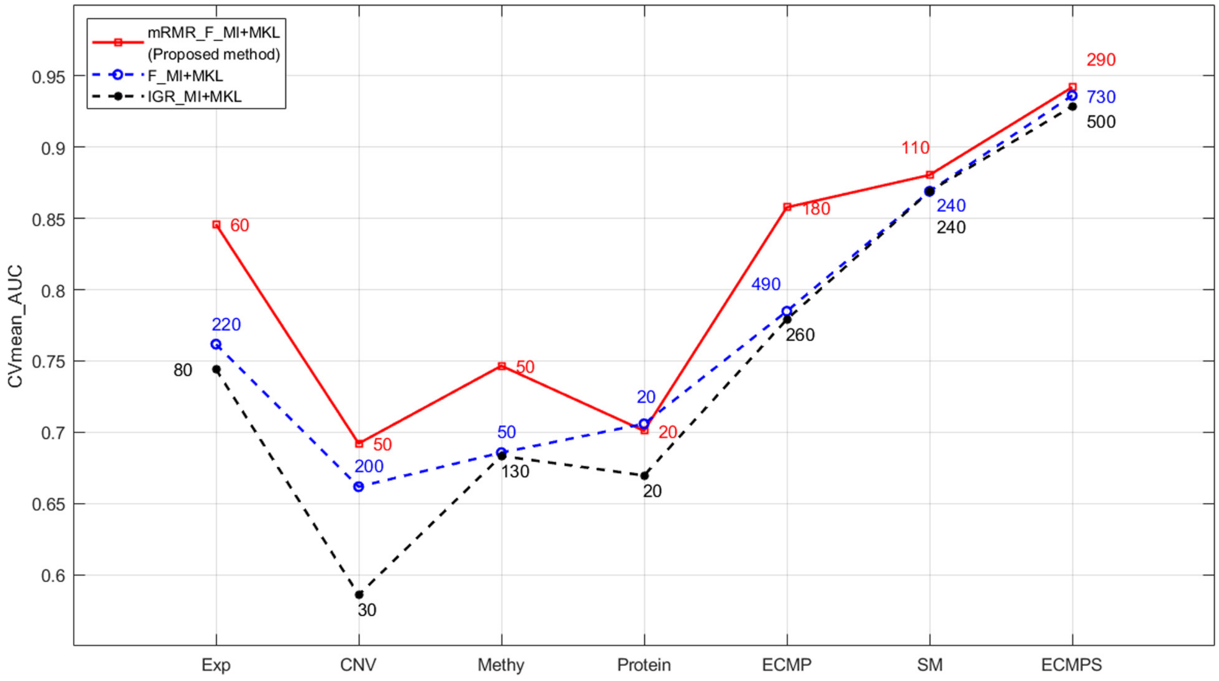


FIGURE 3 | Performance comparison of mRMR and the two k-best methods based on MKL under different data types. The numbers in different colors on the lines indicate the number of optimal features selected by the corresponding method.

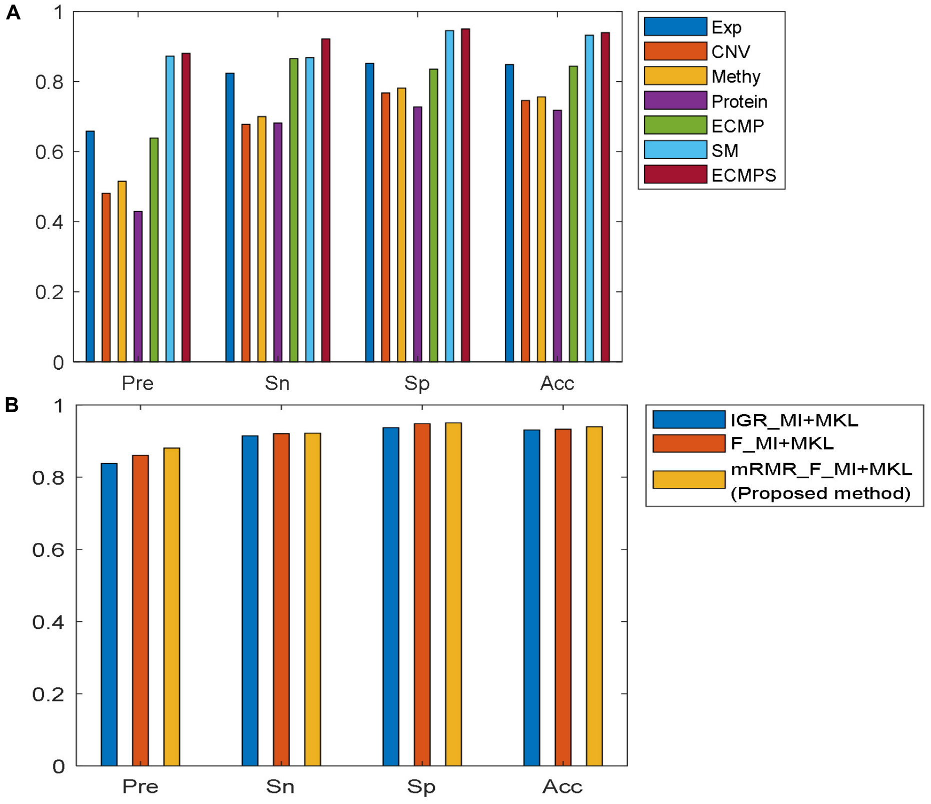


FIGURE 4 | Comparison of performances of the models using different evaluation metrics: Pre, Sn, Sp, and Acc. **(A)** Performance of the proposed method in seven datasets. **(B)** Performance of various feature selection methods based on MKL under the same data type "ECMPS."

previous study (Sun et al., 2018), we set k = [10, 20,. . .,300] in our work and chose the optimal parameter k as the final parameter for each data type in our study based on the prediction result.

The classification outcomes of the five data types under different parameters are presented in **Supplementary File S1**. The optimal feature number was selected based on the position of the maximum AUC value as the final parameter for a model of further integration. Take gene expression for example, as shown in **Figure 5**, the optimal number of features in the gene expression model using the proposed method is 60, which achieves the largest mean value of AUC with 10-fold cross-validation. Finally, we chose k = [60, 50, 50, 20, 110] as the optimal parameters for the five types of molecular data (Exp, CNV, Methy, Protein, and SM), respectively, for further integration analysis, and the total 290 features were obtained for our integrated ECMPS model.

The F-statistic (F) and PCC were used for the mRMR feature selection method to calculate the relevance and redundancy (Radovic et al., 2017), respectively, for four continuous data types, including Exp, CNV, Methy, and Protein, in order to maintain the original information for different types of data. MI was used to calculate both the relevance and redundancy of somatic mutation features, and is short for "mRMR_F_MI." In all cases, the selected features were integrated using MKL classification. To assess the performance of the mRMR feature selection method in the selection of features for our breast cancer survival prediction model, the proposed mRMR feature selection was compared with two commonly used k-best methods, which only consider relevance with the output, based on the same datasets and classification method MKL: (1) F-MI. Compared to the proposed method, it only uses the F-statistic and MI to select the most optimal k-best features for four continuous molecular datasets and discrete somatic mutation. (2) IGR-MI. It adopts a recently used feature selection method, the IGR (Sun et al., 2018), for four continuous molecular datasets and MI for discrete somatic mutation.

The proposed mRMR method outperformed both k-best feature selection methods F-statistics and IGR for four continuous molecular data types and their integration ECMP model according to the results shown in **Figure 3**. For instance, 260 features were selected by IGR based on the ECMP model and the AUC value was 0.7791, which was consistent with previous studies (Sun et al., 2018). Next, 180 features were selected using mRMR and AUC was 0.8578 showing that mRMR can achieve higher predictive accuracy using fewer features. The mRMR method also outperformed MI for discrete somatic mutation returning a smaller number of features. The most optimal result was obtained by mRMR and the total integration model ECMPS. The metrics Pre, Sn, Sp, and Acc were calculated in addition to the AUC for each dataset model, with a more optimal performance by mRMR as compared to the other the two k-best methods (**Figure 4B**). Our findings indicated that the use of proper feature selection methods is crucial to the classification process.

As the red line shows in **Figure 3**, for the integrated ECMPS model, 290 features were selected as more relevant to survival and non-redundant features in the integrated ECMPS mode consisting of 60 Exp, 50 CNV, 50 Methy, 20 Protein, and 110 SM using mRMR, and the most optimal AUC (0.9421) in the present study was achieved. Next, mRMR was applied again for the set of 290 features, which is

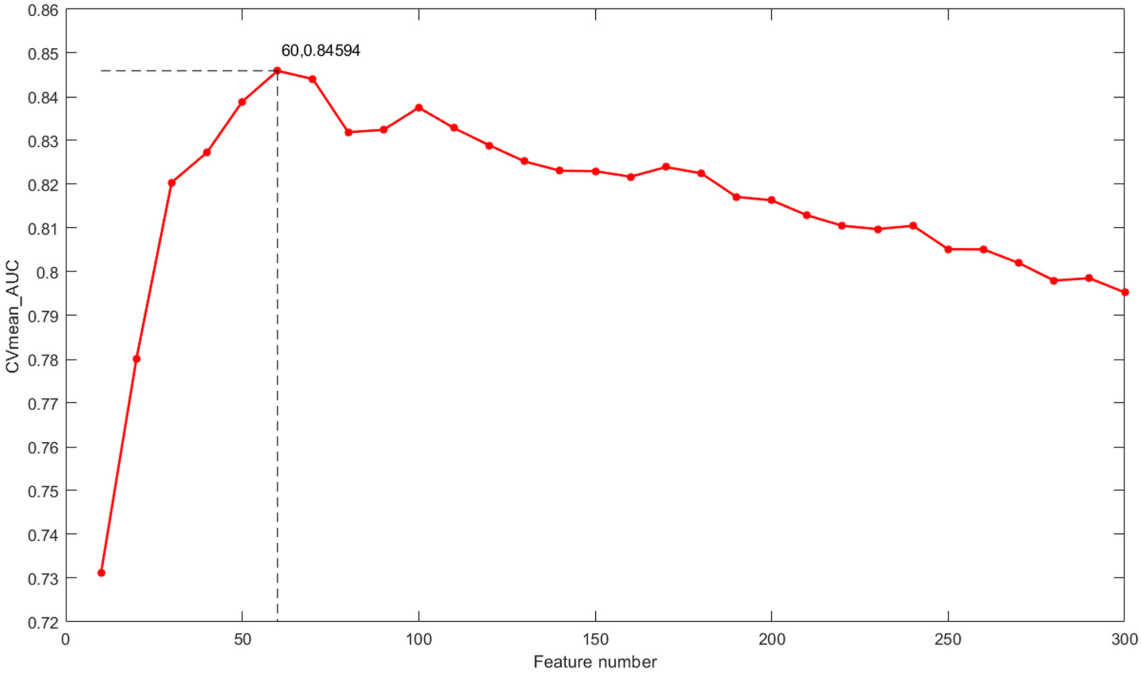


FIGURE 5 | The mean value of the AUC for 10-fold cross-validation (CVmean_AUC) under the feature numbers ranging from 10 to 300 for the model based on gene expression.

TABLE 3 | Comparison of mRMR and 2-mRMR on survival prediction power and feature numbers.

	AUC(ECMPS)	Number of features					
		ECMPS	Exp	CNV	Methy	Protein	SM
mRMR	0.9421 ± 0.0281	290	60	50	50	20	110
2-mRMR	0.9439 ± 0.0264	220	49	22	29	10	110

TABLE 4 | Comparative results of the proposed MKL method and existing traditional classifiers using AUC values under two mRMR selected integrated data models.

	ECMP	ECMPS
RF	0.7135 ± 0.054	0.7916 ± 0.027
SVM	0.8325 ± 0.037	0.9086 ± 0.058
MKL	0.8578 ± 0.049	0.9421 ± 0.028

termed as "2-mRMR," to resolve the redundancy that exists in the selected features of different data types. However, as shown in **Table 3**, the number of SM remained at 110 among the new 220 features, and the AUC value was only marginally improved. These results showed that there is a large internal redundancy within one type of data, while the redundancy between different types of data is small. It further indicated that the importance of somatic mutations to the prognosis is relatively stable. Finally, we retained the integrated 290 features originally selected by mRMR and used them for further classification, considering the stable high performance and simpler simple computational complexity of mRMR. We observed that mRMR outperformed k-best methods, and integrating somatic mutations achieved the most accurate prognosis.

Comparison of MKL With Traditional Classification Methods

The proposed method achieves a stronger performance by integrating somatic mutations compared with those methods incorporating single data types and integrated datasets without somatic mutations. The MKL classification method was compared with two widely used classifiers, SVM and RF, to further verify its ability to combine different types of data. Experiments were conducted in two integrated datasets: ECMP and ECMPS, which were selected by mRMR. The AUC value (mean value and standard error) was used to assess the performance of different methods and the results are provided in **Table 4**. **Table 4** shows that a more optimal performance was obtained from MKL for both integrated datasets compared to other classifiers, and this finding indicated the superiority of MKL in data integration.

In addition, the performances of all the classifiers were improved when employing ECMPS compared with ECMP, which further suggested that somatic mutations can provide adequate supplementary information for survival prediction of breast cancer. Finally, our method achieves the most optimal performance due to its ability to integrate multiple molecular data types, including somatic mutations, and MKL was quite efficient in integrating the data from distinct sources in breast cancer survival prediction.

Analysis of the Most Desirable Features From Somatic Mutation and Gene Expression Data

The top 10 features ranked by mRMR for each molecular data type were further analyzed by conducting a simple analysis on their association with breast cancer.

TABLE 5 | Genes previously associated with breast cancer.

Genes	Reports	References
HCN4	HCN4 was highly correlated with lower survival rates of breast cancer.	Phan et al., 2017
RGPD3	30 most enriched new HOXB7 binding sites on breast cancer cell chromatin for which an annotated nearest gene exists: RGPD3, PIK3R1, etc.	Heinonen et al., 2015
EFCAB13	Variants that induce premature stop codons were identified in the DENND2D, EFCAB13, and TICRR genes.	Määttä et al., 2016
NFATC1	NFATC1 overexpression results in oncogenic BMI1transcriptional upregulation. Co-expression of FUNDC1 and BMI1 in BC patients predicted worse prognosis.	Wu et al., 2019
VAC14	VAC14 selectively prevents rapid degradation of Sac3.	Ikonomov et al., 2013
PRB2	A novel six-gene (TMEM252, PRB2, SMCO1, IVL, SMR3B, and COL9A3) signature was significantly associated with prognosis as an independent prognostic signature.	Lv et al., 2019
HIPK1	The deletion of the miR-200c/141 cluster resulted in increased tumor metastasis and inhibited tumor growth by directly upregulating the target gene HIPK1.	Liu et al., 2018
IRF2	Interferon regulatory factor 1 (IRF-1) and IRF-2 expression in breast cancer tissue microarrays.	Connett et al., 2005
HMGB2	Promotion of breast cancer progression by HMGB2.	Fu et al., 2018
FRMPD1	Rat Mcs5a is associated with breast cancer risk. Mcs5a1 is located within the ubiquitin ligase Fbxo10, whereas Mcs5a2 includes the 5′ portion of FRMPD1.	Samuelson et al., 2007
RPS27	The best ranked cancer immunotherapy proteins related to BC were RPS27, SUPT4H1, and CLPSL2.	López-Cortés et al., 2020
PTPRR	PTPRR and myocyte enhancer factor 2C (MEF2C) genes were upregulated in the classical MAPK and p38 MAPK pathways.	Motaghed et al., 2014

Only features from somatic mutations and gene expression datasets were explored to further assess the effectiveness of our method. The results of this analysis showed that it was previously reported that some of the genes are associated with breast cancer survival. These genes and their references are listed in **Table 5**. It has previously been reported in the literature that seven of the top 10 ranked gene names from the somatic mutation features play critical roles in breast cancer prognosis. For example, the HCN4 gene is highly correlated with lower survival rates of breast cancer (Phan et al., 2017), and the gene PRB2 is significantly related to prognosis as an independent prognostic marker (Lv et al., 2019). On the other hand, five of the top 10 genes selected from gene expression datasets have also been found to be associated with breast cancer. For instance, the expression of IRF2 has been found to be related to breast cancer (Connett et al., 2005), and it has been reported that HMGB2 directly and significantly promotes breast cancer progression (Fu et al., 2018). Thus, the top ranked features were shown to be important for breast cancer prognosis.

Validation

Optimization techniques have been previously applied (Zhang et al., 2016; Zhang A. et al., 2019) to select the most optimal feature subsets in a wrapper feature selection framework. Therefore, experiments were performed on an independent validation dataset to further evaluate our proposed method. Our model was initially trained and tested on a learning dataset containing 390 breast cancer patients, and then, to predict patient survival, it was applied to a 98-patient validation dataset that was not involved in training or testing. The survival of most of the 98 breast cancer patients was correctly classified, and the accuracy of the proposed method on the validation dataset was 0.9808.

DISCUSSION

We integrated somatic mutations and previously used data types, including Exp, CNV, Methy, and protein, using MKL to predict breast cancer patient survival. Applying mRMR-selected features and MKL classification, we found that the integration of somatic mutations enriched the diversity of features and was conducive to the improvement of the prediction model. In all, integrating promising data sources such as somatic mutations and harnessing the powerful feature selection method mRMR and the effective data fusion method MKL can increase the prediction accuracy of breast cancer patient survival.

Although our method is effective and can accurately predict the survival of breast cancer patients, some limitations remain in the prognosis of breast cancer. For instance, there may be more effective methods that can be used to construct kernels for an improved multi-kernel learning method in the future that will further improve the performance in multi-omics data fusing. In addition, our available sample size was limited by the intersection of multiple types of molecular data samples. Thus, the performance of our method could be promoted when a larger population of samples becomes available in the future. Furthermore, somatic mutations are highly heterogeneous among patients, and therefore, further understanding of the mechanism of somatic mutation in cancer may lead to a more accurate prognostic model for breast cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ZH and JZ participated in the design of algorithms and experiments and participated in the design of the whole framework of prediction of breast cancer survival. JZ directed the whole work. YZ participated in the analysis of the performance of the proposed method. JZ and XY conceived of the study and helped edit the manuscript. All authors read the final manuscript and approved the submission.

FUNDING

This research was funded by the Natural Science Foundation of China under Grants 61571341 and 61902430 and the Fundamental Research Funds for the Central Universities under Grant JB190302.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.632901/full#supplementary-material

REFERENCES

Arslanturk, S., Draghici, S., and Nguyen, T. (2020). Integrated Cancer subtyping using heterogeneous genome-scale molecular datasets. *Pac. Symp. Biocomput.* 25, 551–562.

Brennan, C. W., Verhaak, R. G. W., Mckenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2014). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477.

Cancer Genome Atlas Research Network (2013). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 494, 506–506. doi: 10.1038/nature11903

Chen, Q., Lai, D., He, L., Yan, Y., Li, E., Liu, Y., et al. (2019). "ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion," in *Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Piscataway, NJ: IEEE.

Chen, Y., Sun, J., Huang, L.-C., Xu, H., and Zhao, Z. (2015). Classification of cancer primary sites using machine learning and somatic mutations. *Biomed. Res. Int.* 2015, 1–9. doi: 10.1155/2015/491502

Connett, J. M., Badri, L., Giordano, T. J., Connett, W. C., and Doherty, G. M. (2005). Interferon regulatory factor 1 (IRF-1) and IRF-2 expression in breast cancer tissue microarrays. *J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon Cytokine Res.* 25, 587–594. doi: 10.1089/jir.2005.25.587

Dey, S., Gupta, R., Steinbach, M., and Kumar, V. (1990). *Integration of Clinical and Genomic Data: A Methodological Survey*. Minneapolis, MN: University of Minnesota Digital Conservancy.

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.1142/s0219720005001004

Ferlay, J., Héry, C., Autier, P., and Sankaranarayanan, R. (2010). *Global Burden of Breast Cancer*. New York, NY: Springer.

Fu, D., Li, J., Wei, J., Zhang, Z., Luo, Y., Tan, H., et al. (2018). HMGB2 is associated with malignancy and regulates Warburg effect by targeting LDHB and FBP1 in breast cancer. *Cell Commun. Signal.* 16:8.

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., and Moor, A. B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22, e184–e190.

Gönen, M., and Alpaydin, E. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* 12, 2211–2268.

Griffith, O. L., Spies, N. C., Anurag, M., Griffith, M., Luo, J., Tu, D., et al. (2018). The prognostic effects of somatic mutations in ER-positive breast cancer. *Nat. Commun.* 9:3476.

Haricharan, S., Bainbridge, M. N., Scheet, P., and Brown, P. H. (2014). Somatic mutation load of estrogen receptor-positive breast tumors predicts overall survival: an analysis of genome sequence data. *Breast Cancer Res. Treat.* 146, 211–220. doi: 10.1007/s10549-014-2991-x

He, Z., Zhang, J., Yuan, X., Liu, Z., Liu, B., Tuo, S., et al. (2017). Network based stratification of major cancers by integrating somatic mutation and gene expression data. *PLoS One* 12:e0177662. doi: 10.1371/journal.pone.0177662

He, Z., Zhang, J., Yuan, X., Xi, J., Liu, Z., and Zhang, Y. (2019). Stratification of breast cancer by integrating gene expression data and clinical variables. *Molecules* 24:631. doi: 10.3390/molecules24030631

Heinonen, H., Lepikhova, T., Sahu, B., Pehkonen, H., Pihlajamaa, P. I., Louhimo, R., et al. (2015). Identification of several potential chromatin binding sites of HOXB7 and its downstream target genes in breast cancer. *Int. J. Cancer J. Int. Cancer* 137, 2374–2383. doi: 10.1002/ijc.29616

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Ikonomov, O. C., Filios, C., Sbrissa, D., Chen, X., and Shisheva, A. (2013). The PIKfyve-ArPIKfyve-Sac3 triad in human breast cancer: functional link between elevated Sac3 phosphatase and enhanced proliferation of triple negative cell lines. *Other* 440, 342–347. doi: 10.1016/j.bbrc.2013.09.080

Jain, A., and Zongker, D. (1997). Feature selection: evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 153–158. doi: 10.1109/34.574797

Jie, T., Hammond, J. H., Hogan, D. A., and Greene, C. S. (2015). ADAGE analysis of publicly available gene expression data collections illuminates *Pseudomonas aeruginosa*-host interactions. *mSystems* 1:e00025-15. doi: 10.1128/mSystems.00025-15

Khademi, M., and Nedialkov, N. S. (2016). "Probabilistic graphical models and deep belief networks for prognosis of breast cancer," in *Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL.

Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. (2011). lp-norm multiple kernel learning. *J. Mach. Learn. Res.* 12, 953–997.

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). "LDICDL: LncRNA-disease association identification based on collaborative deep learning," in *Proceedings of the IEEE/ACM Trans Comput Biol Bioinform*, Piscataway, NJ: IEEE.

Lanckriet, G., Cristianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.* 5, 27–72.

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Liu, B., Du, R., Zhou, L., Xu, J., Chen, S., Chen, J., et al. (2018). miR-200c/141 regulates breast cancer stem cell heterogeneity via Targeting HIPK1/β-Catenin Axis. *Theranostics* 8, 5801–5813. doi: 10.7150/thno.29380

López-Cortés, A. L., Cabrera-Andrade, A., ázquez-Naya, J. M. V., Pazos, A., and Munteanu, C. R. (2020). Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks. *Entific Rep.* 10:8515.

Lv, X., He, M., Zhao, Y., Zhang, L., and Wei, M. (2019). Identification of potential key genes and pathways predicting pathogenesis and prognosis for triple-negative breast cancer. *Cancer Cell Int.* 19:172.

Määttä, K., Rantapero, T., Lindström, A., Nykter, M., Kankuri-Tammilehto, M., Laasanen, S. L., et al. (2016). Whole-exome sequencing of Finnish hereditary breast cancer families. *Eur. J. Hum. Genet. Ejhg* 25, 85–93. doi: 10.1038/ejhg.2016.141

Mary, G., Brian, C., Teresa, S., Melissa, C., Olena, M., Mark, D., et al. (2014). The UCSC cancer genomics browser: update 2015. *Nucleic Acids Res.* 43, D812–D817.

Mermel, C. H., Schumacher, S. E., Hill, B., and Meyerson, M. L. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12, R41–R41.

Motaghed, M., Al-Hassan, F. M., and Hamid, S. S. (2014). Thymoquinone regulates gene expression levels in the estrogen metabolic and interferon pathways in MCF7 breast cancer cells. *Int. J. Mol. Med.* 33, 8–16. doi: 10.3892/ijmm.2013.1563

Nguyen, C., Yong, W., and Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Ence Eng.* 06, 551–560. doi: 10.4236/jbise.2013.65070

Phan, N. N., Huynh, T. T., and Lin, Y. C. (2017). Hyperpolarization-activated cyclic nucleotide-gated gene signatures and poor clinical outcome of cancer patient. *Transl. Cancer Res.* 6, 698–708. doi: 10.21037/tcr.2017.07.22

Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* 18:9. doi: 10.1186/s12859-016-1423-9

Rakotomamonjy, A., Bach, F. R., Canu, S., and Grandvalet, Y. (2008). SimpleMKL. *J. Mach. Learn. Res.* 9, 2491–2521.

Ronen, J., Hayat, S., and Akalin, A. (2018). Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Sci. Alliance* 2:e201900517. doi: 10.26508/lsa.201900517

Samuelson, D. J., Hesselson, S. E., Aperavich, B. A., Zan, Y., Haag, J. D., Trentham-Dietz, A., et al. (2007). Rat Mcs5a is a compound quantitative trait locus with orthologous human loci that associate with breast cancer risk. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6299–6304. doi: 10.1073/pnas.0701687104

Seoane, J. A., Day, I. N. M., Gaunt, T. R., and Colin, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics* 30, 838–845. doi: 10.1093/bioinformatics/btt610

Sun, D., Li, A., Tang, B., and Wang, M. (2018). Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Comput. Methods Progr. Biomed.* 161, 45–53. doi: 10.1016/j.cmpb.2018.04.008

Wu, L., Zhang, D., Zhou, L., Pei, Y., Zhuang, Y., Cui, W., et al. (2019). FUN14 domain-containing 1 promotes breast cancer proliferation and migration by activating calcium-NFATC1-BMI1 axis. *Ebiomedicine* 41, 384–394. doi: 10.1016/j.ebiom.2019.02.032

Xu, X., Huang, L., Chan, C. H., Yu, T., Miao, R., and Liu, C. (2016). Assessing the clinical utility of genomic expression data across human cancers. *Oncotarget* 7, 45926–45936. doi: 10.18632/oncotarget.10002

Xu, X., Zhang, Y., Liang, Z., Wang, M., and Ao, L. (2013). "A gene signature for breast cancer prognosis using support vector machine, biomedical engineering and informatics (BMEI)," in *Proceedings of the 2012 5th International Conference on BioMedical Engineering and Informatics*, Chongqing: IEEE.

Yan, F., Kittler, J., Mikolajczyk, K., and Tahir, M. A. (2009). "Non-sparse multiple kernel learning for fisher discriminant analysis," in *Proceedings of the IEEE International Conference on ICDM*, Miami, FL.

Ye, Z. L., Guan, W. L., Tang, T., Wang, F., and He, C. Y. (2019). Gene mutation profiling in chinese colorectal cancers patients and its association with clinicopathological characteristics and prognosis. *Ssrn Electron. J.* 9, 745–756. doi: 10.1002/cam4.2727

Yuan, X., Bai, J., Zhang, J., Yang, L., Duan, J., Li, Y., et al. (2020a). CONDEL: detecting copy number variation and genotyping deletion zygosity from single tumor samples using sequence data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1141–1153.

Yuan, X., Gao, M., Bai, J., and Duan, J. (2020b). SVSR: a program to simulate structural variations and generate sequencing reads for multiple platforms. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1082–1091. doi: 10.1109/tcbb.2018.2876527

Yuan, X., Yu, J., Xi, J., Yang, L., Shang, J., Li, Z., et al. (2019). "CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data," in *Proceedings of the IEEE/ACM Trans Comput Biol Bioinform*, Piscataway, NJ: IEEE.

Yuan, X., Zhang, J., and Yang, L. (2017). IntSIM: an integrated simulator of next-generation sequencing data. *IEEE Trans. Biomed. Eng.* 64, 441–451. doi: 10.1109/tbme.2016.2560939

Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* 32, 644–652. doi: 10.1038/nbt.2940

Zhang, A., Li, A., He, J., and Wang, M. (2019). LSCDFS-MKL: a multiple kernel based method for lung squamous cell carcinomas disease-free survival prediction with pathological and genomic data. *J. Biomed. Inform.* 94:103194. doi: 10.1016/j.jbi.2019.103194

Zhang, Y., Li, A., He, J., Wang, M., and Novel, A. (2019). MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data. *IEEE J. Biomed. Health Inform.* 24, 171–179. doi: 10.1109/jbhi.2019.2898471

Zhang, Y., Li, A., Peng, C., and Wang, M. (2016). Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13, 825–835. doi: 10.1109/tcbb.2016.2551745

Zhang, Y., Yang, W., Dan, L., Yang, J. Y., Guan, R., and Yang, M. Q. (2018). Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. *BMC Med. Genomics* 11(Suppl. 5):104. doi: 10.1186/s12920-018-0419-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 January 2021
doi: 10.3389/fgene.2020.608512

Hierarchical Microbial Functions Prediction by Graph Aggregated Embedding

Yujie Hou[1,2†], Xiong Zhang[1,3†], Qinyan Zhou[1,4†], Wenxing Hong[1,5*] and Ying Wang[1,5,6*]

[1] Department of Automation, Xiamen University, Xiamen, China, [2] Department of Automation, University of Science and Technology of China, Hefei, China, [3] School of Automation Science and Engineering, South China University of Technology, Guangzhou, China, [4] Institute of AI and Robotics, Fudan University, Shanghai, China, [5] Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision, Xiamen, China, [6] Fujian Key Laboratory of Genetics and Breeding of Marine Organisms, Xiamen, China

Matching 16S rRNA gene sequencing data to a metabolic reference database is a meaningful way to predict the metabolic function of bacteria and archaea, bringing greater insight to the working of the microbial community. However, some operational taxonomy units (OTUs) cannot be functionally profiled, especially for microbial communities from non-human samples cultured in defective media. Therefore, we herein report the development of Hierarchical micrObial functions Prediction by graph aggregated Embedding (HOPE), which utilizes co-occurring patterns and nucleotide sequences to predict microbial functions. HOPE integrates topological structures of microbial co-occurrence networks with *k*-mer compositions of OTU sequences and embeds them into a lower-dimensional continuous latent space, while maximally preserving topological relationships among OTUs. The high imbalance among KEGG Orthology (KO) functions of microbes is recognized in our framework that usually yields poor performance. A hierarchical multitask learning module is used in HOPE to alleviate the challenge brought by the long-tailed distribution among classes. To test the performance of HOPE, we compare it with HOPE-one, HOPE-seq, and GraphSAGE, respectively, in three microbial metagenomic 16s rRNA sequencing datasets, including abalone gut, human gut, and gut of *Penaeus monodon*. Experiments demonstrate that HOPE outperforms baselines on almost all indexes in all experiments. Furthermore, HOPE reveals significant generalization ability. HOPE's basic idea is suitable for other related scenarios, such as the prediction of gene function based on gene co-expression networks. The source code of HOPE is freely available at https://github.com/adrift00/HOPE.

Keywords: microbial co-occurrence networks, functions prediction, graph embedding, hierarchical multi task learning, deep learning

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Lin Wan,
Academy of Mathematics and Systems Science (CAS), China
Kai Song,
Qingdao University, China

****Correspondence:***
Ying Wang
wangying@xmu.edu.cn
Wenxing Hong
hwx@xmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *21 September 2020*
Accepted: *20 November 2020*
Published: *18 January 2021*

Citation:
Hou Y, Zhang X, Zhou Q, Hong W and Wang Y (2021) Hierarchical Microbial Functions Prediction by Graph Aggregated Embedding. Front. Genet. 11:608512. doi: 10.3389/fgene.2020.608512

INTRODUCTION

The analysis of microbial communities is founded on the characterization of functional diversity, which is increasingly recognized as the bridge linking biodiversity patterns and ecosystem functioning, as a way of explaining the interactions between microbes and their responses to changes in the environment (Bardgett and Der Putten, 2014; Escalas et al., 2019). However, a large proportion of microbes remain uncultivated and, therefore, functionally unknown. However, because of the prevalence of high-throughput sequencing technologies, large-scale 16S rRNA marker gene sequencing of microbes is becoming available. Related approaches, such as PICRUSt (Langille et al., 2013) and Tax4Fun (Ashauer et al., 2015), are proposed to infer functional profiles from genomes and phylogeny. PICRUSt and Tax4Fun identify microbial functions by estimating 16s rRNA marker gene families based on the similarity between 16s rRNA sequencing data and known marker gene databases. They rely on the reference databases on Greengenes (Desantis et al., 2006) and SILVA (Quast et al., 2012). However, owing to the incompleteness of the 16S rRNA marker gene database, large amounts of OTUs cannot be functionally profiled, especially for microbial communities from non-human samples from defective culture media (Pachiadaki et al., 2019; Wang X. et al., 2020).

The protein–protein interaction (PPI) network in protein function prediction gives theoretical insight to our study of microbial functional diversity. More specifically, network representation of the PPI network extracts functional context from topological structure (Gligorijevic et al., 2018; Kulmanov et al., 2018) and achieves better performance than the previous algorithm that only uses sequence data (Wass et al., 2012; Cozzetto et al., 2013). Several researchers found that proteins with interactions in PPI networks have a high possibility of sharing the same or similar functions (Lele et al., 2011; Liu et al., 2017). Inspired by empirical success in using the PPI network, we can build a microbial co-occurrence network to provide new insights into the exploration of microbial functions. Microorganisms do not live in isolation but, rather, interact with the environment through, for example, mutualism, competition, parasitism, and predation. "Co-occurrence" means that microbes have statistically significant associations of abundance in one microbial community. The co-occurrence relationship is generally inferred by abundance correlation over several microbial community samples. The microbial co-occurrence network was designed to describe these relationships among microbes, and those microbes with closely correlated relationships become linked in the microbial co-occurrence network.

A novel method, Hierarchical micrObial functions Prediction by graph aggregated Embedding (HOPE), was proposed to capture potential functions in a microbial co-occurrence network. Our method is built based on the key hypothesis that microbes with co-occurring patterns have a high possibility of sharing the same or similar functions. So, our method tries to use this property to infer unknown microbe functions from its neighbors in the microbial co-occurrence network. HOPE has two main modules: hierarchical multitask learning and graph embedding. Here, the hierarchical multitask learning framework solves the class imbalance problem, and the graph embedding learns the co-occurrence patterns in microbial networks. Two classic strategies have traditionally been performed: resampling (Chawla et al., 2002) and cost-sensitive reweighting (Khan et al., 2018) during our previous experiments. These methods change the training dataset distribution by either undersampling the majority class, oversampling the minority class, or giving a higher cost to misclassification of the minority class. However, neither of these classic methods could ameliorate the negative impact of imbalanced classes during our experiments. Both the majority class and the minority class can be well-classified if they are trained independently; therefore, we were motivated to design a hierarchical multitask training scheme to manage the imbalance of functional datasets with the long-tailed distribution. To accomplish this, we input two graphs with the majority class and the minority class, respectively, into the HOPE algorithm and train the model by multitask learning. A graph embedding model is designed to map the microbial co-occurrence network to a lower-dimensional continuous latent space while maximally preserving the topological relationships among OTU features. HOPE incorporates *k*-mer compositions of microbial sequences and topology of microbial networks, as complementary data sources, to learn an embedding representation of a microbial network. The embedded low-dimensional numerical vector of each OTU node reflects its sequencing features and co-occurrence correlation with its neighbors. After that, the multilayer perceptron (MLP) classifier takes embedding vectors as inputs to predict the function for those OTUs without functional information from the known database.

Cross-validation was designed to evaluate the performance of HOPE on three microbial metagenomic 16s rRNA sequencing datasets from abalone gut, human gut, and gut of *Penaeus monodon*, respectively, and all experiments mentioned above verified the superiority of HOPE in predicting microbial functions. HOPE is compared with its two variants, HOPE-seq and HOPE-one, as well as a well-known graph embedding algorithm, GraphSAGE (Hamilton et al., 2017), in the experiments. HOPE-seq uses only *k*-mer frequency vectors as features with the hierarchical multitask learning framework to train the classifier on majority classes and minority classes. HOPE-one ignores the hierarchical multitask learning but integrates the sequence representation with microbial network topological structure as embedding features for function prediction. In the testing set, we learned that HOPE outperforms HOPE-seq and HOPE-one on almost every measurement. HOPE outperforms HOPE-one by 9.5% in Micro-F1 on the Abalone Gut Microbiota dataset and 15.6% in Macro-F1 on the *P. monodon* intestine dataset. When compared with GraphSAGE, using three different aggregator functions, including a mean aggregator, an LSTM aggregator, and a pooling aggregator, HOPE achieves the highest score in most measurements with a significant margin. Compared with GraphSAGE, HOPE gains a higher accuracy score by 4.4% averagely. Finally, our results show that HOPE demonstrates significant generalization ability since it can be used to predict microbial functions without learning previous information in our experiments.

METHODS

Framework of Microbial Function Prediction With HOPE

HOPE consists of four steps to predict microbial functions, including data input, microbial co-occurrence network construction, graph embedding generation, and function prediction. The input are 16s rRNA sequence reads containing all microbial community information clustered to OTUs for building the microbial co-occurrence network based on the co-occurrence correlation relationships among OTUs. During the graph embedding step, HOPE learns the embedding vectors of the majority class and the minority class with multitask learning to mitigate class imbalance. The HOPE algorithm could distill the high-dimensional information about OTUs and their "neighbor OTUs" and embed the resultant data on topological structure into dense representative vectors. In this way, the original microbial network is converted to a compact embedding space, while a given node's features and the topological structure of its "neighborhood" are preserved. Finally, our approach uses the low-dimensional embedding vectors to identify microbial functions via an MLP classifier. The total pipeline of the framework is shown in **Figure 1**.

Data Preprocessing

Construction of Microbial Co-occurrence Network

Before microbial function prediction can take place, the raw data coming from 16s rRNA sequencing data will be input to the framework, which may have millions of reads and cause numerous computations. The 16s rRNA sequences are grouped into OTU bins based on the sequence alignment similarity, which is a step that can reduce the number of OTUs for faster calculation. During the experiments, sequences are clustered to OTUs satisfying the following criteria via the UPARSE-OTU algorithm (Edgar, 2013). The sequences in the same cluster (OTU) should have more than 97% pairwise sequence alignment similarity, and the sequences in a different cluster (OTU) should have more than 3% pairwise sequence alignment dissimilarity. The OTU representative sequence is the most abundant contig in the OTU cluster and is selected to represent the cluster for the following processing. Then, the "co-occurrence" patterns, which were revealed as the co-occurrence interaction of two species or any taxonomically relevant units in habitats, are calculated via the OTU table and the correlation algorithm. The OTU table describes the abundances of OTUs in samples by the USEARCH algorithm (Edgar, 2010), and the correlation score is computed for each OTU pair by the SparCC algorithm (Friedman and Alm, 2012). OTUs with a higher correlation score than the threshold are considered proof of having a strong co-occurrence correlation, and these OTUs will be connected with an edge in the microbial network. The microbial co-occurrence network is constructed to preserve the interaction patterns where each node represents an OTU, and each edge represents a pairwise association between them and the pipeline as shown in **Figure 1B**.

The OTU representative sequences offer sequence signatures and potential information about their functions. *K*-mer means nucleotide sequences of length *k*. The *k*-mer frequency is the number of occurrences of *k*-mer within the whole sequence(s) normalized by the total number of occurrences in the vector for each data. The *k*-mers frequency is adopted as OTU features, whose statistical distribution of frequency reflects the sequence signatures. The short sequence representation, *k*-mers frequency, further reduces calculation and reflects the compositional distribution of DNA sequence(s). Previous studies have shown that *k*-tuple frequencies are similar across different regions of the same genome but differ between genomes (Karlin et al., 1997), which offers the theoretical basis to measure the dissimilarity between contigs. The length of k has a significant impact on the final results. When $k \geq 20$ bp (long *k*-mer), *k*-mer reflects more detail and local biological information in the nucleotide sequences, but the high sparsity of the frequency vector lead by too long *k*-mer would lose the statistical power (Wang et al., 2014, 2018; Wang Y. et al., 2020). However, when $k \leq 10$ bp (short *k*-mer), the frequency of *k*-mers reflects the global compositional distribution of the whole sequences (Ren et al., 2016). In our study, the representative sequence of each OTU is $\sim 10^3$ bp; generally, *k* should be set from 4 to 10 (Wang et al., 2014). After testing on the different length of *k*-mer, the *k*-mer length of 7–10 has no much impact on performance. Therefore, we select $k = 7$ to reduce the running time of *k*-mer counting.

Function Labeling in Co-occurrence Network

As supervised learning, the labels of OTUs in the training set and the validation set need to be annotated. The multiclass classification means that there are more than two classes in the classification problem, and in our study, existence of multiple KEGG Orthology (KO) functions means multiple classes (Kanehisa and Goto, 2000). Multilabel means that a sample might belong to multiple classes, and in our study, there would be multiple KO functions for one OTU. Therefore, the function prediction task is formulated as a multiclass, multilabel classification problem. The label vectors containing the ground truth of OTU's functions utilize multihot encoding. This encoding approach could convert the useful information into a binary string with a single bit value of 1 or 0. If the OTU is annotated on the K00001 and K00003 function, then we will assign 1 to the first position and the third position in the binary string as a positive sample for this function. Every unique function category is represented as a binary value at a specific position in the labeled vector.

Working Principle of HOPE

The HOPE algorithm includes two critical modules, a hierarchical multitask learning scheme and a graph embedding module (**Figure 2**). In the hierarchical multitask learning part, the HOPE model is trained on the majority class and the minority class, respectively, wherein the majority class means this class exists in more than half of OTUs, and the minority class only appears in less than half of OTUs. Then the graph embedding module learns embedding vectors of OTUs by propagating nodes' neighbor feature to the nodes along the edges and aggregating the topological structure of nodes' neighborhood with the *k*-mer representation of OTUs, along with the microbial co-occurrence network.

Hierarchical Multitask Learning Scheme

During the learning of embedded representation of nodes, the highly skewed distribution of the functional class is observed

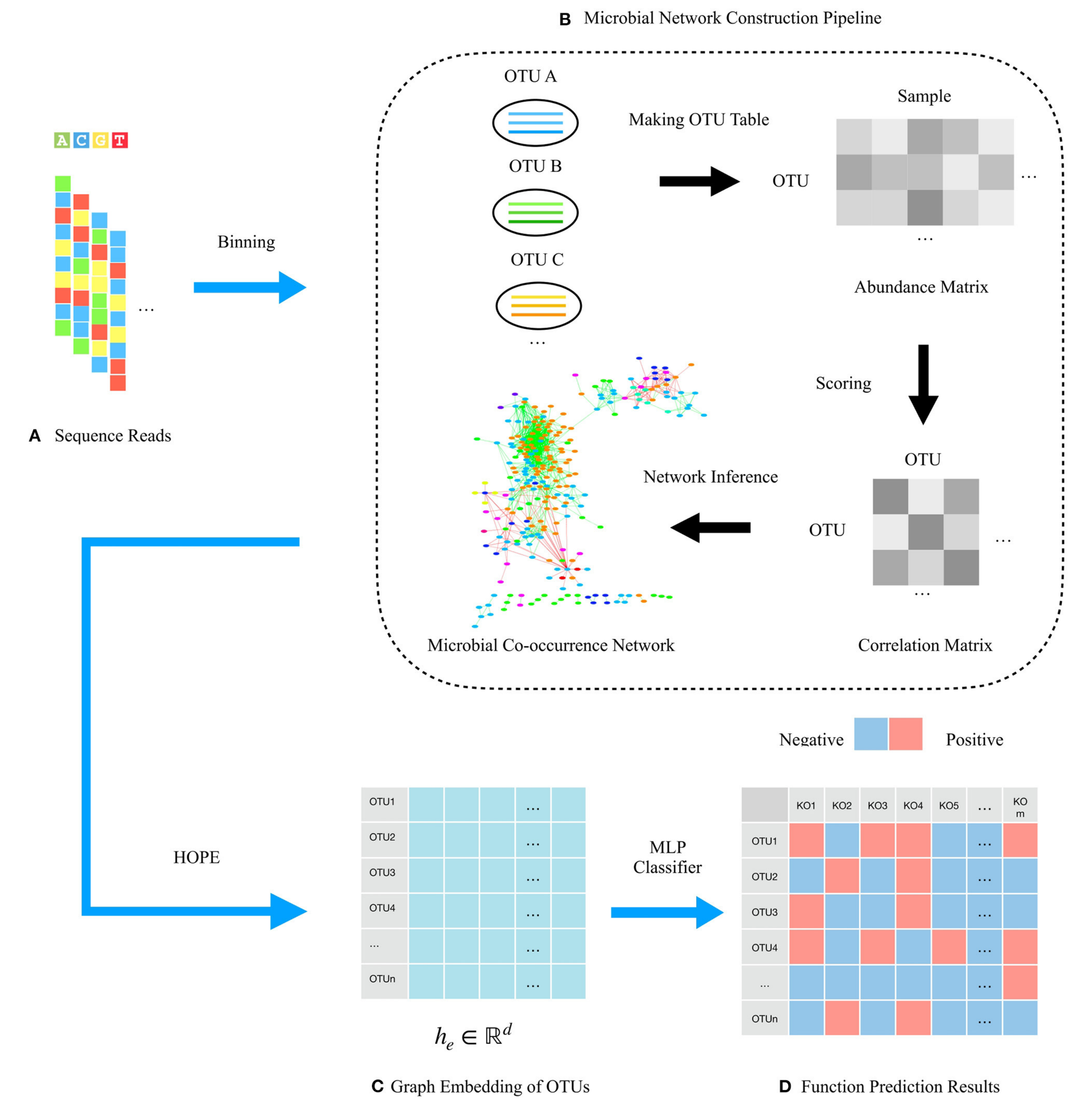


FIGURE 1 | Schematic illustration of the framework for predicting microbial functions using HOPE. **(A)** 16s rRNA sequencing reads from a microbial community are adopted for network construction. **(B)** Pipeline for constructing microbial networks. OTUs are binned by clustering reads from the same source population. Then, the abundance matrix that describes the relative abundance of OTUs in every microbiota sample is calculated. Pairwise scores between OTUs are then computed gaining the correlation matrix, and OTU pairs with correlation score over the threshold are connected by an edge. Gray areas in the correlation matrix indicate similarity of OTUs. Finally, the whole microbial community is visualized as a network wherein nodes represent OTUs, and edges represent the correlation between them. **(C)** Embedding representations of each OTU via the HOPE algorithm. **(D)** Function prediction matrix of OTUs. Different colors indicate different KO functions.

(**Figure 3**), which will cause a class imbalanced problem. Long-tailed and skewed distributions among different functions cause the classifier to ignore the minority classes (Huang et al., 2016). The majority class will influence the classifier to be biased toward the majority class so that the minority class will be overwhelming, wherein the majority class means this class exists

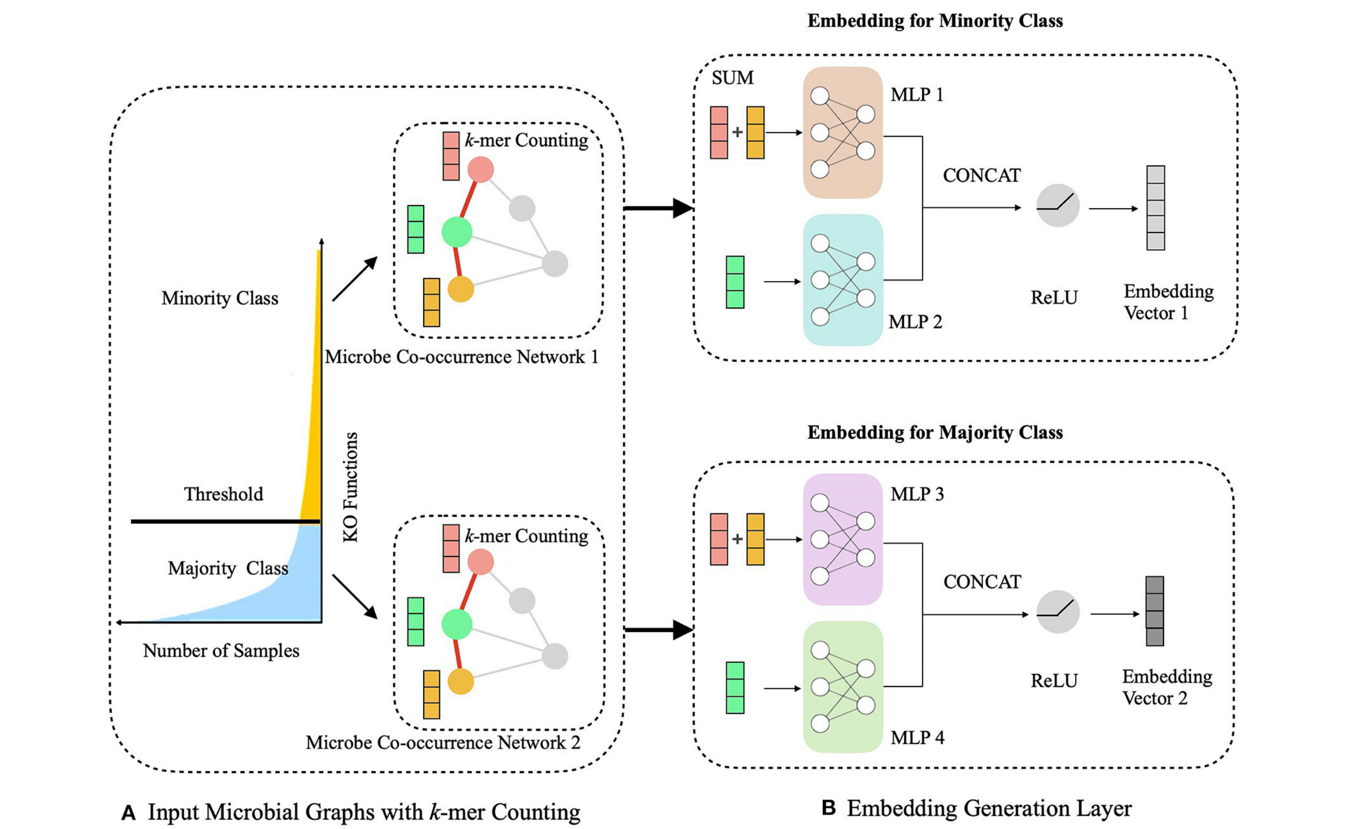

Embedding for Minority Class
SUM
MLP 1
k-mer Counting
CONCAT
Minority Class
ReLU
Embedding Vector 1
MLP 2
Microbe Co-occurrence Network 1
Embedding for Majority Class
KO Functions
Threshold
MLP 3
k-mer Counting
CONCAT
Majority Class
Number of Samples
ReLU
Embedding Vector 2
MLP 4
Microbe Co-occurrence Network 2
A Input Microbial Graphs with k-mer Counting
B Embedding Generation Layer

FIGURE 2 | Schematic illustration of generating the embedding representations of microbial network with sequence k-mer counting. **(A)** We process the long-tailed distribution class with hierarchical multitask learning, which learns the majority class and the minority class independently with two microbial networks. **(B)** The embedding generation layer learns the embedding vector of OTUs via aggregating the sequence information from current nodes and their neighbors.

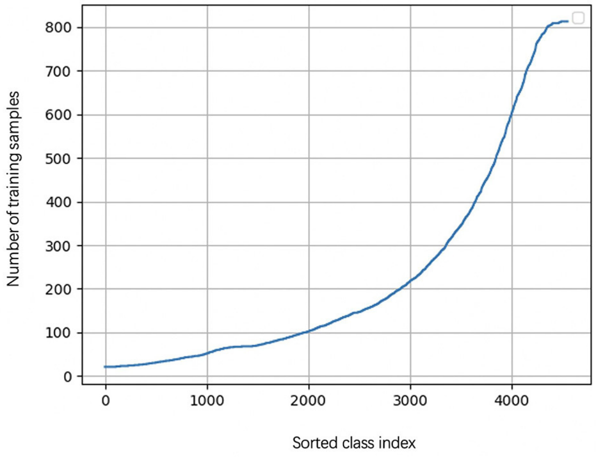


FIGURE 3 | The KO function number of appearance cure. The cure is highly skewed because a few dominant function classes claim most of the samples, while most of the other function classes are represented by relatively a few samples.

in more than half of OTUs, and the minority class only appears in less than half of OTUs. Traditional class rebalancing strategies, such as resampling and reweighting solutions, perform poorly on our tasks and slow down the training process. In fact, both majority and minority classes can be classified when trained independently. This motivated us to develop a hierarchical multitask training scheme designed to account for the poor prediction performance of minority classes (**Figure 2A**). The hierarchal multitask learning trains on the majority class and the minority class, respectively, so that the majority class in one task will not interfere the other task to learn the minority class and finally ameliorates the negative impact of the class imbalanced problem. The threshold of identifying a class belonging to a majority class or a minority class is a parameter that should be determined before model training, and the best threshold lets the model achieve the highest measurements on the validation set. Assume that the dataset is represented by space $V \times Y$, where V indicates an OTU set with n OTU, and Y indicates the corresponding KO function set. The KO function set is then divided into the majority class Y_{ma} and the minority class Y_{mi} by the number of samples. The OTU set and the KO function set can be shown as

$$V = \{v_1, v_2, \ldots, v_{n-1}, v_n\} \quad (1)$$

$$Y = Y_{ma} + Y_{mi} \quad (2)$$

The goal is to learn two functions, f_1, f_2, that classify every input data point to the proper classes:

$$y_{mai} \approx f_1(v_i), i \in \{1, \ldots, n\} \quad (3)$$

$$y_{mii} \approx f_2(v_i), i \in \{1, \ldots, n\} \quad (4)$$

Two models are considered to learn the majority class and the minority class from the long-tailed dataset separately and simultaneously. The input data v_i are the same for all tasks, but the output values y_i are different for each task so that the novel method can mitigate the biased tendency of the classifier toward either the majority or minority class. The HOPE approach uses cross-entropy loss function as the feedback information to train to learn the embedding vector. A drop in the loss value means less bias between predicted values and observed targets:

$$Loss_{ma} = -\sum_{i=1}^{n} (y_{mai} log(f_1(v_i)) + (1 - y_{mai}) log(1 - f_1(v_i))) \quad (5)$$

$$Loss_{mi} = -\sum_{i=1}^{n} (y_{mii} log(f_2(v_i)) + (1 - y_{mii}) log(1 - f_2(v_i))) \quad (6)$$

Embedding of Microbial Co-occurrence Network

Based on the hierarchical learning framework, HOPE also computes the embedding vectors of microbial co-occurrence networks (**Figure 2B**). Our learning model for graph embedding builds upon the GraphSAGE (Hamilton et al., 2017) algorithm, which performs learnable aggregation to replace full-graph Laplacian and finds the embedding map for a large graph. Our algorithm integrates topological information for neighbors of each node with its own sequence information and conserves the useful graph data as completely as possible. Embedding vectors not only save node information but also save the graph's edge information. HOPE maps two nodes to close points in the embedding space if and only if their features are highly similar and their neighborhoods are topologically similar. These closed OTUs in the embedding space have a high probability of similar functions. Thus, the embedding vectors could be used intuitively for classification.

Graph embedding involves two key steps. First, randomly select the neighbor nodes of the target nodes and aggregate the features of these nodes with those of the target nodes via a SUM function. The microbial co-occurrence network has a feature set $h = \{h_1, h_2, \ldots, h_n\}, h_i \in \mathbb{R}^f$, where n denotes the number of nodes in the graph. We uniformly sampled N nodes to pick out a fixed-size set of neighbor nodes V_N:

$$V_N = N(v) \quad (7)$$

A sum aggregator function is used to combine these features of neighboring nodes, and we gain the aggregated representation of neighbor h_N:

$$h_N = Aggregator\left(\{h_i, \forall i \in V_N\}\right) \quad (8)$$

The node's neighborhood embedding should be unique when no isomorphic neighborhoods exist. To aim this target, the aggregator function in the graph embedding algorithm has to be injective to achieve the upper bound method, the Weisfeiler–Lehman (WL) graph isomorphism test (Xu et al., 2019). Although the WL test has powerful capability in discriminating different graph structures, it does not know how to learn the intrinsic properties of nodes in a graph and generates unsuitable node features, which might be quite essential for function prediction task in testing. Thus, the WL test has poor generalization and would not be used in our study. The SUM aggregator that is used in this work is injective so that our method could be maximally powerful from a theoretical perspective and have well generalization. After aggregating features of the neighboring nodes, we then concatenate the target nodes feature, h_T, with the aggregated neighbor feature, h_N, and the concatenated vector is imported into the MLP layer with non-linear activation function σ:

$$h_E = \sigma\left(\left[W_1 \cdot h_T\right] CONCAT \left[W_2 \cdot h_N\right]\right) \quad (9)$$

$W_p, p \in \{1, 2\}$ are a set of weight matrices containing trainable weights that can be learned by back-propagation. This embedding generation process will be iterated in a loop as the searching depth deepens, K. For each iteration, target nodes will aggregate features from neighboring nodes to update the representation of a node, and the target node will gradually capture more and more information from further reaches of the nodes of the graph after two iterations of aggregation. After aggregating feature information from neighboring nodes in depth K, the layer outputs new embedding node features, as $h_{EK} = \{h_{e1}, h_{e2}, \ldots, h_{en}\}, h_{ei} \in \mathbb{R}^d, d < F$. Then in the next iteration, the outputs feature h_{EK} from the previous depth would be considered as the neighboring features in depth K-1, $h_{N(K-1),}$ and they will be aggregated with new target features, $h_{T(K-1)}$, for updating. Thus, for the embedding representation vector, we get a target node feature after iterations. **Figure 4** shows an example of an aggregating target node with its neighbors in two depths.

RESULTS

Experimental Design

The Experimental Datasets

In this study, three 16s rRNA sequencing datasets from abalone gut, human gut from early pregnancy, and *P. monodon* gut are tested in the experiments. The three datasets are available in NCBI with accession IDs ERP017548, SRP266217, and SRP261546. We constructed the microbial networks of the datasets by co-occurrence correlation, and the detail of these networks is shown in **Table 1**.

Experimental Strategies

Before making a comparison of specific methods, we first take steps to confirm our key hypothesis, i.e., that the co-occurrence relationship among microbes, together with neighborhood topological structures in the microbe network,

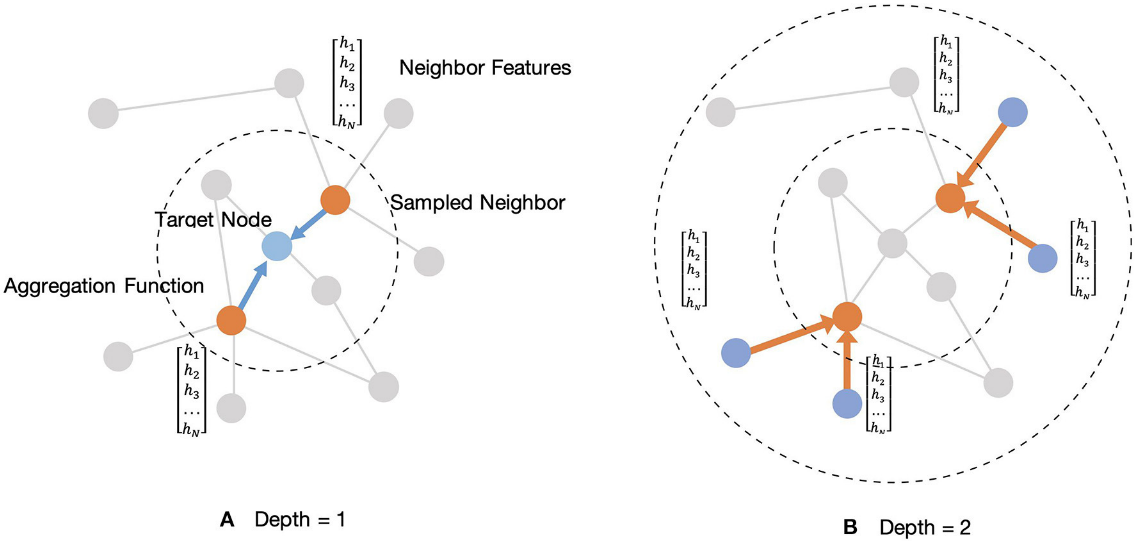

h_1 h_2 h_3 ... h_N
Neighbor Features
Sampled Neighbor
Target Node
Aggregation Function
h_1 h_2 h_3 ... h_N
A Depth = 1
h_1 h_2 h_3 ... h_N
h_1 h_2 h_3 ... h_N
h_1 h_2 h_3 ... h_N
h_1 h_2 h_3 ... h_N
B Depth = 2

FIGURE 4 | Illustration of sampled two-hop neighborhood and aggregation of features for these nodes. **(A)** Example of aggregating one-hop sampled neighborhood. **(B)** Example of aggregating two-hop sampled neighborhood.

TABLE 1 | Summary of the datasets used in our experiments.

	Abalone gut	**Human gut**	***Penaeus monodon* gut**
Nodes	15,796	3,254	42,84
Average edges	41.4	6.08	18.05
Classes	4,075	4,289	5,144
Training nodes	11,058	2,278	2,999
Validation nodes	3,159	649	855
Test Nodes	1,579	327	430

provides a fully functional context for function prediction. After that, to further evaluate the performance of our method, we design two experimental strategies: function prediction within a microbial community and function prediction across the microbial community. In the first strategy, both the training set and the test set come from the same microbial community to check the normal prediction ability of our method. The second strategy trains the model on one type of microbial community and then tests that model on a different, but related, microbial community, aiming to test the generalization ability of HOPE. Thus, HOPE must learn the universal knowledge on the training set and the validation set to make sound prediction results on the test set.

Function prediction for each OTU is modeled as a multiclass, multilabel supervised classification problem. In our study, the experimental dataset is divided into three distinct parts, including the training set, the validation set, and the testing set. We randomly split 20% of all OTUs into an independent testing set and designed an eight-fold cross-validation on the remaining 80% of all OTUs. The cross-validation is applied to learn the appropriate parameters in the weight matrices and select the appropriate hyperparameters. The goal is to develop the best model on both training and validation sets to achieve the highest prediction performance on the testing set. All methods mentioned above use rectified linear units (ReLUs) as the non-linearity functions to evaluate all datasets.

Hyperparameters of the Training Process

In training, we use the cross-entropy loss function for multiclass, multilabel classification together with the Adam optimizer (Kingma and Ba, 2015). The cross-entropy loss function treats each class independently and measures the difference between the ground truth label and predicted labels. The ground truth label of each class is 0 or 1, and the predicted result of each class is the probability between 0 and 1. When the predicted probability is far from the ground truth label, the loss value will be large.

We set $K = 2$ as the neighborhood region and sample sizes S1 $=$ 25 and S2 $=$ 10 at each hop of a neighborhood leading to the best performance during the graph embedding step. Adam and L2 regularization are adopted for model optimization with the size of mini batch at 128 and a learning rate of 0.01. To avoid overfitting, dropout is set as $=$ 0.4. All experiments use ReLUs as activation functions. The experiments are run on a single machine with 4 NVIDIA GeForce GTX1080 TI with CUDA Version 10.2, Intel(R) Xeon(R) CPU (E5-2620 v4 @ 2.10 GHz), and 128 Gb of RAM.

Hypothesis Verification

As noted above, this work is driven by the hypothesis that microbes with strong correlations, or strong neighborhood topology profiles, have similar, or highly correlated, functions. Therefore, we designed the following experiments to confirm that the topological structure of a neighbor node is predictive, or not, by comparing the similarity of KOs between OTU nodes with similar and different neighbors.

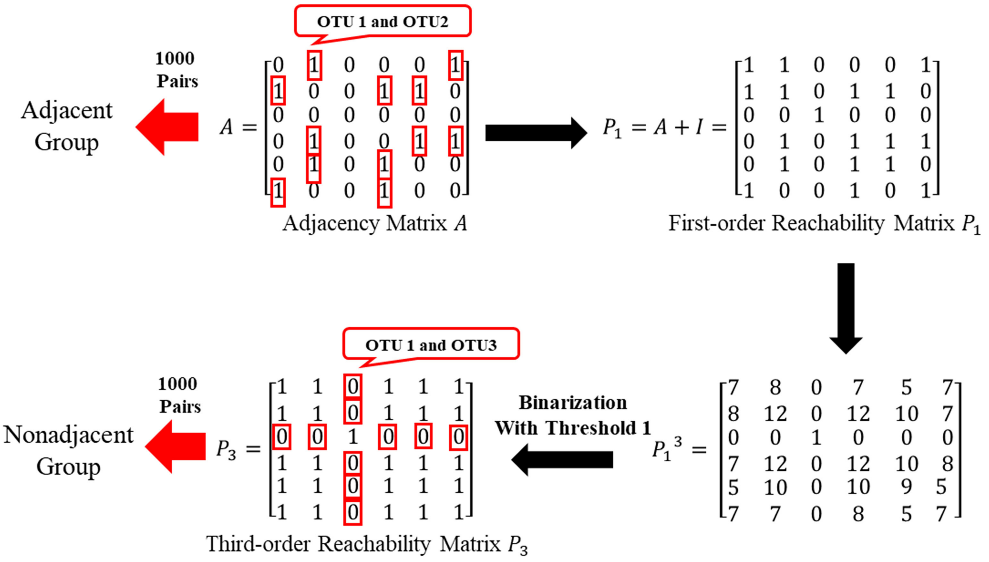


FIGURE 5 | Sketch of extracting "Adjacent Group" and "Non-adjacent Group".

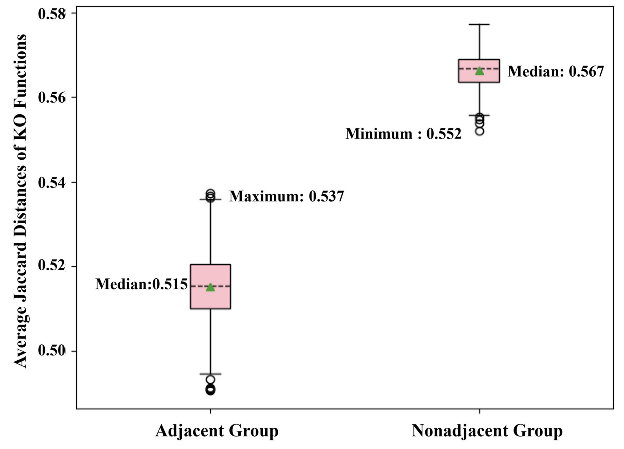


FIGURE 6 | Box plot comparing KO similarities 1,000 times with each test extracting 1,000 pairs of nodes for two groups and calculating the average Jaccard distance.

Verify Whether Two Adjacent Neighbors Share Highly Correlated Functions

By the adjacent matrix of the microbial network, 1000 pairs of adjacent OTU nodes are randomly extracted as the "Adjacent Group." Meanwhile, a three-step reachability matrix of OTUs is computed from the adjacency matrix. The two OTU nodes that cannot reach each other within three steps are extracted as the "Non-adjacent Group," which ensures that the node pairs are sufficiently far from each other. The extraction process of the two groups is shown in **Figure 5**. OTU functions are represented by a row of KO function vectors using 0/1 to indicate whether the OTU possesses the KO function or not. The distance between function vectors of two OTUs is calculated by the Jaccard distance. The function distances between the "Adjacent Group" and the "Non-adjacent Group" are calculated and averaged, respectively, and the tests are repeated 1,000 times. **Figure 6** shows the average function distances from the "Adjacent Group" and the "Non-adjacent Group" over the course of 1,000 respective tests. The median of average Jaccard distances of the "Adjacent Group" is 0.515, which is significantly lower than that of the "Non-adjacent Group." Even the maximum average distance from the "Adjacent Group" is smaller than the minimum average distance from the "Non-adjacent Group," which suggests that the adjacent relationships of OTU nodes contain the information required to predict KO functions.

Verify Whether Two Nodes Sharing Similar Neighbors Would Have Highly Correlated Functions

To further confirm that two nodes sharing similar neighbors have highly correlated functions, we use the corresponding row of the adjacent matrix to present the neighbor structure of each OTU. Hamming distance between every two rows of the adjacent matrix is adopted to evaluate neighborhood similarity between two corresponding OTUs. The smaller the Hamming distance between the two rows is, the more similar the neighborhood of the two nodes is. As shown in **Figure 7**, we selected the 10,000 pairs of nodes with the smallest Hamming distance in the neighborhoods as the "Similar Group" and the 10,000 pairs with the largest Hamming distance in the neighborhoods as the "Different Group." Therefore, OTU pairs in the "Similar Group" share similar neighbors, and the other pairs in the "Different Group" do not. Function similarity is also measured by the Jaccard distance between KO function vectors. Function distances for the "Similar Group" and the "Different Group" are calculated and plotted as boxplots, as shown in **Figure 8**. It is clear that KO functions are closer to each other for OTUs sharing

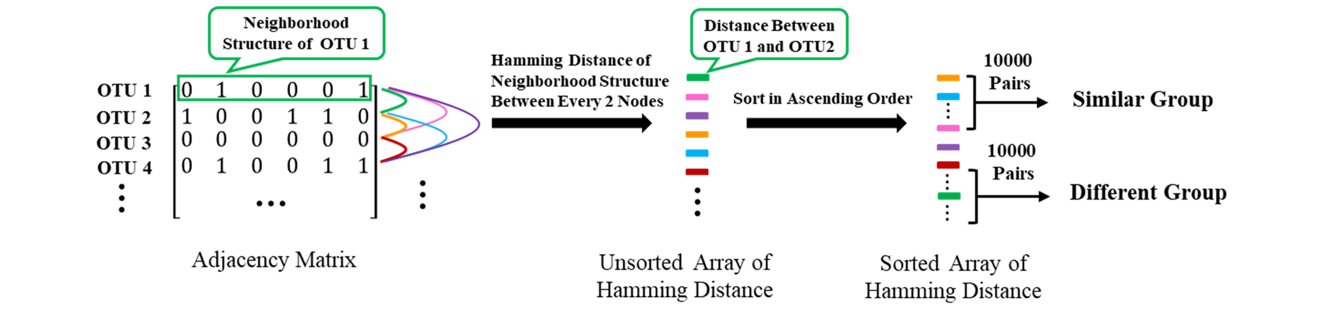

Neighborhood Structure of OTU 1
OTU 1
OTU 2
OTU 3
OTU 4
0 1 0 0 0 1
1 0 0 1 1 0
0 0 0 0 0 0
0 1 0 0 1 1
Adjacency Matrix
Hamming Distance of Neighborhood Structure Between Every 2 Nodes
Distance Between OTU 1 and OTU2
Unsorted Array of Hamming Distance
Sort in Ascending Order
Sorted Array of Hamming Distance
10000 Pairs
Similar Group
10000 Pairs
Different Group

FIGURE 7 | The extraction of “Similar Group” and “Different Group”.

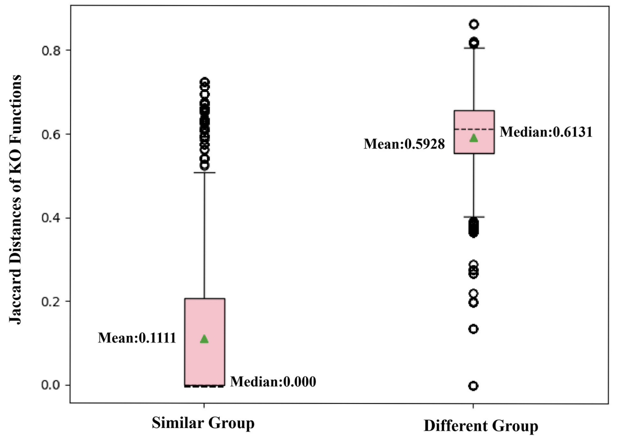


FIGURE 8 | Box plot comparing KO similarities by calculating Jaccard distances with each group, including 10,000 pairs of OTUs with similar or different neighbors.

common neighbors. For OTUs with highly different neighbors, KO functions are farther apart. The mean of Jaccard distances of the “Similar Group” is 0.1111, which is much smaller than that of the “Different Group.” Based on the two verification experiments, we infer that the OTUs that are adjacent to, or share, common neighbors would have highly similar KO functions. Therefore, learning the topological structure of the microbial co-occurrence network would provide clear and beneficial information for function predictions.

Evaluations and Comparisons of Experimental Results

HOPE, its two variants, HOPE-seq and HOPE-one, and GraphSAGE are applied to three datasets to evaluate by comparison the performance of HOPE. Recall that HOPE features hierarchical multitask learning to solve the highly skewed class distribution problem, and it incorporates information of both microbe sequences and microbe interactions in a co-occurrence network. Therefore, the variant HOPE-seq only uses the microorganism sequence representations as input but utilizes hierarchical multitask learning to train the classifier on majority and minority classes. HOPE-one ignores class imbalance problems but integrates the vector representation of sequences with microbial network information as an input feature. Both of the variant methods use the same hyperparameters and training strategies as parent HOPE. GraphSAGE (Hamilton et al., 2017) is a well-known and widely used graph embedding algorithm that provides an inductive framework to generate embeddings by sampling and aggregating features from a node’s local neighborhood. The aggregation function can have various forms, and the authors suggest three aggregator functions: a mean aggregator, an LSTM aggregator, and a pooling aggregator (shown as GS-Mean, GS-LSTM, and GS-Pooling, respectively). The mean aggregator simply takes the elementwise mean of the node’s features. The LSTM aggregator is built on a standard LSTM architecture (Hochreiter and Schmidhuber, 1997) to aggregate the nodes’ neighbors, which are listed to a random permutation, to embedding representations. The detailed description of the LSTM aggregator can be found in the study of GraphSAGE (Hamilton et al., 2017). In the pooling aggregator, an elementwise max-pooling operation is applied to aggregate information across the node’s neighbors.

In our experiments, we use four different measurements, including micro-averaged F1 score, macro-averaged F1 score, accuracy, and ROC-AUC score, to judge the comparison results. Micro-averaged F1 score and macro-averaged F1 score are both F1 scores, but they differ in the averaging method.

The micro-F1 score will aggregate the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) of all classes to compute the average F1-score. Assuming n classes, the TP-value, FP-value, and TF value of the ith class are represented as TP_i, FP_i, and FN_i, respectively:

$$precision_{mi} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \quad (10)$$

$$recall_{mi} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \quad (11)$$

$$micro - F1 = 2\frac{recall_{mi} \times precision_{mi}}{recall_{mi} + precision_{mi}} \quad (12)$$

On the other hand, the macro-F1 score will compute the F1-score independently for each class and then take the average as

$$precision_i = \frac{TP_i}{TP_i + FP_i} \tag{13}$$

$$precision_{ma} = \frac{\sum_{i=1}^{n} precision_i}{n} \tag{14}$$

$$recall_i = \frac{TP_i}{TP_i + FN_i} \tag{15}$$

$$recall_{ma} = \frac{\sum_{i=1}^{n} recall_i}{n} \tag{16}$$

$$macro - F1 = 2\frac{recall_{ma} \times precision_{ma}}{recall_{ma} + precision_{ma}} \tag{17}$$

Accuracy is the ratio of correct predictions to total input samples. The ROC-AUC score is defined as the area under the ROC curve. It provides an aggregate measure of performance across all possible classification thresholds. The ROC-AUC score varies between 0 and 1, and the closer it is to 1, the better the performance of the classifier.

Functional Prediction Within the Same Microbial Community

In this part, we use the training, validation, and testing data from the same microbial community and calculate various measurements for every experiment (see **Table 2**). HOPE is compared against its variants and GraphSAGE. According to the results, HOPE nearly outperforms all baselines on various measurements, especially the micro-F1 score and macro-F1 score. However, the performance of HOPE is comparable to its two variants in terms of accuracy and ROC-AUC. For example, HOPE outperforms HOPE-one by 9.5% in the micro-F1 score on the Abalone Gut Microbiota dataset and 15.6% in the macro-F1 score on the *P. monodon* intestine dataset. In some parameters, the performance of HOPE-one is better than that of HOPE-seq, like accuracy and ROC-AUC, but HOPE-seq can improve upon HOPE-one by a margin of 5.7% in the micro-F1 score on the Abalone Gut Microbiota. Since HOPE integrates sequence information with microbial network information via graph embedding, thus combining the advantages of its two variants, it nearly achieves the highest performance. **Table 2** also shows the performance results of HOPE compared to the variants of GraphSAGE on the benchmark datasets. HOPE nearly achieves the highest score in all measurements and outperforms two baselines by a significant margin. According to **Table 2**, we find that HOPE-one achieves better results on accuracy and ROC-AUC than HOPE on three datasets because HOPE sacrifices some performance on the majority class to learn the minority class well.

Functional Prediction Across Different Microbial Communities

We further consider generalizing across different microbial communities, which requires our model to learn the context of common functions from one microbe to infer the functions of other organisms. Some researchers may want to know novel microbial functions but have only information about related microbial functions. In this case, the generalization ability of the algorithm is very important. Therefore, in this part, we design experiments with the different test sets to evaluate the generalization ability of HOPE.

We first set the training and validation data from the abalone gut microbiota and use human feces and shrimp intestine microbiota to construct microbial networks as a test set. In these scenarios, we evaluate the performance of our model when the training data are different from the data used in the test set.

TABLE 2 | The performance of HOPE and its variants and GraphSAGE within the same microbial community for training and testing.

Method	Abalone gut microbiota				Human feces				***Penaeus monodon*** **intestine**			
	Mi- F1*	**Ma-F1***	**Accuracy**	**ROC-AUC**	**Mi-F1**	**Ma-F1**	**Accuracy**	**ROC-AUC**	**Mi-F1**	**Ma-F1**	**Accuracy**	**ROC-AUC**
HOPE-seq	0.786	0.500	0.887	0.840	0.742	0.236	0.861	0.807	0.907	0.701	0.941	0.923
HOPE-one	0.729	0.544	**0.921**	**0.881**	0.742	0.199	0.883	0.816	0.941	0.675	0.963	**0.955**
GS-mean	0.727	0.433	0.861	0.822	0.672	0.217	0.843	0.780	0.872	0.515	0.918	0.908
GS-LSTM	0.677	0.290	0.843	0.777	0.711	0.205	0.867	0.800	0.713	0.263	0.829	0.787
GS-pooling	0.735	0.387	0.879	0.806	0.747	0.183	**0.888**	**0.816**	0.738	0.263	0.845	0.804
HOPE	**0.824**	**0.592**	0.908	0.869	**0.758**	**0.309**	0.870	0.811	**0.941**	**0.831**	**0.963**	0.949

**Mi-F1 and Ma-F1 represent micro-F1 score and macro-F1, respectively. The bold values mean the best performance of each column of index.*

TABLE 3 | Evaluation of generalization performance of HOPE across different microbial communities.

Training set organism	Test set organism	Mi-F1	Ma-F1	Accuracy	ROC-AUC
Human feces	Human feces	0.758	0.309	0.842	0.811
Abalone gut	Human feces	0.728	0.217	0.867	0.819
Penaeus monodon intestine	*Penaeus monodon* intestine	0.941	0.831	0.963	0.949
Abalone gut	*Penaeus monodon* intestine	0.837	0.529	0.877	0.866

The first and second columns list the microbial communities used in the training set and the testing set, respectively. The first and third rows list the baseline of performance when training and test sets are from the same microbial community.

Table 3 summarizes the performance of HOPE with different test sets. Compared to baselines, experiments utilizing a test set different from the training set achieve lower scores but within an acceptable range. We train the model on the abalone gut microbiota dataset and test the model on datasets from human feces and shrimp intestine microbiota. Although using training sets from a different source, results show that HOPE achieves nearly 90% performance of experiments when the training set and test set data belong to the same species. HOPE achieves high performance in generalization, which means that our approach can learn the fundamental knowledge from known microbial functions and infer the functions of unseen microorganisms.

Discussion for Class Imbalance Problem

We find that KO functions with a large number of annotation samples generally outperform KO functions with a few annotations. Further experiments explore the relationship between the number of training samples and the variance in predictive performance and plot the result in **Figure 9**. It can be seen that the predictive performance is strongly correlated with the number of instances in the training set. We build linear regressions for the measurement scores and the number of samples for every KO function, and the coefficients of the explanatory variable in all regressions are significantly greater than zero ($P = 0.0000$). The statistical results prove that KO functions with rich training sample annotations perform better than KO functions represented by only a few samples.

In all experiments mentioned above, we observe that some specific KO functions are easily classified to wrong places, causing low scores across the measurements evaluated. Even though some specific KO functions have been learned by a large amount of training samples, like K00096, K02080, and K10014, their F1 scores are nearly zero. Owing to the hierarchical nature of KOs, these bad KOs are defined as low-level, or rare, existing functions. In the future, additional weights based on the general level of KO should be assigned to each class to achieve better performance.

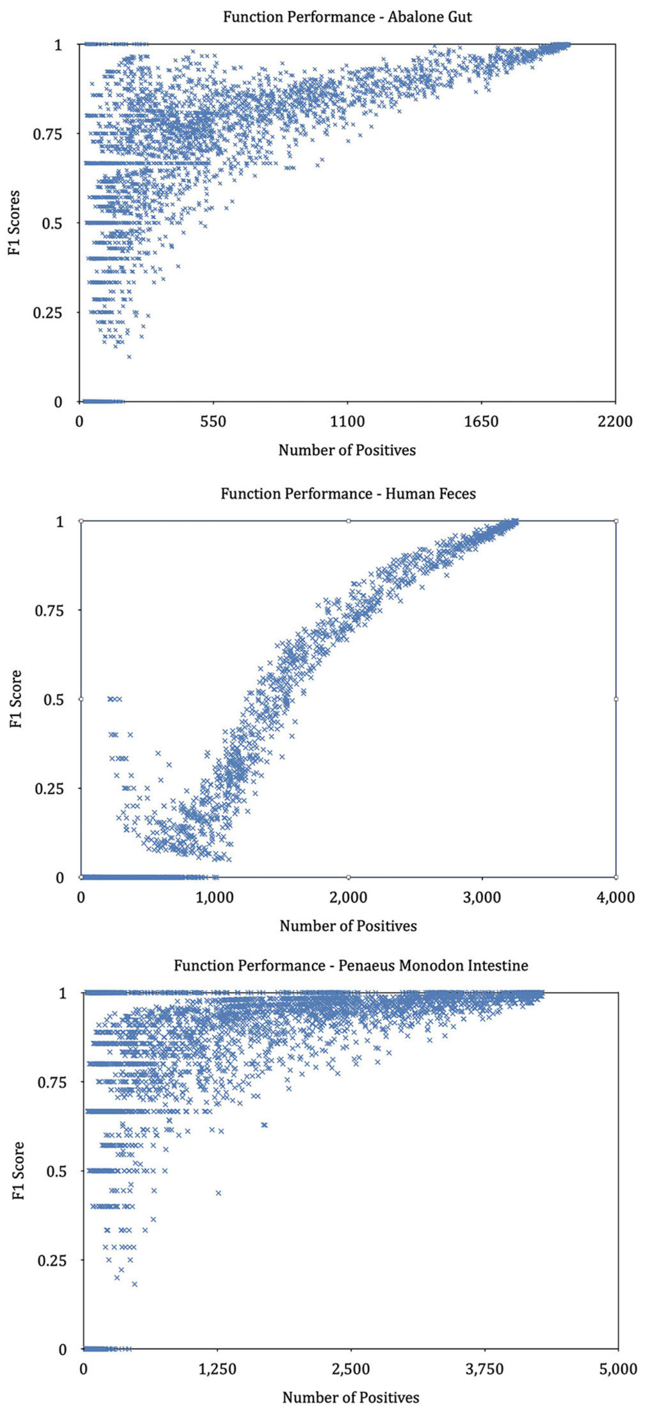


FIGURE 9 | Performance of KOs with different annotated samples. The graphs plot the predictive performance of each KO in our method as a function of the number of training samples.

CONCLUSION AND DISCUSSION

In this paper, a pipeline for the HOPE method is proposed for the analysis of microbial functions. The method leverages hierarchical multitask learning and graph embedding to extract features from sequence compositional signatures and topological patterns in non-linear microbial interaction networks. The hierarchical multitask learning module is to cope with class imbalanced datasets and achieve significant performance gains on predicting functions that appear in a few training samples. Using the graph embedding model, HOPE integrated the sequence compositional signatures and co-occurrence relationship among OTUs in microbial communities with the *k*-mer frequency feature in each node and topological patterns in microbial networks. Therefore, HOPE outperforms baselines on almost all indexes in all experiments. In detail, the percentage of macro-F1 scores reached from our classifier has an increased score of at least seven percentage points compared to the other methods. Experiment results also showed that HOPE has satisfactory generalization ability when it predicts functions across different microbial communities. Because the graph embedding of microbial co-occurrence networks conserves the interactions and similarities among OTUs, which are useful for inferring unknown functions, HOPE demonstrates significant generalization ability. Several potential improvements are possible. In the future, during the construction of the microbial network, the threshold of defining an edge between two OTUs could change to a learnable value. The training of HOPE is more time-consuming than the previous algorithms because the extra MLP layer for function prediction requires the optimization of much more parameters.

Although the primary purpose of HOPE is the prediction of microbial functions on the microbial co-occurrence network, the framework can be used on other related scenarios, such as the prediction of gene function based on the gene co-expression network.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YW, YH, XZ, and QZ planned the project. YW and YH designed the model and experiments. YH, XZ, and QZ performed the experiments. YW, YH, and QZ analyzed the data. YH and WH contributed the materials and analysis tools. YW, YH, and QZ wrote the main manuscript. All authors read and approved the final manuscript.

FUNDING

The research was supported by National Natural Science Foundation of China (61673324), National Key Research and Development Program of China (2018YFD0901401), Open Fund of Engineering Research Center for Medical Data Mining and Application of Fujian Province (MDM2018002), and Natural Science Foundation of Fujian (2018J01097).

REFERENCES

Ashauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 31, 2882–2884. doi: 10.1093/bioinformatics/btv287

Bardgett, R. D., and Der Putten, W. H. V. (2014). Belowground biodiversity and ecosystem functioning. *Nature* 515, 505–511. doi: 10.1038/nature13855

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Cozzetto, D., Buchan, D. W. A., Bryson, K., and Jones, D. T. (2013). Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinformatics* 14:S1. doi: 10.1186/1471-2105-14-S3-S1

Desantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Escalas, A., Hale, L., Voordeckers, J. W., Yang, Y., Firestone, M. K., Alvarezcohen, L., et al. (2019). Microbial functional diversity: from concepts to applications. *Ecol. Evol.* 9, 12000–12016. doi: 10.1002/ece3.5670

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Gligorijevic, V., Barot, M., and Bonneau, R. (2018). deepNF: deep network fusion for protein function prediction. *Bioinformatics* 34, 3873–3881. doi: 10.1093/bioinformatics/bty440

Hamilton, W. L., Ying, Z., and Leskovec, J. (2017). "Inductive representation learning on large graphs," in *Paper Presented at the Neural Information Processing Systems* (Long Beach, CA).

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780.

Huang, C., Li, Y., Loy, C. C., and Tang, X. (2016). "Learning deep representation for imbalanced classification," in *Paper Presented at the Computer Vision and Pattern Recognition* (Las Vegas, NV).

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopaedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Karlin, S., Mrázek, J., and Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* 179, 3899–3913. doi: 10.1128/jb.179.12.3899-3913.1997

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F., and Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Netw.* 29, 3573–3587. doi: 10.1109/TNNLS.2017.2732482

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *Paper Presented at the International Conference on Learning Representations* (San Diego, CA).

Kulmanov, M., Khan, M. A., and Hoehndorf, R. (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 34, 660–668. doi: 10.1093/bioinformatics/btx624

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Lele, H., Tao, H., Xiaohe, S., Wen-Cong, L., Yu-Dong, C., Kuo-Chen, C., et al. (2011). Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. *PLoS ONE* 6:e14556. doi: 10.1371/journal.pone.0014556

Liu, L., Tang, L., He, L., Yao, S., and Zhou, W. (2017). Predicting protein function via multi-label supervised topic model on gene ontology. *Biotechnol. Biotechnol. Equip.* 31, 630–638. doi: 10.1080/13102818.2017.1307697

Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., et al. (2019). Charting the complexity of the marine microbiome through single-cell genomics. *Cell* 179:1623. doi: 10.1016/j.cell.2019.11.017

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, 590–596. doi: 10.1093/nar/gks1219

Ren, W., Yin, J., Chen, S., Duan, J., Liu, G., Li, T., et al. (2016). Proteome analysis for the global proteins in the jejunum tissues of enterotoxigenic *Escherichia coli*-infected piglets. *Sci. Rep.* 6, 25640. doi: 10.1038/srep25640

Wang, X., Tang, B., Luo, X., Ke, C., Huang, M., You, W., et al. (2020). Effects of temperature, diet and genotype-induced variations on the gut microbiota of abalone. *Aquaculture* 524:735269. doi: 10.1016/j.aquaculture.2020.735269

Wang, Y., Chen, Q., Deng, C., Zheng, Y., and Sun, F. (2020). KmerGO: a tool to identify group-specific sequences with k-mers. *Front. Microbiol.* 11:2067. doi: 10.3389/fmicb.2020.02067

Wang, Y., Fu, L., Ren, J., Yu, Z., Chen, T., and Sun, F. (2018). Identifying group-specific sequences for microbial communities using long k-mer sequence signatures. *Front. Microbiol.* 9:872. doi: 10.3389/fmicb.2018.00872

Wang, Y., Liu, L., Chen, L., Chen, T., and Sun, F. (2014). Comparison of metatranscriptomic samples based on k-tuple frequencies. *PLoS ONE* 9:e84348. doi: 10.1371/journal.pone.0084348

Wass, M. N., Barton, G., and Sternberg, M. J. E. (2012). CombFunc: predicting protein function using heterogeneous data sources. *Nucleic Acids Res.* 40, 466–470. doi: 10.1093/nar/gks489

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). "How powerful are graph neural networks," in *Paper Presented at the International Conference on Learning Representations* (New Orleans, LA).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 January 2021
doi: 10.3389/fgene.2020.626500

Identification of Protein Subcellular Localization With Network and Functional Embeddings

Xiaoyong Pan [1,2†], Hao Li [3†], Tao Zeng [4], Zhandong Li [3], Lei Chen [5], Tao Huang [6] and Yu-Dong Cai [1*]*

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Key Laboratory of System Control and Information Processing, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Ministry of Education of China, Shanghai, China, [3] College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [4] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China, [5] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [6] Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Peng Zhang,
Shanghai University of Medicine and Health Sciences, China
Yun Li,
University of Pennsylvania, United States

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *06 November 2020*
Accepted: *21 December 2020*
Published: *20 January 2021*

Citation:
Pan X, Li H, Zeng T, Li Z, Chen L, Huang T and Cai Y-D (2021) Identification of Protein Subcellular Localization With Network and Functional Embeddings. Front. Genet. 11:626500. doi: 10.3389/fgene.2020.626500

The functions of proteins are mainly determined by their subcellular localizations in cells. Currently, many computational methods for predicting the subcellular localization of proteins have been proposed. However, these methods require further improvement, especially when used in protein representations. In this study, we present an embedding-based method for predicting the subcellular localization of proteins. We first learn the functional embeddings of KEGG/GO terms, which are further used in representing proteins. Then, we characterize the network embeddings of proteins on a protein–protein network. The functional and network embeddings are combined as novel representations of protein locations for the construction of the final classification model. In our collected benchmark dataset with 4,861 proteins from 16 locations, the best model shows a Matthews correlation coefficient of 0.872 and is thus superior to multiple conventional methods.

Keywords: protein subcellular localization, network embedding, functional embedding, gene ontology, KEGG pathway

INTRODUCTION

The functions of proteins are closely related to their subcellular locations in cells. In studying proteins, determining their locations in cells is usually the first step, and these locations are used as guides for designing drugs. Thus, many experimental methods for identifying protein locations have been developed, such as *in situ* hybridization. Through these methods, a large number of proteins have been verified and recorded in biological databases, such as the Swiss-Prot database. In addition, these data serve as benchmark datasets for developing machine learning methods and useful in the computational identification and investigation of protein locations.

Many computational methods based on machine learning for predicting protein subcellular locations have been proposed. For example, Chou and Cai (2002) proposed a support vector machine-based method for predicting protein locations with the use of functional domain data. LocTree2 (Goldberg et al., 2012) presents a hierarchical model for classifying 18 protein locations. To further improve prediction effectiveness, LocTree3 incorporates homology information into the models (Goldberg et al., 2014). Hum-mPloc 3.0 trains an ensemble classifier by integrating sequence and gene ontology information (Zhou et al., 2017). Recently, deep learning has achieved remarkable results in computational biology, particularly in identifying protein subcellular locations. In classification tasks, deep learning automatically learns high-level features rather than hand-designing features. For example, DeepLoc (Almagro Armenteros et al., 2017) presents a recurrent neural network with attention mechanism for the identification of protein locations, using sequences alone. rnnloc (Pan et al., 2020a) combines network embeddings and one-hot encoded functional data to predict protein locations with the use of a recurrent neural network. In Hum-mPloc 3.0 and rnnloc, functional data demonstrate strong discriminating power for different subcellular locations. However, both methods encode functional data into a high-dimensional one-hot encoded vector, which may cause feature disaster, especially when the number of training samples is smaller than the number of features.

For the above issues, embedding-based methods can be applied to the transfer of high-dimensional one-hot encoding into distributed vectors for sequential and network data. Given that interacting proteins generally share similar locations, node2vec (Grover and Leskovec, 2016) can be used in learning network embeddings for individual proteins from a protein–protein network, which help better represent the protein interaction information into feature vectors.

In this study, we present an embedding-based method for predicting protein locations. It learns network embeddings from a protein–protein network and functional embeddings of GO/KEGG terms. Then, these learned embeddings are used to represent proteins and further selected using feature selection methods. Finally, an optimal feature subset and classifier are obtained for the classification of protein subcellular localization, and the optimal classifier is superior to multiple conventional methods.

MATERIALS AND METHODS

In this study, we first collect a benchmark dataset for protein localization. Then we learn network embeddings from a protein–protein network, using node2vec and functional embeddings from KEGG/GO functional data and word2vec. Then, the learned embeddings are used to represent each protein. To obtain refined combined embeddings, we use two-step feature selection methods in determining the optimal features and classifiers in predicting protein locations. The whole process is illustrated in **Figure 1**.

Datasets

The original 5,960 protein sequences are retrieved from a previous study (Li et al., 2014), which are extracted from Swiss-Prot (http://cn.expasy.org/, release 54.0). The protein sequences do not include proteins with <50 amino acids or more than 5,000 amino acids and unknown amino acids. The included proteins are processed through CD-HIT (Li and Godzik, 2006). Sequence similarity between each pair of proteins is <0.7. Given that we extract features from gene ontology (GO) terms and KEGG pathways of proteins through natural language processing methods, we exclude proteins without GO terms and KEGG pathways, finally obtaining a total of 4,861 proteins. The proteins are classified into 16 categories according to their subcellular locations. The number of proteins from each location is listed in **Table 1**.

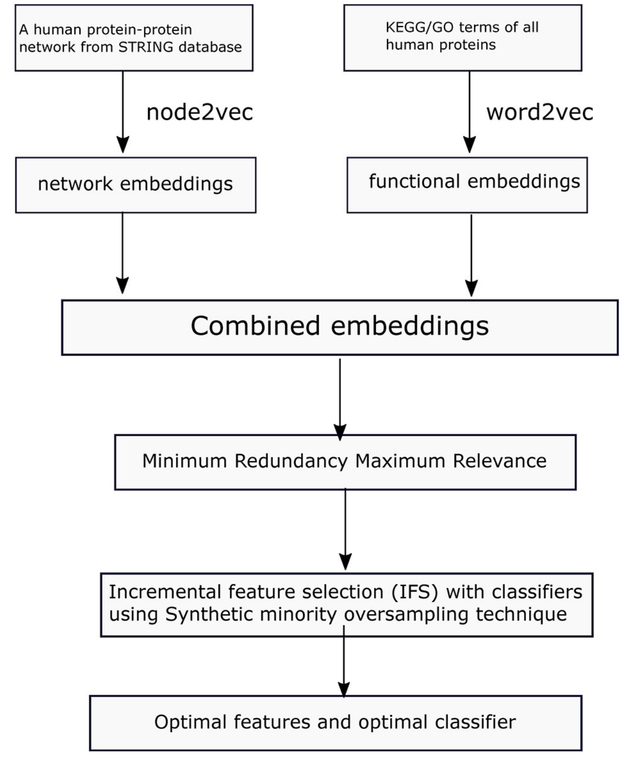

A human protein-protein network from STRING database
KEGG/GO terms of all human proteins
node2vec
word2vec
network embeddings
functional embeddings
Combined embeddings
Minimum Redundancy Maximum Relevance
Incremental feature selection (IFS) with classifiers using Synthetic minority oversampling technique
Optimal features and optimal classifier

FIGURE 1 | Flowchart of the proposed method in this study.

TABLE 1 | Number of proteins in each category.

Category	Number of proteins
Biological membrane	1,483
Cell periphery	33
Cytoplasm	488
Cytoplasmic vesicle	69
Endoplasmic reticulum	188
Endosome	25
Extracellular space or cell surface	636
Flagellum or cilium	3
Golgi apparatus	95
Microtubule cytoskeleton	48
Mitochondrion	326
Nuclear periphery	31
Nucleolus	108
Nucleus	1,229
Peroxisome	45
Vacuole	54

Protein Representation

One-Hot Encoded Representation Based on GO Term and KEGG Pathway

The GO terms and KEGG pathways are the essential properties of proteins. A representation containing this information is an excellent scheme for encoding each protein.

A protein p can be encoded into a binary vector, $V_{GO}(p)$, based on its GO terms, which is formulated by

$$V_{GO}(P) = [g_1, g_2, \cdots, g_m]^T, \tag{1}$$

where m represents the number of GO terms and g_i is defined as follows:

$$g_i = \begin{cases} 1 & if\ p\ is\ annotated\ by\ the\ i-th\ GO\ term \\ 0 & Otherwise \end{cases} \tag{2}$$

In this study, 22,729 GO terms are included, which induced a 22,729-D vector for each protein.

Moreover, with its KEGG pathways, it can also be encoded into a vector, $V_{KEGG}(p)$, with the formula

$$V_{GO}(P) = [k_1, k_2, \cdots, k_n]^T, \tag{3}$$

where n represents the number of KEGG pathways and k_i is defined as follows:

$$k_i = \begin{cases} 1 & if\ p\ is\ annotated\ by\ the\ i-th\ KEGG\ pathway \\ 0 & Otherwise \end{cases} \tag{4}$$

Here, 328 KEGG pathways are used, inducing a 328-D vector for each protein.

Two vectors based on GO terms and KEGG pathways are concatenated to a final vector. Thus, each protein is represented by a 23,057-D vector. We use Boruta feature selection (Kursa and Rudnicki, 2010) to reduce the computational burden and retain relevant features.

Functional Embeddings Based on GO Term and KEGG Pathway

Given that the one-hot encoded representation of GO (Ashburner et al., 2000) and KEGG (Ogata et al., 1999) terms is highly dimensional, the method by which they are mapped into low-dimensional embeddings is extremely important. GO/KEGG terms co-occur in a protein frequently and may thus be similar in distance, although distances among GO/KEGG terms vary. Thus, we apply word2vec (Mikolov et al., 2013) to learn an in-depth representation of GO/KEGG terms, representing each GO/KEGG term with a vector containing continuous values.

We first collect whole human proteins with GO/KEGG terms. Each GO or KEGG term is a word, and each protein is a sentence. The set of human proteins is a corpus. We run Word2vec program in genism (https://github.com/RaRe-Technologies/gensim) on this corpus to learn the embeddings of each GO/KEGG term.

Each protein contains multiple GO and KEGG terms. After obtaining the embeddings for each KEGG/GO term, we average the embeddings of KEGG/GO terms within a protein as the functional embeddings of this protein.

Network Embeddings From a Protein–Protein Network

In a protein–protein network, each node is a protein, and the edge is whether the two proteins interact or not. We first download a human protein–protein network from STRING (version 9.1) (Szklarczyk et al., 2017), and the network consists of 2,425,314 interaction pairs and 20,770 proteins.

node2vec is designed to learn embeddings from a graph through a flexible sampling approach and maximizes the log probability of nodes, given the learned embeddings:

$$\max_e \sum_{v \in V} \log P\left(N\left(v|e\left(v\right)\right)\right) \tag{5}$$

where v is the node, $N(v)$ is the neighborhood of the node v, and e is the mapping function from nodes to embeddings.

In this study, we use node2vec implemented at https://snap.stanford.edu/node2vec/, and the dimension of the learned embeddings is set at 500. Finally, the network embeddings of each protein are obtained.

Feature Selection

Instead of directly using combined features from network embeddings and functional embeddings for each protein, we further use minimum redundancy maximum relevance (mRMR) (Peng et al., 2005) to analyze these embedding features, which has wide applications in tackling different biological problems (Wang et al., 2018; Li et al., 2019, 2020; Zhang et al., 2019, 2020; Chen et al., 2020). This method has two criteria to evaluate the importance of features. One is the maximum relevance to class labels and the other is the minimum redundancy to other features. Based on these two criteria, mRMR method generates a feature list, named mRMR feature list. Regardless of relevance and redundancy, mutual information (MI) is adopted in this method to make evaluation. For two variables x and y, their MI is computed by

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy, \tag{6}$$

where $p(x)$ and $p(x, y)$ denote the marginal probabilistic density and joint probabilistic density, respectively. A high MI indicates the strong associations of two variables. The mRMR feature list is produced by adding features one by one. Initially, it is empty. In each round, for each of features not in the list, its relevance to class labels, evaluated by MI value of the feature and class labels, and redundancy to features in the list, assessed by the mean of MI values of the feature and those in the list, are calculated. A feature with maximum difference of relevance to class labels and redundancy to features in the list is picked up and appended to the list. When all features are in the list, the mRMR feature list is complete. Here, we used the mRMR program provided in http://penglab.janelia.org/proj/mRMR/. It is executed with its default parameters.

The mRMR method can only output a feature list. Which features are optimum is still a problem. In view of this, the incremental feature selection (IFS) (Liu and Setiono, 1998) method is employed. This method can extract an optimum feature combination for a given classification algorithm. In detail, from the mRMR feature list yielded by mRMR, IFS generates a series of feature subsets with a step 1, that is, the top feature in the list comprises the first feature subset, the top two features comprise the second feature subset, and so forth. For each feature subset, a classifier is built with a given classification algorithm and samples represented by features in the subset. All constructed classifiers are evaluated by a cross-validation method (Kohavi, 1995). We select the classifier with the best performance and call it as the optimum classifier. The corresponding feature subset is termed as the optimum feature subset and features in this feature subset are denoted as the optimal features.

Synthetic Minority Oversampling Technique

The number of proteins from different locations varies, resulting in a data imbalance problem. To reduce the impact of data imbalance on classification model construction, we apply synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) to generate some synthesized samples for minority classes. For each location, except the location with the largest number of proteins, we synthesize new proteins and add them to this location until each location has almost the same number of proteins.

Classification Algorithm

In this study, we test four classification algorithms to select the best one for our task: decision tree (DT) (Safavian and Landgrebe, 1991), K-nearest neighbors (KNN) (Cover and Hart, 1967), random forest (RF) (Breiman, 2001), and support vector machine (SVM) (Cortes and Vapnik, 1995).

K-Nearest Neighbors

KNN is a simple intuitive method for classifying samples. Given a query sample, it calculates the distance between a query sample and training samples. Then. it selects k training samples with the least distance, and the label of the query sample is determined by major voting, which assign a label with the most votes to the query sample.

Decision Tree

The DT is an interpretable classifier method, which can automatically learn classification rules from data. It uses a greedy strategy to build a flow-like structure; each internal node is determined by a feature to go to the left or right child node. The leaf node represents the outcome labels. The DT in Scikit-learn implements the CART algorithm with Gini index. It is used in this study.

Random Forest

RF (Breiman, 2001; Jia et al., 2020; Liang et al., 2020; Pan et al., 2020b) is a meta predictor with multiple DTs, which are grown from the bootstrap samples consisting of randomly selected features. Given a new sample, RT first uses its multiple trees for the prediction of sample labels, and then majority voting is used in determine the label of the new sample.

Support Vector Machine

SVM (Cortes and Vapnik, 1995; Chen et al., 2018a,b; Liu et al., 2020; Zhou et al., 2020) is a supervised classifier based on statistical theory, and it builds a hyperplane with a maximum margin between two classes. It first transforms nonlinear data from a low-dimensional space to a linear high-dimensional space with a kernel trick, then the margin between two classes in the high-dimensional space is maximized for acquisition of SVM parameters. Given a test sample, SVM determines the label according to the side of the hyperplane where it is located.

In this study, we use the Scikit-learn package to implement above four classification algorithms.

Baseline Methods

BLAST

To indicate the utility of the proposed method, we further employ basic local alignment search tool (BLAST) (Altschul et al., 1990) to construct a baseline method and make comparisons. In a given protein sequence, BLAST search the most similar protein sequences, measured with an alignment score, in the training dataset. The method based on BALST directly assigns the class of the most similar protein sequence to a given protein sequence as its predicted class. Such method is evaluated with a Jackknife test.

DeepLoc

DeepLoc (Almagro Armenteros et al., 2017) is another deep learning based method for predicting protein locations from sequences. We use DeepLoc downloaded from https://github.com/ThanhTunggggg/DeepLoc with default parameters.

RESULTS AND DISCUSSION

In this section, we first visualize the learned embeddings, using T-SNE, then we evaluate the effectiveness of different classifiers with different input embedding features. Finally, we compare our proposed method with baseline methods.

Visualization of the Learned Functional and Network Embeddings

To demonstrate the power of the learned embeddings, we visualize these embeddings, one-hot encoded features, and the combined network and functional embeddings, respectively. As shown in **Figure 2**, the embeddings can distinguish proteins from different locations to some extent. The learned functional embeddings (**Figure 2B**) shows higher discriminate power on some locations (e.g., for discriminating biological membrane) than the one-hot encoded representation based on functional data (**Figure 2A**). As shown in **Figure 2C**, the network embeddings have some discriminate power for some locations, for example, endosomes, which cannot be easily separated by functional embeddings. Also, the combined embeddings (**Figure 2D**) of functional and network embeddings have strong discriminating power. Intuitively, the four types of embeddings have similar discriminating power on the whole.

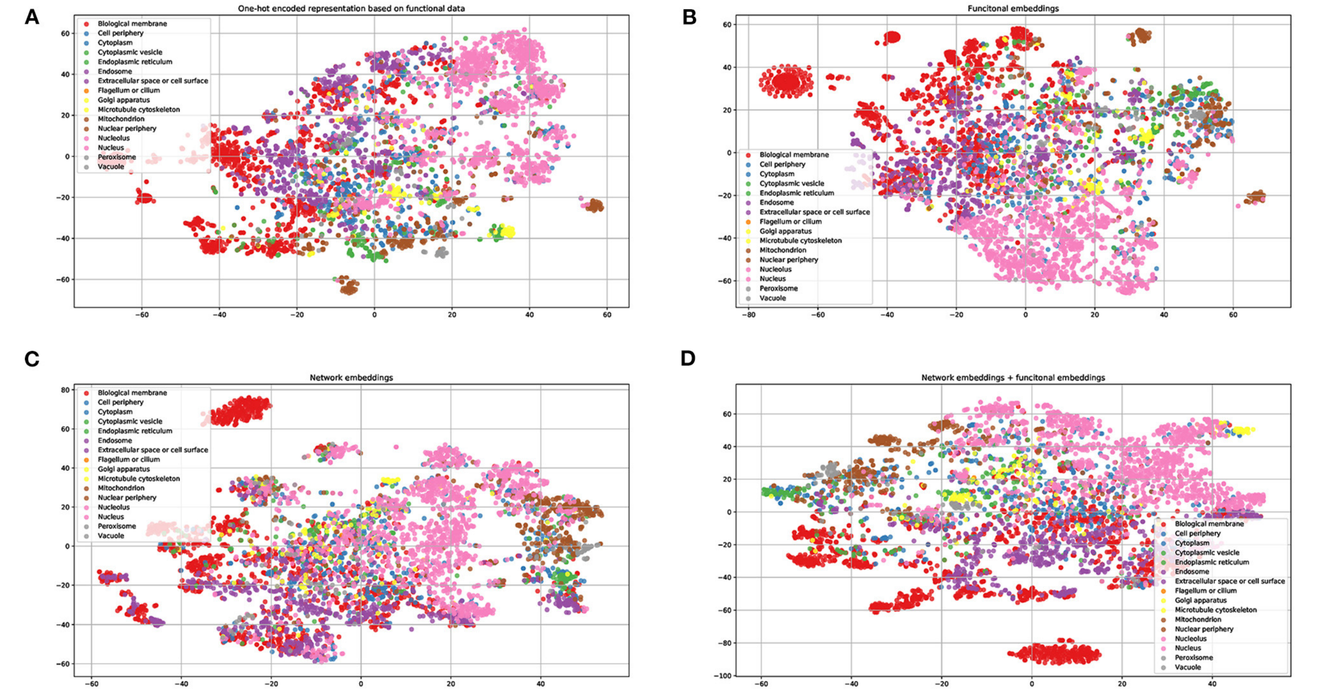


FIGURE 2 | Visualization of one-hot encoded functional features and the learned embeddings. **(A)** One-hot encode representation of functional data KEGG/GO terms; **(B)** the learned network embeddings from a protein-protein network; **(C)** the learned embeddings of functional data using word2vec; **(D)** the combined network and functional embeddings.

Effectiveness of Different Classifiers With Different Input Embedding Features

We evaluate the effectiveness of different classifiers with different input features, including one-hot encoded representations of GO/KEGG terms, functional embeddings, network embeddings, and the combination of network and functional embeddings. All the input features are first reordered using mRMR, resulting in the mRMR feature list. Then, a series of feature subsets are generated based on such list. On the series of feature subsets, several classifiers are built with a given classification algorithm. Each constructed classifier is assessed by 10-fold cross-validation. The measurements for each classifier, including accuracies on 16 categories, overall accuracy (ACC) and Matthew correlation coefficient (MCC) (Matthews, 1975; Gorodkin, 2004), are provided in **Supplementary Tables 1–4**. For each feature type and each classification algorithm, a curve is plotted with MCC as Y-axis and number of used features as X-axis, as shown in **Figure 3**. The MCC values change with the number of features for each classification algorithm. Clearly, for each feature type, RF outperforms other three classification algorithms.

For one-hot encoded representations of GO/KEGG terms, the corresponding curves are illustrated in **Figure 3A**. The optimum RF classifier yielded the MCC of 0.858, which uses the top 511 features. The corresponding ACC is 0.885 (**Table 2**). The MCCs of the optimum DT and SVM classifiers are 0.763 and 0.832, respectively, and the corresponding ACCs are 0.805 and 0.864. They are all lower than those of the optimum RF classifier. The MCC of the optimum KNN classifier is also 0.858, however, the ACC is only 0.882, lower than that of the optimum RF classifier. The accuracies of 16 categories yielded by four optimum classifiers are shown in **Figure 4A**, further confirming the superiority of RF.

Of the functional embeddings, **Figure 3B** shows the curve for each classification algorithm. It can be observed that the four optimum classifiers yield the MCCs of 0.697, 0.837, 0.762, and 0.876, respectively. The corresponding ACCs are 0.743, 0.860, 0.799, and 0.897 (**Table 2**), respectively. Likewise, RF still yields the best performance. The detailed performance (accuracies on 16 categories) of four optimum classifiers is listed in **Figure 4B**. Again, the optimum RF classifier produces the most high accuracies, indicating the advantage of RF.

For the third feature type (network embeddings), we also plot four curves, one curve corresponds one classification algorithm, as shown in **Figure 3C**. The highest MCCs for four classification algorithms are 0.612, 0.755, 0.618, and 0.803, respectively. Corresponding ACCs are 0.669, 0.786, 0.669, and 0.835 (**Table 2**), respectively. Also, the optimum RF classifier yields the best performance. We further list the accuracies on all categories produced by four optimum classifiers in **Figure 4C**. Clearly, the optimum RF classifier is superior to other optimum classifiers.

As for the last feature type (the combination of network and functional embeddings), four curves are plotted in **Figure 3D**. The optimum RF classifier generates the MCC of 0.872 and ACC of 0.893 (**Table 2**). The optimum KNN classifier yields a

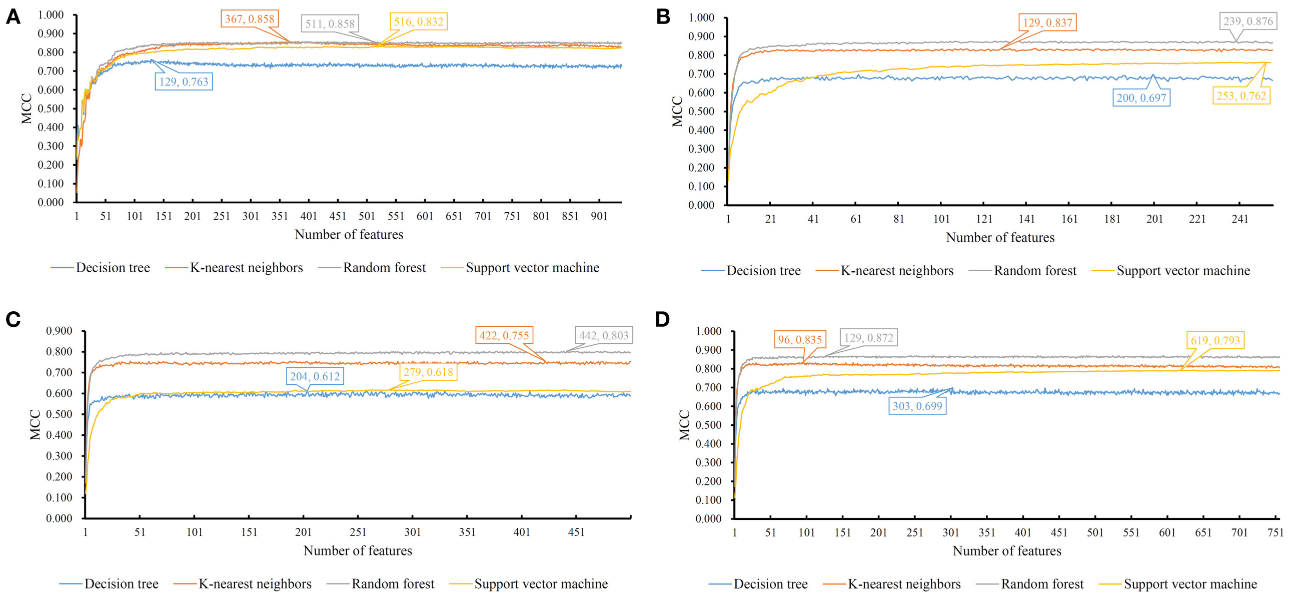


FIGURE 3 | MCC changes with the number of features for IFS with different classification algorithms. **(A)** one-hot encoded representation derived from KEGG/GO terms; **(B)** functional embeddings from KEGG/GO terms; **(C)** network embeddings from a protein-protein network; **(D)** the combined functional and network embeddings.

TABLE 2 | Comparisons of different classifiers with or without feature selection.

Feature type	Classification algorithm	ACC		MCC	
		With feature selection	Without feature selection	With feature selection	Without feature selection
One-hot encoded representations	Decision tree	0.805	0.776	0.763	0.726
	K-nearest neighbors	0.882	0.854	0.858	0.826
	Random forest	0.885	0.878	0.858	0.849
	Support vector machine	0.864	0.859	0.832	0.825
Functional embeddings	Decision tree	0.743	0.717	0.697	0.666
	K-nearest neighbors	0.860	0.852	0.837	0.828
	Random forest	0.897	0.889	0.876	0.867
	Support vector machine	0.799	0.798	0.762	0.760
Network embeddings	Decision tree	0.669	0.648	0.612	0.588
	K-nearest neighbors	0.786	0.785	0.755	0.754
	Random forest	0.835	0.827	0.803	0.795
	Support vector machine	0.669	0.661	0.618	0.609
Functional and network embeddings	Decision tree	0.746	0.720	0.699	0.670
	K-nearest neighbors	0.858	0.832	0.835	0.805
	Random forest	0.893	0.884	0.872	0.861
	Support vector machine	0.825	0.823	0.793	0.791

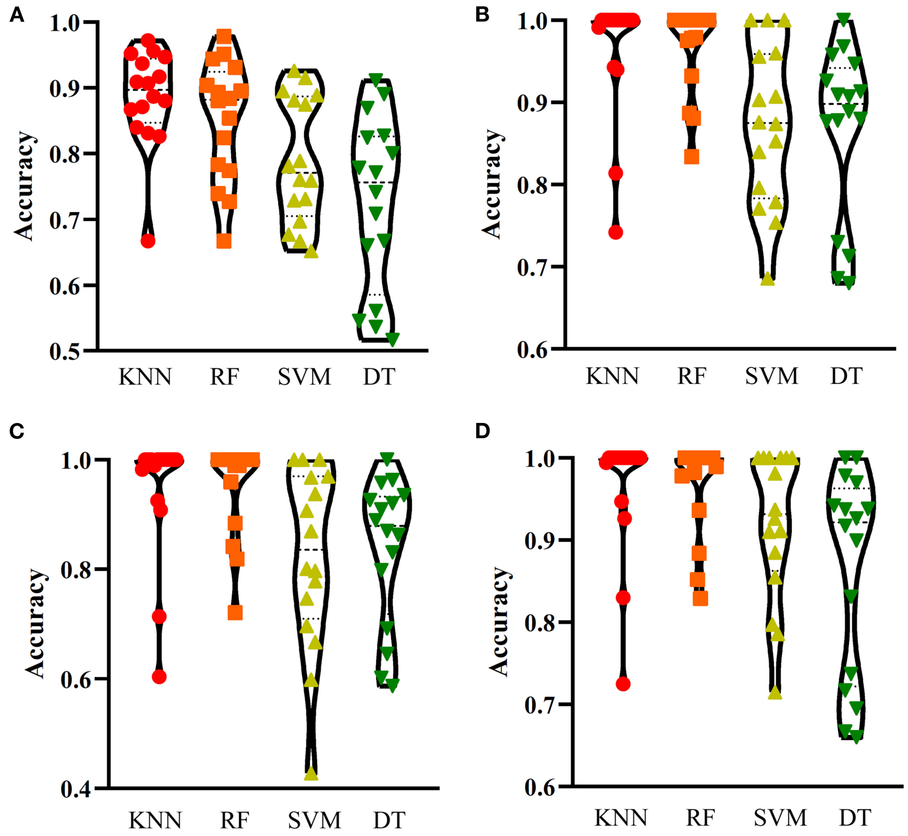


FIGURE 4 | Performance of the optimum classifiers on 16 categories with different feature types. **(A)** one-hot encoded representation derived from KEGG/GO terms; **(B)** functional embeddings from KEGG/GO terms; **(C)** network embeddings from a protein-protein network; **(D)** the combined functional and network embeddings. KNN, K-nearest neighbors; RF, random forest; SVM, support vector machine; DT, decision tree.

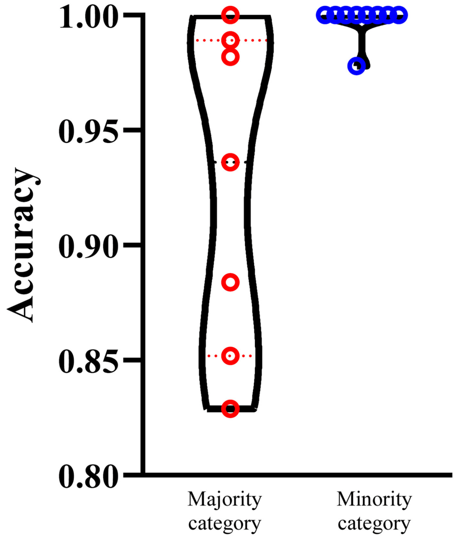


FIGURE 5 | Performance of the optimum RF classifiers on majority and minority categories using the combined functional and network embeddings. The majority categories contain more than 100 proteins, whereas the minority categories consist of <100 proteins. The performance on minority categories is not lower than that on the majority categories.

high MCC of 0.835. However, the other two optimum classifiers produce much lower MCC (lower than 0.800). The ACCs shows the same results (see **Table 2**). Accuracies on all categories are shown in **Figure 4D**. Similarly, the optimum RF classifier provides the best performance.

As mentioned above, the optimum RF classifier is all best for four different feature types. The optimum RF classifier on functional embeddings derived from KEGG/GO terms yields the best MCC value (0.876). This classifier is based on the top 239 features. The optimum RF classifier with the combined embeddings only yields an MCC value of 0.872 and is a little worse than the RF with only functional embeddings. However, it uses only the top 129 features, which is nearly half of the number of features of the optimum RF classifier with only functional embeddings (239). Thus, in this study, we use the combined network and functional embeddings as the final input features.

We select the optimum RF classifier with the combined embeddings as the proposed method. As the sizes of 16 categories are of great difference, it is necessary to investigate the performance of such classifier on majority and minority categories. We set 100 as the threshold, that is, categories containing more than 100 proteins are deemed as majority categories, whereas other categories are termed as minority categories. In this case, we obtain seven majority categories

TABLE 3 | Performance of BLAST, DeepLoc, and our proposed method.

Class	BLAST	DeepLoc	Ours
Biological membrane	0.843	–	0.829
Cell periphery	0.424	–	1.000
Cytoplasm	0.455	–	0.852
Cytoplasmic vesicle	0.232	–	1.000
Endoplasmic reticulum	0.532	–	0.989
Endosome	0.280	–	1.000
Extracellular space or cell surface	0.739	–	0.936
Flagellum or cilium	0.000	–	1.000
Golgi apparatus	0.379	–	1.000
Microtubule cytoskeleton	0.333	–	1.000
Mitochondrion	0.356	–	0.982
Nuclear periphery	0.097	–	1.000
Nucleolus	0.241	–	1.000
Nucleus	0.733	–	0.884
Peroxisome	0.289	–	0.978
Vacuole	0.333	–	1.000
Overall accuracy	0.660	0.659	0.893
MCC	0.576	0.568	0.872

and nine minority categories. The performance on majority and minority category of the proposed classifier is shown in **Figure 5**. It is surprising that the performance on minority categories is not lower than that on the majority categories. This result indicates that the performance of such classifier is not influenced by the imbalanced problem after SMOTE is applied.

Comparison of Classifiers With or Without Feature Selection

In this study, we employed a feature selection procedure to improve the performance of different classification algorithms. **Table 2** lists the performance of different classification algorithms on four feature types with or without feature selection. It can be observed that the performance of DT is enhanced most by the feature selection. MCC is improved about 3% and ACC is enhanced about 2.5%. The improvement on the performance of KNN yielded by feature selection is quite different for different feature types. For one-hot encoded representations and combined functional and network embeddings, the performance is evidently enhanced, while the performance is improved limited for other two feature types. As for other two classification algorithms (RF and SVM), the improvement is not very evident (almost all <1% for both ACC and MCC). Anyway, it can be confirmed that the employment of feature selection can improve the performance of all classification algorithms.

Proposed Method Is Superior to State-of-the-Art Methods

To demonstrate the power of our proposed method, we compare our method with published methods, including BLAST and Deeploc. The results are listed in **Table 3**. BLAST and DeepLoc nearly have the same level of performance and are inferior to our proposed method. Of the 16 locations, our method can achieve 100% accuracy on nine locations. Here, DeepLoc has the worst performance, and a potential reason is that our benchmark dataset is heavily imbalanced and results in biased preference for majority classes. To resolve the data imbalance issue, our method applies SMOTE in the construction of a balanced training set.

CONCLUSION

In this study, we present an embedding-based method to predict protein subcellular locations by integrating protein interactions and functional information. The proposed method first learns network embeddings from a protein–protein network and functional embeddings from associations between proteins and GO/KEGG terms. We demonstrate that our proposed method is superior to state-of-the-art methods, and the learned embeddings offer valuable biological insights.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the (Swiss-Prot) (http://cn.expasy.org/).

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. XP, HL, and TZ performed the experiments. XP, HL, ZL, and LC analyzed the results. XP and HL wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), National Key R&D Program of China (2017YFC1201200), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.626500/full#supplementary-material

Supplementary Table 1 | The performance of four classifiers change with the number of one-hot encoded representation derived from KEGG/GO terms.

Supplementary Table 2 | The performance of four classifiers change with the number of functional embeddings from KEGG/GO terms.

Supplementary Table 3 | The performance of four classifiers change with the number of network embeddings from a protein-protein network.

Supplementary Table 4 | The performance of four classifiers change with the number of combined functional embeddings and network embeddings.

REFERENCES

Almagro Armenteros, J. J., Sonderby, C. K., Sonderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 3387–3395. doi: 10.1093/bioinformatics/btx431

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, L., Li, Z., Zeng, T., Zhang, Y.-H., Liu, D., Li, H., et al. (2020). Identifying robust microbiota signatures and interpretable rules to distinguish cancer subtypes. *Front. Mol. Biosci.* 7:604794. doi: 10.3389/fmolb.2020.604794

Chen, L., Pan, X., Hu, X., Zhang, Y.-H., Wang, S., Huang, T., et al. (2018a). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143, 1731–1740. doi: 10.1002/ijc.31554

Chen, L., Zhang, S., Pan, X., Hu, X., Zhang, Y.-H., Yuan, F., et al. (2018b). HIV infection alters the human epigenetic landscape. *Gene Ther.* 26, 29–39. doi: 10.1038/s41434-018-0051-6

Chou, K. C., and Cai, Y. D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* 277, 45765–45769. doi: 10.1074/jbc.M204161200

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machi. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transact. Inform. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Goldberg, T., Hamp, T., and Rost, B. (2012). LocTree2 predicts localization for all domains of life. *Bioinformatics* 28, i458–i465. doi: 10.1093/bioinformatics/bts390

Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., et al. (2014). LocTree3 prediction of localization. *Nucleic Acids Res.* 42, W350–355. doi: 10.1093/nar/gku396

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Grover, A., and Leskovec, J. (2016). "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA: ACM).

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-Based Machine Learning Model for Predicting the Metabolic Pathways of Compounds. *IEEE Access* 8, 130687–130696. doi: 10.1109/ACCESS.2020.3009439

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence* (Mahwah, NJ: Lawrence Erlbaum Associates Ltd.), 1137–1145.

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J. Stat. Softw.* 36, 1–13. doi: 10.18637/jss.v036.i11

Li, B.-Q., Huang, T., Chen, L., Feng, K.-Y., and Cai, Y.-D. (2014). "Prediction of human protein subcellular locations with feature selection and analysis," in *Frontiers in Protein and Peptide Sciences*, ed B. M. Dunn (Soest: Bentham Science Publishers), 206–225.

Li, J., Lu, L., Zhang, Y., Liu, M., Chen, L., Huang, T., et al. (2019). Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell. Biochem.* 120, 405–416. doi: 10.1002/jcb.27395

Li, M., Pan, X. Y., Zeng, T., Zhang, Y. H., Feng, K. Y., Chen, L., et al. (2020). Alternative polyadenylation modification patterns reveal essential posttranscription regulatory mechanisms of tumorigenesis in multiple tumor types. *Biomed. Res. Int.* 2020:6384120. doi: 10.1155/2020/6384120

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of drug side effects with a refined negative sample selection strategy. *Comput. Math. Methods Med.* 2020:1573543. doi: 10.1155/2020/1573543

Liu, H., Hu, B., Chen, L., and Lu, L. (2020). Identifying protein subcellular location with embedding features learned from networks. *Curr. Proteom.* doi: 10.2174/1570164617999201124142950

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230. doi: 10.1023/A:1008363719778

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations* (Scottsdale, AZ).

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34. doi: 10.1093/nar/27.1.29

Pan, X., Lu, L., and Cai, Y. D. (2020b). Predicting protein subcellular location with network embedding and enrichment features. *Biochim. Biophys. Acta Proteins Proteom.* 1868:140477. doi: 10.1016/j.bbapap.2020.140477

Pan, X. Y., Zeng, T., Zhang, Y. H., Chen, L., Feng, K. Y., Huang, T., et al. (2020a). Investigation and prediction of human interactome based on quantitative features. *Front. Bioeng. Biotechnol.* 8, 730. doi: 10.3389/fbioe.2020.00730

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transact. Pattern Anal. Mach. Intell.* 1226–1238. doi: 10.1109/TPAMI.2005.159

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transact. Syst. Man Cybernet.* 21, 660–674. doi: 10.1109/21.97458

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Wang, S., Zhang, Q., Lu, J., and Cai, Y.-D. (2018). Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm. *Curr Bioinform.* 13, 3–13. doi: 10.2174/1574893611666160608075753

Zhang, S., Pan, X., Zeng, T., Guo, W., Gan, Z., Zhang, Y.-H., et al. (2019). Copy number variation pattern for discriminating MACROD2 states of colorectal cancer subtypes. *Front. Bioeng. Biotechnol.* 7:407. doi: 10.3389/fbioe.2019.00407

Zhang, S. Q., Zeng, T., Hu, B., Zhang, Y. H., Feng, K. Y., Chen, L., et al. (2020). Discriminating origin tissues of tumor cell lines by methylation signatures and dys-methylated rules. *Front. Bioeng. Biotechnol.* 8:507. doi: 10.3389/fbioe.2020.00507

Zhou, H., Yang, Y., and Shen, H. B. (2017). Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics* 33, 843–853. doi: 10.1093/bioinformatics/btw723

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020). iATC-FRAKEL: a simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. *Bioinformatics* 36, 3568–3569. doi: 10.1093/bioinformatics/btaa166

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DeCban: Prediction of circRNA-RBP Interaction Sites by Using Double Embeddings and Cross-Branch Attention Networks

Liangliang Yuan[1] *and Yang Yang*[1,2*]

[1] *Department of Computer Science and Engineering, Center for Brain-Like Computing and Machine Intelligence, Shanghai Jiao Tong University, Shanghai, China,* [2] *Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai, China*

Circular RNAs (circRNAs), as a rising star in the RNA world, play important roles in various biological processes. Understanding the interactions between circRNAs and RNA binding proteins (RBPs) can help reveal the functions of circRNAs. For the past decade, the emergence of high-throughput experimental data, like CLIP-Seq, has made the computational identification of RNA-protein interactions (RPIs) possible based on machine learning methods. However, as the underlying mechanisms of RPIs have not been fully understood yet and the information sources of circRNAs are limited, the computational tools for predicting circRNA-RBP interactions have been very few. In this study, we propose a deep learning method to identify circRNA-RBP interactions, called DeCban, which is featured by hybrid double embeddings for representing RNA sequences and a cross-branch attention neural network for classification. To capture more information from RNA sequences, the double embeddings include pre-trained embedding vectors for both RNA segments and their converted amino acids. Meanwhile, the cross-branch attention network aims to address the learning of very long sequences by integrating features of different scales and focusing on important information. The experimental results on 37 benchmark datasets show that both double embeddings and the cross-branch attention model contribute to the improvement of performance. DeCban outperforms the mainstream deep learning-based methods on not only prediction accuracy but also computational efficiency. The data sets and source code of this study are freely available at: https://github.com/AaronYll/DECban.

Keywords: circular RNAs, RNA binding proteins, deep learning, double embeddings, attention network

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Xiangxiang Zeng,
Hunan University, China
Qi Zhao,
University of Science and Technology Liaoning, China

****Correspondence:***
Yang Yang
yangyang@cs.sjtu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *24 November 2020*
Accepted: *23 December 2020*
Published: *22 January 2021*

Citation:
Yuan L and Yang Y (2021) DeCban: Prediction of circRNA-RBP Interaction Sites by Using Double Embeddings and Cross-Branch Attention Networks. Front. Genet. 11:632861. doi: 10.3389/fgene.2020.632861

1. INTRODUCTION

Circular RNAs (circRNAs) are a special kind of non-coding RNA molecules. Different from linear RNAs, circRNA molecules have closed-ring structures, which are not affected by RNA exonuclease, and their expression is more stable (Pamudurti et al., 2017; Li et al., 2018). Although natural circRNAs were discovered more than two decades ago, their important roles in gene regulation and disease development have just been revealed in recent years (Hansen et al., 2013; Li et al., 2015).

Emerging studies have shown that circRNAs can bind to various types of proteins to affect protein localization, regulate protein expression, or influence protein-protein-interactions. The circRNA-binding-proteins (circRBPs) include transcription factors, RNA processing proteins, proteases, and common RNA-binding-proteins (RBPs) that can be bound with linear RNAs. Understanding the interactions between circRNAs and proteins is helpful for revealing the biological functions of circRNAs (Du et al., 2017; Zang et al., 2020). For the past decade, high-throughput experimental technologies have been widely used to detect the interactions between RNAs and proteins, like cross-linking and immunoprecipitation followed by RNA sequencing (CLIP-Seq) (Yang et al., 2015). The large-scale experimental data makes it possible to predict RNA-protein interactions (RPIs) based on machine learning methods (Li et al., 2013). Compared with expensive and time-consuming wet experiments, the computational methods have considerably sped up the identification of interactions, thus the automatic prediction of RPI has been a hot topic in the bioinformatics field (Pan et al., 2019).

The existing prediction tools include both RNA-oriented or protein-oriented, i.e., identifying the binding sites in the RNA chain and protein chain, respectively (Yan et al., 2016). Benefitting from the abundant domain knowledge from protein databases, many studies perform prediction based on protein information. By contrast, much fewer studies focus on the binding sites on circRNAs (Ju et al., 2019; Zhang et al., 2019; Jia et al., 2020; Wang and Lei, 2020). The reasons are two-folds. For one thing, compared with other non-coding RNAs, like microRNAs and long non-coding RNAs, research on circRNAs has been largely lagged and their data is scarce. For another thing, the prediction for circRNAs is a very difficult task, due to the long sequences, sparsely distributed binding sites and limited information that could be extracted.

As circRNAs have attracted more and more attention, experimental data of circRNAs has increased rapidly. Till now, a lot of circRNA-protein interactions have been revealed and released in public databases, e.g., CircInterome that houses the RBP/miRNA-binding sites on human circRNAs (Dudekula et al., 2016). Thanks to the fast-growing circRNA data and the rise of deep learning, methods for predicting circRNA-RBP binding sites are emerging. For instance, Zhang et al. (2019) proposed a method called CRIP to predict circRNA-RBP binding sites, which is a hybrid architecture of convolutional neural networks (CNNs) and recurrent neural networks (RNNs); Jia et al. (2020) proposed an ensemble classifier, PASSION, which combines various statistical sequence features and performs feature selection to enhance the prediction accuracy.

Note that learning long sequences has still been an open problem for neural networks. Biological sequences are much longer than natural language sentences, conventional learning models, including long short-term memory networks (LSTMs) which were designed to handle long-term dependencies (Hochreiter and Schmidhuber, 1997), do not work well for extremely long sequences. Therefore, most of the existing predictors take short segments instead of full-length non-coding RNAs as input to identify the binding sites (Pan and Shen, 2017, 2018; Pan et al., 2018; Zhang et al., 2019), i.e., they divid the RNA sequences into short fragments and predict whether a fragment is a binding site or not. Obviously, such simplification does not accord with the real scenario. For one thing, RPIs are usually determined by the full-length RNA information rather than short fragments; and for the other thing, the binding regions only make up a tiny proportion in the whole RNA sequences, while the fragment-based prediction often constructs relatively balanced datasets, leading to a high false-positive-rate. Therefore, to address the sparse distribution of binding sites and reduce false positive predictions, this study aims to develop a model which allows full-length circRNA sequences as input and provides reliable predictions.

Generally, the performance of machine learning methods depends on two factors, namely feature extraction and learning model. In traditional learning methods, RNA sequences are represented by statistical features, like the frequency of *k*-mers and secondary structure elements (Zhang et al., 2011; Chen et al., 2014). With the rise of deep learning, hand-crafted feature extraction has been largely replaced by automatic feature learning and pre-training via large-scale unlabeled datasets (Clauwaert and Waegeman, 2019; Meher et al., 2019). Word embedding is an emerging technique for representing biological sequence features. Unlike traditional features or one-hot encoding, word embedding is a kind of continuous distributed features. Commonly used word embedding methods include Word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), ELMo (Peters et al., 2018), GPT (Radford et al., 2018), and Bert (Devlin et al., 2018). The first two models yield static embedding, i.e., the embedding vector for each word is context-independent and fixed after training (Peters et al., 2018), while the latter three methods yield context-dependent embedding vectors.

At present, static embeddings learned by shallow models have been widely used in biological sequence analysis, while only a few studies applied dynamic embedding, like Elmo and Bert. One reason is that the models based on deep learning models such as Elmo and Bert are very computation-intensive. Especially, non-coding RNA sequences are much longer than protein sequences, thus learning dynamic embedding for RNA sequences may require more complex model. In this study, we also adopt static word embedding method to represent circRNA sequences. To better mine the sequence information, we propose a double-embedding method to expand the feature space, which is further learned by deep neural networks to extract abstract features for classification.

As circRNAs are usually thousands of nucleotides, to handle the extremely long sequences, specialized model design is also required. Previous studies mainly used CNN (Alipanahi et al., 2015), RNN, or CNN-RNN hybrid models (Pan and Shen, 2017; Zhang et al., 2019). As aforementioned, these models take short fragments as input and construct balanced datasets, while true binding sites are very rare. In this study, we design a new model called DeCban (Double embedding and Cross-branch attention network) to predict the presence of RBP-binding sites on full-length circRNAs. This predictor is featured by not only a new sequence encoding scheme, i.e., double embedding, but also a cross-branch attention neural network. The network extracts sequence features of different abstract levels and different granularities, and the attention module allows the network to focus on important features for discrimination.

TABLE 1 | Experimental datasets.

RBP	Train#	Test#	RBP	Train#	Test#
AGO1	33547	14377	IGF2BP2	59467	25485
AGO2	57697	24724	IGF2BP3	83120	35622
AGO3	8570	3672	LIN28A	50769	21757
ALKBH5	4497	1927	LIN28B	21601	9257
AUF1	3045	1305	METTL3	9033	3871
C17ORF85	6225	2667	MOV10	6309	2703
C22ORF28	15680	6720	PTB	67963	29127
CAPRIN1	15503	6643	PUM2	4903	2101
DGCR8	57651	24707	QKI	3036	1300
EIF4A3	25017	10721	SFRS1	36563	15669
EWSR1	13253	5679	TAF15	3580	1534
FMRP	79392	34024	TDP43	2610	1118
FOX2	2756	1180	TIA1	5127	2197
FUS	60699	26013	TIAL1	9613	4119
FXR1	2908	1246	TNRC6	3876	1660
FXR2	15400	6600	U2AF65	16236	6958
HNRNPC	2588	1108	WTAP	1517	649
HUR	73352	31436	ZC3H7B	30175	12931
IGF2BP1	66355	28437			

Compared with the existing RPI prediction tools and mainstream deep learning models, DeCban has great advantages on both prediction accuracy and computational efficiency.

2. METHODOLOGY

2.1. Datasets

To evaluate the prediction performance of DeCban, we collect circRNAs and their interacting proteins from Circular RNA Interactome (https://circinteractome.nia.nih.gov/) (Dudekula et al., 2016). The sequence redundancy is removed by CD-Hit (Fu et al., 2012) with threshold 0.8, resulting into 32,216 circRNA sequences, which are bound to a total of 37 RBPs. We train a binary prediction model for each RBP and construct 37 datasets. The positive-to-negative ratio of each data set is 1:1, where the positive samples are the circRNAs binding to the RBP and negative samples are the remaining ones. The circRNAs in this set range from 100 to 30,000 nt in length, 90% of which are 500~7,000 nt. Therefore, to avoid the potential bias brought by too short and too long sequences, we only include the sequences falling in the range of 500~7,000 nt in the final data set. The data statistics are shown in **Table 1**.

2.2. Model Architecture

Figure 1 shows the model architecture. The feature vectors generated by double embeddings are fed into a CNN-based neural network with multiple branches of different granularities. We introduce the self-attention mechanism to automatically integrate the semantic information extracted from different branches at each abstract level (an abstract level corresponds to a convolutional layer), and combine multiple levels of semantic information to determine whether binding sites exist in the RNA equences.

2.2.1. Double Embeddings

To work with deep neural networks (DNNs), input sequences are usually converted into numerical vectors by encoding schemes, such as one-hot, which encodes each nucleotide by a four-dimensional binary vector with only one element equal to 1. One-hot is unable to express the association between different nucleotides or context information, and the low dimensionality of its feature space limits the performance of further learning by DNN. By contrast, word embeddings, that are continuous dense vectors capturing semantic association of words, have been a mainstream method to represent words and sentences in natural language processing. The training of word embeddings is based on the language modeling task, like next-word prediction, which does not require sequence labels. Thus, the training of embeddings can be performed on large-scale unlabeled corpus.

In recent years, word embeddings for *k*-mers have emerged in various bioinformatics applications. Here we also adopt word embeddings to represent circRNA sequence features. Besides, we notice that the word embedding technology has been applied more and achieved better performance in protein classification tasks, perhaps due to the bigger alphabet size and much shorter length of amino acid sequences compared with DNA/RNA sequences. To expand the alphabet, Zhang et al. (2019) developed a codon-based encoding scheme for circRNA sequences. A major advantage of this scheme lies in the enlarged feature space, as the classic one-hot has only 4 symbols while the codon-based encoding has 21 symbols, which are a combination of 3 nucleotides. The genetic codes define not only the alphabet of the new symbol system, but also the rules of correspondence between combinations of nucleotides and new symbols. Zhang et al. (2019) also showed that the three-nucleotide combinations defined by codons, are superior to random combinations defined in other encoding systems. Inspired by this idea, we convert RNA segments into pseudo-peptides and obtain word embeddings for them (we call them "pseudo-" because them are not real peptides). Then, we combine the two kinds of embeddings to generate the input features of our model. We call the new feature extraction method as double embeddings.

For a circRNA fragment of length *k*, there are $(k - 2)$ consecutive codons, where the codons are translated in an overlapping manner to retain more local context information. Then we perform pre-training of the word embeddings for *k*-mer RNA segments and $(k - 2)$-mer peptides, respectively. Since circRNA sequences are very long, to reduce the length of sentences, we need to set a large *k*, and long fragments also contain more local sequence information. However, training long words will require intensive computation resource. As a tradeoff, we set *k* to 7. We treat the segmented *k*-mers as words and adopt the GloVe algorithm to train their embeddings. Like NLP applications, to produce good embedding vector for words, a large corpus of text is required. Here we adopt the whole human genome as the corpus for RNA sequences (we replace "T" with "U" to convert DNAs to RNAs) and UniRef50 (https://www.uniprot.org/help/uniref) as the corpus for amino acid sequences.

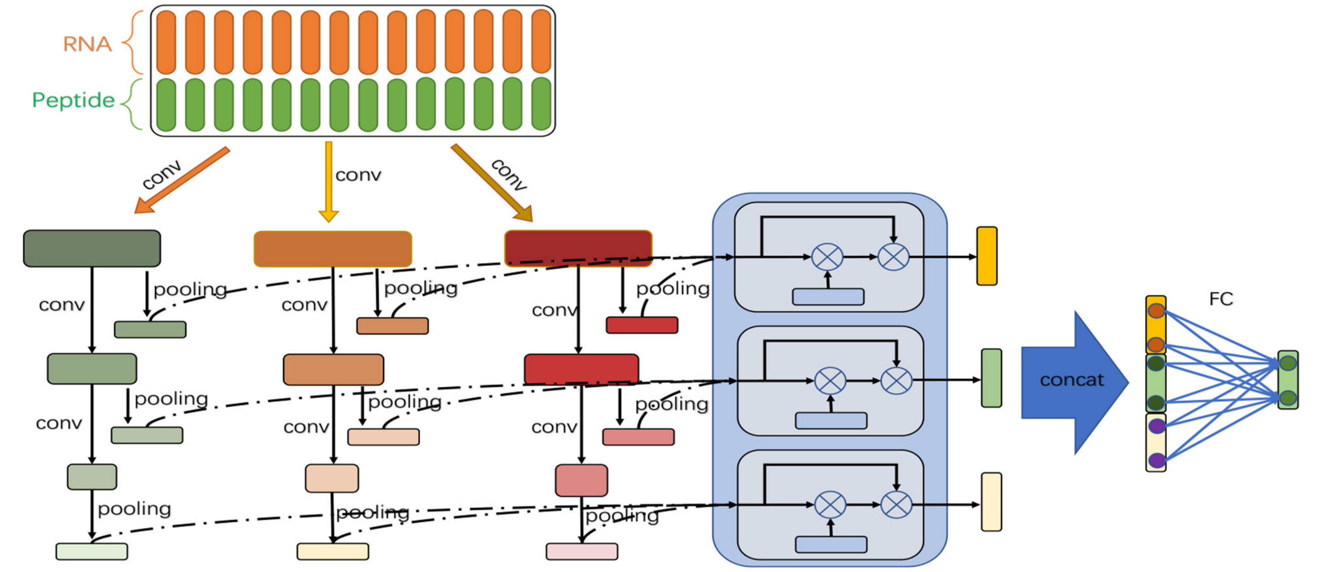


FIGURE 1 | Model architecture of DeCban. The network consists of three convolutional layers and three branches (shown in green, orange, and red, respectively). An attention layer (shown in light blue) is used to integrate the outputs of the three branches. Then, the feature embeddings learned by the three layers are concatenated and fed to a fully connected layer to yield the final output.

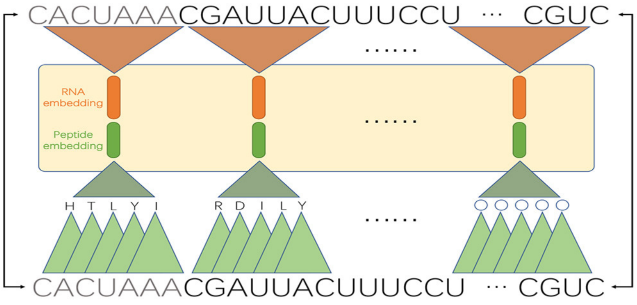


FIGURE 2 | An example of double embeddings. An RNA sequence is segmented into 7-mers, and each 7-mer is converted into an embedding vector; meanwhile, the 7-mer is mapped to a pseudo-peptide, which is also converted into an embedding vector. The two embedding vectors are concatenated as a whole input.

Finally, we construct the input matrix by using pre-trained word embeddings. Specifically, for each 7-mer fragment of a circRNA sequence, we concatenate the RNA embedding and the corresponding pseudo-peptide embedding. For example, as shown in **Figure 2**, the first 7-mer "CACUAUA" contains the codons CAC, ACU, CUA, UAU, and AUA, which encode the amino acids H, T, L, Y, and I, respectively. Then, the embedding vectors of "CACUAUA" and "HTLYI" are concatenated to represent the feature vector of "CACUAUA."

Formally, for a given circRNA, let its length be L, which is divided into m segments $(m = \lfloor L/k \rfloor\})$. Let the RNA and peptide embedding vectors for w_i are R_i and P_i, whose dimensions are p and q, respectively. Then the double embedding for w_i is defined as,

$$D_i = R_i \oplus P_i, i \in \{1, 2, \cdots, m\}, \tag{1}$$

where $\oplus$ denotes the concatenation operation. Then the circRNA is represented by a matrix of size $(p + q) \times m$, i.e., $[D_1, D_2, \cdots, D_m]$.

2.2.2. Cross-Branch Attention Network

As shown in **Figure 1**, the network has multiple branches, which have the same number of convolutional layers but vary in convolution kernel size. Thus, the branches can extract features at different granularities.

Besides, at the same layer of all branches, we introduce the self-attention mechanism. As the length of the input sequences varies greatly, the best features extracted from different sequences may come from different branches. The self-attention module enables the network to assign weights to the branches and obtain weighted average features. We introduce such modules in each layer to extract features of different abstract levels. Therefore, we name the model cross-branch attention network.

Formally, let the input of the network be X, and the first layer outputs of the three branches be X_1^1, X_2^1, and X_3^1, respectively, which can be expressed as,

$$X_j^1 = f(W_j^1 * X + b_j^1), j \in \{1, 2, 3\}. \tag{2}$$

Similarly, for each subsequent layer i, the outputs X_j^i are computed as,

$$X_j^i = f(W_j^{i-1} * X_j^{i-1} + b_j^i), i \in \{2, 3\}, j \in \{1, 2, 3\}. \tag{3}$$

The X_j^is are further processed via a maximum pooling operation, i.e.,

$$Y_j^i = h(X_j^i), \tag{4}$$

where $h(\cdot)$ is the max-pooling function. Then, the self-attention module works on each layer to integrate the outputs of three branches,

$$Y^i_{attn}(W_a, Y^i_{1:3}) = SoftMax(g(W_a * (Y^i_{1:3})^T)) * Y^i_{1:3}, \tag{5}$$

where $g(\cdot)$ denotes the activation function, and Y^i_{attn} is the output yielded by the attention module for the i-th layer. The outputs of the three layers are combined as y_{out},

$$Y_{out} = concat(Y^1_{attn}, Y^2_{attn}, Y^3_{attn}). \tag{6}$$

Finally, the output O of the network is obtained through a FC layer,

$$O = SoftMax(g(W_{fc} * Y_{out} + b_{fc})). \tag{7}$$

3. EXPERIMENTAL RESULTS

3.1. Experimental Settings

The DeCban model has three branches, and the sizes of their convolution kernels are 3, 5, 7, respectively. Each branch has three convolutional layers and each layer has 100 filters. The initial parameters of each attention module are randomly generated with normal distribution. We use Adam optimizer with learning rate of 0.001 to optimize the model. The number of early stopping rounds is set to 10, and the training-to-test ratio is 7:3.

3.2. Baseline Methods

To assess the performance of DeCban, we compare it with not only the existing predictor for RNA-protein interactions but also mainstream deep neural networks, including a recent method called CRIP (Zhang et al., 2019), recurrent neural networks (RNNs), and convolutional neural networks (CNNs). Note that CRIP performs prediction on short fragments (i.e., 101-nt), thus for a full-length RNA sequence, we first divide it into fragments and use CRIP to predict for each fragment, and then merge the results to get the prediction for the whole sequence. The other baseline models fall into five groups. Each group contains three methods with the same backbone model but different feature representations, namely RNA embeddings, peptide embeddings, and double embeddings. In addition, the performance of DeCban working with RNA or peptide embeddings alone is evaluated. The specification of baseline models is as follows.

- Group 1—LSTM: a vanilla long short-term memory network (Hochreiter and Schmidhuber, 1997).
- Group 2—BiLSTM with attention: a bidirectional LSTM with attention mechanism (Zhou et al., 2016)[1].
- Group 3—TextCNN: a TextCNN (Kim, 2014) model.
- Group 4—ResNet18 base: a basic ResNet18 model (He et al., 2016).
- Group 5—ResNet18 small: a simplified ResNet18 model, which has the same architecture as ResNet18 but fewer convolutional kernels on each layer.

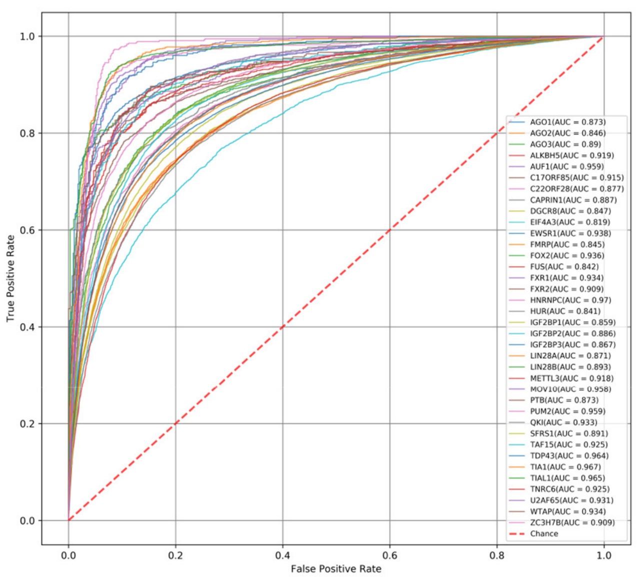


FIGURE 3 | The ROC curves obtained by DeCban for 37 circRNA data sets.

- CRIP: a CNN-RNN hybrid model for the prediction of RBP-bindings sites on RNAs (Zhang et al., 2019).

3.3. Experimental Results and Analysis

For a comprehensive comparison, we consider not only the prediction accuracy but also computational efficiency. The accuracy is evaluated by the common metrics of machine learning models, F_1 and AUC score (Area under the ROC Curve). The efficiency is assessed by the number of parameters and speedup.

First, we compare the AUC scores of DeCban and CRIP on all 37 data sets. The ROC curves are shown in **Figures 3**, **4**, respectively. The AUC scores range from 0.819 to 0.970, and the average AUC is 0.905. The lowest, highest, and average AUCs of these two methods are 0.819 vs. 0.734, 0.970 vs. 0.917, and 0.905 vs. 0.821, respectively. DeCban has an obvious advantage over CRIP.

Second, we compare the F_1 scores for all baseline models. **Table 2** shows the average F_1, number of parameters and speedup. As can be seen, DeCban achieves the highest average F_1 of 0.841, and the second best model is BiLSTM with attention, whose average F_1 is 0.827. The detailed scores for all 37 data sets are listed in **Supplementary Table 1**. DeCban obtains the highest F_1 scores on all of the datasets. Meanwhile, DeCban has a lightweight architecture. Compared with the second best model BiLSTM, DeCban has a significant reduction on model parameters. The detailed comparison results are discussed in sections 3.3.1–3.3.5.

3.3.1. Comparison of the Sequence Encoding Methods

From **Table 2**, it can be observed that double embeddings can improve the performance of both baseline models and DeCban.

[1] An advanced model structure based on LSTM, which has achieved state of art results on multiple NLP tasks.

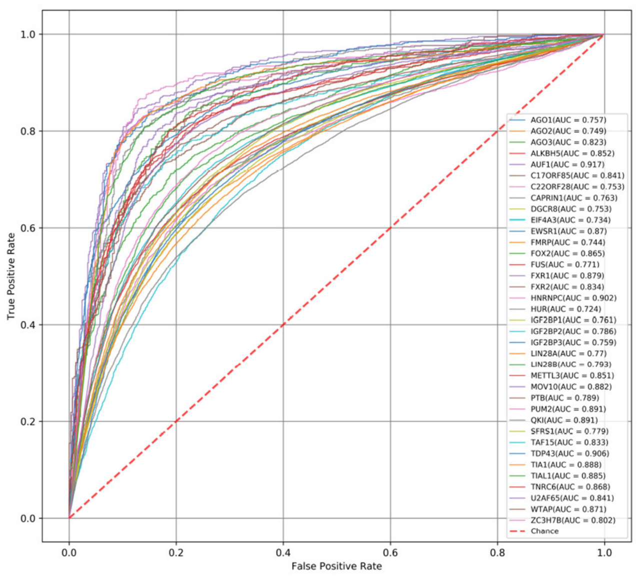


FIGURE 4 | The ROC curves obtained by CRIP for 37 circRNA data sets.

TABLE 2 | Experimental results of different models[a].

Model		**Param[b]**	**Avg F[a]**	**Speedup[c]**
LSTM-base	RNA	118 K	0.685	1.8x
	Peptide	132 K	0.685	2.0x
	Double	183 K	0.692	3.3x
BiLSTM-attention	RNA	647 K	0.817	6.4x
	Peptide	676 K	0.810	6.4x
	Double	778 K	0.827	8.2x
CNN-base	RNA	26 K	0.796	1.0x
	Peptide	30 K	0.793	1.2x
	Double	46 K	0.806	2.3x
ResNet-18-base	RNA	3,914 K	0.811	2.7x
	Peptide	3,927 K	0.803	2.6x
	Double	3,972 K	0.814	3.7x
ResNet-18-small	RNA	254 K	0.770	1.7x
	Peptide	255 K	0.761	1.8x
	Double	261 K	0.773	2.7x
CRIP	–	900 K	0.766	5.7x
DeCban	One-hot	33 K	0.822	9.6x
	RNA	79 K	0.833	1.8x
	Peptide	93 K	0.826	2.0x
	Double	141 K	0.841	3.2x

[a]RNA, Peptide, Double denote the RNA embedding, Peptide embedding, and double embedding, respectively.
[b]Param denotes the number of parameters in the model.
[c]Speedup measures the relative performance of two methods processing the same problem in terms of speed. We use CNN-base with RNA embedding as the basic reference, i.e., its speedup is 1.0x.

Compared to original RNA embeddings, double embeddings increase F_1 by around 1%. In the meantime, using double embeddings do not significantly increase the complexity of the model. The total number of parameters of DeCban using double embeddings has the same order of magnitude as that of the model with RNA embeddings or amino acid embeddings. Taking ResNet-18-base as an example, the number of parameters is increased by <1.5%, while the average F_1 on 37 data sets is increased by nearly 1 percentage.

The results suggest that the combination of RNA information and pseudo-peptide information can improve the data representation ability, although the "peptides" are not biological meaningful. A major reason for the performance improvement is the enlarged feature space. Moreover, the new encoding method traverses the RNA *k*-mers sequentially in an overlapping manner, thus retaining some local context information, which may be helpful for capturing the dependency relationship of nucleotides.

In addition, we replace the double embedding encoding with the traditional one-hot encoding for comparison. The average F_1 on 37 data sets is 0.822, and the training speed is significantly slower than double embedding. This result shows the advantages of double embedding over traditional one-hot encoding.

3.3.2. Comparison of Model Architectures

As DeCban is a convolutional neural network, we compare it with the state-of-the-art CNN model, ResNet-18. The number of layers and parameters of ResNet is much larger than that of DeCban. Specifically, the parameter amount of ResNet-18-base is 28 times of DeCban, while the F_1 score is 2.5% lower than DeCban. Considering that ResNet might overfit the data due to the large model size, we implement a lightweight version of ResNet-18, namely the ResNet-18-small, by reducing the number of convolutional kernels for each layer, then the amount of parameters is at the same order of magnitude as DeCban. However, after the simplification, the prediction accuracy drops significantly. Comparing with ResNet-18-base, the F_1 scores of three embedding methods are decreased by 0.038, 0.043, and 0.044, respectively. By contrast, benefitting from the multi-branch and self-attention mechanism, DeCban can extract features of different scales, and achieve better accuracy with much higher efficiency. Even using only RNA word embeddings, DeCban outperforms all baseline models, demonstrating the superiority of the new model architecture.

Besides CNN models, we also consider the widely-used RNN model, LSTM. Although LSTM was designed to address the gradient vanishing issue and long-term dependencies, it is still difficult for LSTM to handle very long sequences. It can be seen from the experimental results that the performance of vanilla LSTM is poor. When using the double embeddings, the average F_1 on 37 datasets is 0.692, which is much lower than that of basic CNN (0.806). The gap of performance between these two kinds of models may be attributed to the large difference in the sequence length.

As the input sequences vary greatly in length, a large number of meaningless zeros are filled at the end of short sequences.

The padding operation affects the training of LSTM, while CNN has more flexibility in extracting features from sequences with varying length. In this case, attention mechanism becomes a necessary part to enable the model focus on informative regions, thus the BiLSTM model with attention improves the performance of LSTM significantly, even better than basic CNNs and ResNets.

As for the training speed, RNN models generally need longer training time compared with CNN-based models. BiLSTM-attention becomes the most time-consuming model. By contrast, although ResNet-18 has the most parameters, it takes only less than half of the training time of BiLSTM-attention. Thus, the CNN-based DeCban model also achieves high efficiency. Taking the DeCban using double embedding as an example, the parameters are only one fifth of those of BiLSTM-attention, but the average F_1 value is increased by 1.4%, which shows that the proposed network can achieve better performance with less computing resources.

3.4. Comparison With the Latest Models

In addition to the baseline models with common model architectures, we compare DeCban with the existing predictors for RBP-RNA interactions. Currently, the predictors for circRNAs are very few. CRIP (Zhang et al., 2019) and PASSION (Jia et al., 2020) are two recently developed models. We compare them with DeCban in terms of feature extraction, model architecture, and input, as described in the following.

CRIP also uses the 3-nucleotide codons to convert RNAs into pseudo-amino acids, i.e., the stacked-codon encoding scheme. However, CRIP presents the pseudo-amino acids as one-hot vectors, while DeCban uses word embeddings for both original RNAs and the converted pseudo-amino acids. PASSION incorporates some traditional statistical features in addition to CRIP's features. Therefore, a major difference between DeCban and the previous studies is using continuous dense feature encoding instead of sparse discrete features. Besides, the double embeddings contain the information of both RNA segments and pseudo-peptides, so as to strengthen the representation of raw sequences.

As for the model architecture, CRIP adopts a CNN-LSTM hybrid network, and PASSION proposes an ensemble classifier, which combines the hybrid network with an artificial neural network (consisting of fully-connected layers). DeCban is a CNN-based multi-branch attention network. As shown in **Table 2**, the parameter quantity of CRIP is 900 K, and PASSION has more parameters due to the ensemble nature; while DeCban with double embedding uses only one seventh of the parameters of CRIP.

Finally, both CRIP and PASSION perform prediction on short fragments, i.e., 101-nt segments. The incomplete sequences may lose some characteristics of original RNA molecules and lead to more false positive predictions, as mentioned in Zhang et al. (2019), while DeCban handles full-length sequences. **Figure 4** shows the ROC curve of CRIP. The average AUC value of the CRIP model on 37 data sets is 0.821, while DeCban is 0.905. DeCban gets significantly higher AUC value than that of CRIP on nearly all datasets. And, according to the results reported in Jia et al. (2020), PASSION's AUC is about 0.01 higher than that of CRIP. As both these two methods' inputs are short fragments with balanced positive-to-negative ratio, they may have close performance when handling full-length circRNAs.

4. DISCUSSION

Circular RNAs are a special kind of non-coding RNAs, which play an important role in gene regulation and disease development. Studying the interactions between circRNAs and RBPs can reveal the functions of circRNAs. However, the prediction of binding sites on circRNAs faces many challenges.

First, the length range of circRNA sequences is very large, from tens to over 100,000 nt, which adds great difficulty to the learning models. Thus, it is important to design a network to adapt to the large variance of input sequences. The multi-branch design of DeCban aims to extract features from different ranges of sequence regions, as the branches differ in kernel sizes, leading to different receptive fields. For instance, assume that step length is 3, with 0 padding and 0 dilation. When the convolution kernel size is 3, the receptive field sizes of the features output by the first and second layers are 3 and 5, respectively. When the convolution kernel size is 5, the receptive field sizes of the features output by the first and second layers are 5 and 9. Thus, different convolution kernel sizes can extract features of different scales.

The second challenge is that RBP-binding sites are extremely sparsely located in the whole RNA sequences, i.e., the number of binding sites are few and the binding regions are very short compared to full-length sequences. Thus, this is a severely imbalanced learning task, as most of the regions have no binding affinity. The attention mechanism in DeCban can alleviate this problem to a certain extent, which enables the model focus on key regions in long sequences.

The third challenge arises from the data side. Compared with linear RNAs, domain knowledge or information sources other than sequences are lacked. By utilizing the codon-based mapping between RNA and peptides, and performing large-scale pre-training of word embeddings for both RNA segments and peptides, we propose a new feature representation method for circRNAs, called double embeddings. Experiments show that this method effectively improves the representation ability for raw sequences.

Compared with the existing circRNA-RBP prediction methods, DeCban has the following advantages:

(1) The prediction can be performed on full-length circRNA sequences instead of short segments.
(2) The model is highly efficient, whose training has a low cost on computation resources.
(3) The high prediction accuracy makes it a useful tool for studying circRNA-RBP interactions.

5. CONCLUSION

In this study, we propose a method called DeCban to predict the binding relationship between RNA-binding-proteins and circRNAs. Different from the existing tools which can only handle short segments of circRNAs, DeCban is able to predict whether a binding site is present on full-length circRNAs. In order to solve the problem of large length span and sparse distribution of binding sites, we design a multi-branch and multi-layer convolutional neural network with an attention module. Moreover, to enhance the input data representation, we propose the double embedding encoding scheme, which is superior to the traditional single RNA embedding due to the introduction of amino-acid-level sequence information. We perform experiments on 37 data sets, corresponding to 37 RBPs. The experimental results show that our method achieves the best results compared with a variety of advanced deep learning structures. DeCban will be a useful tool for studying the interactions between RBP and circRNA.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LY and YY designed the model, analyzed the results, and wrote the manuscript. LY conducted the experiments. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61972251).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.632861/full#supplementary-material

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33:831. doi: 10.1038/nbt.3300

Chen, W., Lei, T. Y., Jin, D. C., Lin, H., and Chou, K. C. (2014). Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition. *Anal. Biochem.* 456:53. doi: 10.1016/j.ab.2014.04.001

Clauwaert, J., and Waegeman, W. (2019). Novel transformer networks for improved sequence labeling in genomics. *bioRxiv [Preprint]*. 836163. doi: 10.1101/836163

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Du, W. W., Zhang, C., Yang, W., Yong, T., Awan, F. M., and Yang, B. B. (2017). Identifying and characterizing circRNA-protein interaction. *Theranostics* 7:4183. doi: 10.7150/thno.21299

Dudekula, D. B., Panda, A. C., Grammatikakis, I., De, S., Abdelmohsen, K., and Gorospe, M. (2016). Circinteractome: a web tool for exploring circular RNAs and their interacting proteins and microRNAs. *RNA Biol.* 13, 34–42. doi: 10.1080/15476286.2015.1128065

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-hit. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Hansen, T. B., Kjems, J., and Damgaard, C. K. (2013). Circular RNA and MIR-7 in cancer. *Cancer Res.* 73, 5609–5612. doi: 10.1158/0008-5472.CAN-13-1568

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Doha. doi: 10.1109/CVPR.2016.90

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Jia, C., Bi, Y., Chen, J., Leier, A., Li, F., and Song, J. (2020). PASSION: an ensemble neural network approach for identifying the binding sites of RBPs on circRNAs. *Bioinformatics* 36, 4276–4282. doi: 10.1093/bioinformatics/btaa522

Ju, Y., Yuan, L., Yang, Y., and Zhao, H. (2019). Circslnn: Identifying rbp-binding sites on circrnas via sequence labeling neural networks. *Front. Genet.* 10:1184. doi: 10.3389/fgene.2019.01184

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*. doi: 10.3115/v1/D14-1181

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2013). starbase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale clip-seq data. *Nucl. Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248

Li, X., Yang, L., and Chen, L.-L. (2018). The biogenesis, functions, and challenges of circular RNAs. *Mol. Cell* 71, 428–442. doi: 10.1016/j.molcel.2018.06.034

Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., et al. (2015). Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.* 25:981. doi: 10.1038/cr.2015.82

Meher, P. K., Sahu, T. K., Gahoi, S., Satpathy, S., and Rao, A. R. (2019). Evaluating the performance of sequence encoding schemes and machine learning methods for splice sites recognition. *Gene* 705, 113–126. doi: 10.1016/j.gene.2019.04.047

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *Comput. Sci. arXiv preprint* arXiv:1301.3781.

Pamudurti, N. R., Bartok, O., Jens, M., Ashwalfluss, R., Stottmeister, C., Ruhe, L., et al. (2017). Translation of circrnas. *Mol. Cell* 66, 9–21.e7. doi: 10.1016/j.molcel.2017.02.021

Pan, X., Rijnbeek, P., Yan, J., and Shen, H.-B. (2018). Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* 17:582. doi: 10.1186/s12864-018-4889-1

Pan, X., and Shen, H.-B. (2017). RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics* 18:136. doi: 10.1186/s12859-017-1561-8

Pan, X., and Shen, H.-B. (2018). Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics* 34, 3427–3436. doi: 10.1093/bioinformatics/bty364

Pan, X., Yang, Y., Xia, C.-Q., Mirza, A. H., and Shen, H.-B. (2019). Recent methodology progress of deep learning for RNA-protein interaction prediction. *Wiley Interdisc. Rev.* 10:e1544. doi: 10.1002/wrna.1544

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: Global vectors for word representation," in *Proceedings of EMNLP*, Doha, 1532–1543. doi: 10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. *arXiv preprint* arXiv:1802.05365. doi: 10.18653/v1/N18-1202

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Wang, Z., and Lei, X. (2020). Matrix factorization with neural network for predicting circrna-rbp interactions. *BMC Bioinformatics* 21:229. doi: 10.1186/s12859-020-3514-x

Yan, J., Friedrich, S., and Kurgan, L. (2016). A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues. *Brief. Bioinformatics* 17, 88–105. doi: 10.1093/bib/bbv023

Yang, Y.-C. T., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., et al. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* 16:51. doi: 10.1186/s12864-015-1273-2

Zang, J., Lu, D., and Xu, A. (2020). The interaction of circRNAs and RNA binding proteins: an important part of circRNA maintenance and function. *J. Neurosci. Res.* 98, 87–97. doi: 10.1002/jnr.24356

Zhang, K., Pan, X., Yang, Y., and Shen, H.-B. (2019). Crip: predicting circRNA-RBP interaction sites using a codon-based encoding and hybrid deep neural networks. *RNA* 25:rna.070565.119. doi: 10.1261/rna.070565.119

Zhang, Y., Wang, X., and Kang, L. (2011). A k-mer scheme to predict pirnas and characterize locust piRNAs. *Bioinformatics* 27:771. doi: 10.1093/bioinformatics/btr016

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., et al. (2016). "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Doha. doi: 10.18653/v1/P16-2034

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/fgene.2021.635451

An Efficient Computational Model for Large-Scale Prediction of Protein–Protein Interactions Based on Accurate and Scalable Graph Embedding

Xiao-Rui Su[1,2,3], Zhu-Hong You[1,2,3], Lun Hu[1,2,3], Yu-An Huang[1], Yi Wang[1,2,3] and Hai-Cheng Yi[1,2,3]*

[1]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Ürümqi, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]Xinjiang Laboratory of Minority Speech and Language Information Processing, Ürümqi, China

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Tiantian He,
Nanyang Technological University, Singapore
Liang Zhao,
Hubei University of Medicine, China

****Correspondence:***
Zhu-Hong You
zhuhongyou@ms.xjb.ac.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 November 2020*
Accepted: *25 January 2021*
Published: *26 February 2021*

Citation:
Su X-R, You Z-H, Hu L, Huang Y-A, Wang Y and Yi H-C (2021) An Efficient Computational Model for Large-Scale Prediction of Protein–Protein Interactions Based on Accurate and Scalable Graph Embedding. Front. Genet. 12:635451. doi: 10.3389/fgene.2021.635451

Protein–protein interaction (PPI) is the basis of the whole molecular mechanisms of living cells. Although traditional experiments are able to detect PPIs accurately, they often encounter high cost and require more time. As a result, computational methods have been used to predict PPIs to avoid these problems. Graph structure, as the important and pervasive data carriers, is considered as the most suitable structure to present biomedical entities and relationships. Although graph embedding is the most popular approach for graph representation learning, it usually suffers from high computational and space cost, especially in large-scale graphs. Therefore, developing a framework, which can accelerate graph embedding and improve the accuracy of embedding results, is important to large-scale PPIs prediction. In this paper, we propose a multi-level model LPPI to improve both the quality and speed of large-scale PPIs prediction. Firstly, protein basic information is collected as its attribute, including positional gene sets, motif gene sets, and immunological signatures. Secondly, we construct a weighted graph by using protein attributes to calculate node similarity. Then GraphZoom is used to accelerate the embedding process by reducing the size of the weighted graph. Next, graph embedding methods are used to learn graph topology features from the reconstructed graph. Finally, the linear Logistic Regression (LR) model is used to predict the probability of interactions of two proteins. LPPI achieved a high accuracy of 0.99997 and 0.9979 on the PPI network dataset and GraphSAGE-PPI dataset, respectively. Our further results show that the LPPI is promising for large-scale PPI prediction in both accuracy and efficiency, which is beneficial to other large-scale biomedical molecules interactions detection.

Keywords: large-scale, protein-protein interaction, GraphZoom, weighted graph, graph embedding

INTRODUCTION

Over the past years, with the rapid development of biomedical researches as well as computer technologies, an increasing number of biomedical data, such as biomedical entities and their relationships, have been extracted from unconstructed data (Su et al., 2018). As an important and pervasive data carrier, a graph is considered the most suitable structure to present biomedical entities and their relationships. Both the availability of biomedical data and the researches of graphs have greatly facilitated biomedical graph studies, such as graph embedding, node properties prediction, and link prediction.

As the material basis of life, proteins are involved in every cell and almost every primary cellular process (Gavin et al., 2002). Analyzing protein–protein interactions (PPIs) can provide valuable insights into the molecular mechanisms underlying a living cell (Ma et al., 2011). Due to the rapid research in high-throughput technologies and biomedical studies, millions of PPI data have been collected from various experiments. Many databases have been constructed accordingly. However, too much data brings a few problems, such as high false-positive rates, low coverage, and high cost. Therefore, it is very meaningful to propose a high-efficiency computing method to identify PPIs.

Much work has been done in predicting PPIs. According to the method, it generally can be categorized into two groups based on either (1) feature extraction or (2) based on machine learning and deep learning. For the first group, they concentrate on the feature design. Features are extracted from kinds of sources, including protein sequence, functional domain information, physicochemical properties, and the fusion of feature sources. For example, Shen et al. (2007) predicted PPIs using conjoint-triad feature extracted from protein amino acids to represent protein. His work achieved a promising accuracy of 83.90% when applied to a 16,000 diverse PPI pairs dataset. On the basis of a protein sequence, functional domain information was necessary for the understanding of biological processes. Hence, Mudita and Resat (2007) proposed a method based on quantitative score measuring domain-domain interactions derived from available PPI database, then used the obtained score to predict interaction probability between two proteins. Chen et al. (2019) designed three types of protein-pair features based on physicochemical properties of amino acids, gene ontology annotations, and interaction network topologies. Then they introduced an ensemble learning approach for PPI prediction integrating three kinds of features. As for the second group, they concentrate on the design of classifier or neural network. Both machine learning methods and deep learning methods are based on statistics theories. Machine learning methods utilize classifiers to predict PPIs, such as naïve Bayes (NB), logistic regression (LR), random forest (RF), and support vector machine (SVM). Methods based on deep learning tend to apply neural networks to address PPI prediction, such as convolution neural network (CNN), recurrent neural network (RNN), and long short-term memory (LSTM). For instance, Romero-Molina et al. (2019) predicted the protein-protein interactions using SVM based on the sequence of proteins. Wang et al. (2017a,b,c) explored the protein evolutionary features from the angle of the image processing techniques in order to open a new way of researching protein sequences. Sequence-based approaches typically represent protein sequence as a vector using feature representation method, then the vector as an input of classification algorithm. All of these methods have achieved a promising result. However, they tend to concentrate on protein feature extraction and the design of neural networks and not the complex relationships that the proteins have, such as graph topology. More specifically, proteins collaborate and interact with each other to perform biological functions, leading to many protein interactions, which can be integrated and modeled as a graph/network structure. Therefore, it is important to detecting PPIs from the perspective of graph structure.

Analyzing and modeling the biomedical data with graph structure rely on a thorough understanding of graph topology. Numerous network-based learning methods have been developed to explore the interactions between proteins. They are classified into three categories, based on (1) network diffusion, (2) handcrafted graph features, and (3) graph representation learning. For the first group, the diffusion methods employ random walk techniques for influence propagation in different networks, such as integrating PPI networks into disease gene prediction (Luo et al., 2019). For the second group, various features for proteins are extracted and then fed into traditional machine learning methods. Other tasks also benefit from various features, especially when processing graph structure data. For example, graph clustering task (He et al., 2019a,b) utilize these multiview features to detect biological module. Graph clustering is also conducive to graph representation learning tasks because such methods are able to decrease the graph scale, and they can then improve the efficiency of the representation learning model. As for the third group, instead of a handcrafted feature, graph representation learning methods learn features automatically. This kind of method aims to learn a low-dimension representation for each node. Representative methods include Matrix Factorization-based model, Random Walk-based model, and Neural Network-based model. MF-based model (Belkin and Niyogi, 2003) learns graph representation by factorizing the matrix of input data into lower dimensional matrices. RW-based model (Perozzi et al., 2014; Grover and Leskovec, 2016) learns representation by generating a sequence of nodes randomly. The NN-based model integrates neural networks into representation learning. For example, (Kipf and Welling, 2016) proposed that graph convolutional networks (GCN) are perhaps the most representative graph neural network models, having a strong ability in the task of semi-supervised classification. The key issue in GCN is about the filter design in fact since it has a huge influence on the efficiency of model. Additionally, with the widely used of attention mechanism, attention-based graph neural network born, namely graph attention networks (GATs; Velikovi et al., 2017). Compared with GCN, GATs are more flexible and efficient since less parameters are used and can be parallelized. Although graph embedding is the most popular among these three methods, it usually suffers from high computational and space cost, owing to high dimensionality, sparsity of the network, and rapid expansion of the network. Therefore, developing an efficient framework, which can accelerate graph embedding and improve the embedding results accuracy, is important to both PPI and other molecular interactions.

In this paper, we proposed a multi-level model LPPI to improve both the quality and speed of large-scale PPIs prediction. LPPI consists four parts: (i) data collecting, (ii) graph embedding, (iii) embedding enhancement, and (iv) results prediction. Data collecting contains attribute feature extracting. We adopt the fundamental information as the attribute feature, such as positional gene sets, motif gene sets, and immunological signatures. In addition, the protein attribute is used to reconstruct a weighted graph by calculating node similarity. Then, graph embedding is used to learn the topology feature for each node. During this process, GraphZoom (Deng et al., 2019) is applied to accelerate the embedding process by reducing the size of the graph. After enhancing embedding, the classifier is used to predict interactions between protein pairs.

Our contributions are 2-fold. Firstly, LPPI integrates protein attribute into graph embedding task. More than that, LPPI adds weight to the link by calculating node similarity adopting the protein attribute. In this way, multi-view information is used when learning node representation, which is conducive to the improvement of accuracy. Secondly, we reconstruct the graph by using the GraphZoom algorithm in order to reduce the size of the graph. In this way, we can accelerate the efficiency of any network embedding algorithms. By combining the above two aspects, LPPI can save execution time without losing accuracy. Experiments on PPI network dataset and GraphSAGE-PPI dataset demonstrate that LPPI compares favorably both in classification accuracy and efficiency (measured in CPU time) against baseline models for large-scale PPI prediction.

MATERIALS AND METHODS

Benchmark Dataset

In order to validate the efficiency of our model, we collected two datasets with different sizes, which are the PPI network dataset and the GraphSAGE-PPI dataset. The statistics of the datasets are in **Table 1** in which the Density is defined as

$$\frac{2 * \# Links}{\# Nodes^2}$$

The positive PPI network dataset was downloaded from Stanford Large Network Dataset Collection (PPI Network, May 2017 version). This version of the PPI Network contains 818,716 protein-protein pairs of experimentally verified PPIs from 23,997 different human proteins. After eliminating self-interactions and duplicate interactions, we finally obtain 663,954 unique positive protein-protein pairs. The dataset is available at http://snap.stanford.edu/graphsage/ppi.zip.

The positive GraphSAGE-PPI (Sep 2018 version) dataset was collected following (Hamilton et al., 2017), which was also constructed by Stanford University. This version data set was used as the benchmark to train GraphSAGE. The data resource was the same as the PPI network. However, differently from PPI network, GraphSAGE-PPI contains fewer nodes and links, which are numbered at 6,370 and 186,421, respectively. The dataset is available at http://github.com/williamleif/GraphSAGE/example_data.

TABLE 1 | Statistics of the datasets.

Dataset	#Nodes	#Links	Density
PPI network	23,997	663,954	0.23%
GraphSAGE-PPI	6,370	186,421	0.92%

One of the common ways to construct the negative data set is to consider two proteins with different cellular compartments nor interacting. In this study, we adopted the same strategy to construct a negative dataset for two benchmark datasets. We followed this idea and constructed each benchmark dataset according to the following criteria: (1) the number of negative samples was equal to that in the positive dataset; (2) we constructed a complementary graph; (3) we removed the interactions from the same cellular compartments; and (4) we randomly selected noninteracting protein pairs from the complementary graph. After that, a negative dataset had been constructed, which was trained with a positive dataset together. Five-fold cross-validation was adopted when training the model, and, therefore, a negative dataset was constructed at each fold.

Protein Attribute Extraction

In order to represent protein nodes, we extracted the protein attribute features following (Hamilton et al., 2017). Using positional gene sets, motif gene sets, and immunological signatures as features, collected from the Molecular Signatures Database (Subramanian et al., 2005). Positional gene sets corresponding to each human chromosome and each cytogenetic band that has at least one gene. There are 326 positional gene sets in total. As for the motif gene sets, they represent potential targets of regulation by transcription factors or microRNAs. The sets consist of genes grouped by short sequence motifs they share in their non-protein coding regions. Immunological signatures represent cell states and perturbations within the immune system. The signatures are generated by manual curation of published studies in human and mouse immunology. Finally, the protein attribute feature is obtained.

Graph Embedding

Graph embedding methods aim to automatically learn a low-dimensional feature representation for each node in the graph (Wang et al., 2016). Traditionally, a low-dimensional feature is considered as the structural information of the graph. Therefore, it can be used in various downstream tasks. Since the concept of graph embedding proposed, graph embedding methods can be categorized into three groups: MF-based, RW-based, and NN-based (Su et al., 2018; Yue et al., 2019).

For the sake of efficiency improvement, we adapted the RW-based method, which was inspired by the word2vec model (Mikolov et al., 2013). The RW-based method tries to learn node representation by generating node sequence through random walk in graphs. In this way, topological information can be preserved into a low-dimensional vector. As the two representative methods based on random walk, DeepWalk (DW; Perozzi et al., 2014) and Node2vec (Grover and Leskovec, 2016) were applied to learn latent features. DW considers the paths as sentences and implements Skip-Gram to learn the embedding of each node. Specifically, the DeepWalk algorithm first generates a random walk path $P^1_{v_i}, P^2_{v_i}, P^3_{v_i}, \ldots, P^n_{v_i}$ by taking v_i as the root node, the symbol n represents the length of random walk path. Therefore, the aim is to predict the next node according previous sequence:

$$\Pr = (P^m_{v_i} | P^1_{v_i}, P^2_{v_i}, P^3_{v_i}, \ldots, P^{m-1}_{v_i})$$

However, it is difficult to calculate an order sequence in the experiment. In order to solve this problem, Skip-Gram is used to learn the random walk path. This algorithm does not take the sequence order into consideration but sets a sliding window of length n, using target words to predict context. Therefore, the objective function of optimization is as follows:

$$\min_{p^m_{v_i}} - \log \Pr\left(\left\{p^{m-1}_{v_i}, \ldots, p^{m-n}_{v_i}, p^{m+1}_{v_i}, \ldots, p^{m+n}_{v_i}\right\} | p^m_{v_i}\right)$$

Compared to DeepWalk, Node2vec introduces the probability of controlling the walk direction. Therefore, the objective function of optimization is as follows:

$$\max_f \sum_{u \in V} \log \Pr\left(N_s\left(U\right) | f\left(u\right)\right)$$

In this formulation, u represents the current node and $N_s\left(U\right)$ represents the nodes selected by strategy s. In Node2vec, it adapts the breadth-first search (BFS) and the depth-first search (DFS) into the generation process of the random walk sequence by introducing return hyperparameter p and ahead hyperparameter q to control the probability of a walk. The probability $\alpha\left(m,\theta\right)$ from current node m to next node θ is defined as follows:

$$\alpha\left(m,\theta\right) = \begin{cases} \dfrac{1}{p}, if\ d_{m\theta} = 0 \\ 1, if\ d_{m\theta} = 1 \\ \dfrac{1}{q}, if\ d_{m\theta} = 2 \end{cases}$$

Breadth-first search focuses on neighboring nodes and characterizes a relatively local network representation. DFS reflects the homogeneity between nodes at a higher level. Specifically, BFS explores the structural properties of the graph, while DFS explores the similarity in content or similarity between adjacent nodes.

GraphZoom

Owing to the scalable of PPIs data, it is essential to accelerate the graph embedding process. In this section, GraphZoom (Deng et al., 2019) is applied to improve the accuracy and efficiency of graph embedding. GraphZoom is a multi-level framework for improving both the accuracy and scalability of unsupervised graph embedding algorithms. There are four components in it: (1) graph fusion, (2) spectral graph coarsening, (3) graph embedding, and (4) embedding refinement.

For the first step, original graph topology and attribute information are combined to construct a weighted graph, which has the same number of nodes as the original graph. Graph topology can be represented by the adjacency matrix $A_{topo} \in R^{N \times N}$, and cosine similarity on attribute feature is used to calculate edge weight A_{feat}. Then, the fused graph can be represented by a weighted sum:

$$A_{fusion} = A_{topo} + \beta A_{feat}$$

The second step is spectral coarsening, which is the core part of GraphZoom. In order to improve the embedding speed, a fused graph constructed before is coarsened into a much smaller graph by merging nodes with high spectral similarities. Inspired by signal processing, simple smoothing (low-pass graph filtering) function is applied to k random vectors to obtain smoothed vectors for k-dimensional graph embedding instead of calculating the eigenvectors of the original graph Laplacian. Gauss-Seidel iteration method is used to solve k linear equations to obtain initial random k-dimensional feature representation. x represents a random vector calculated by Gauss-Seidel, which is expressed with a linear combination of eigenvectors u of the graph Laplacian. Smoothed vector $\tilde{u}$ is obtained by applying the smoothing function. Then the nodes with a higher spectral affinity $a_{p,q}$ are locally clustered, and a graph with fewer nodes (adjacency matrix) is obtained so repeatedly. This method can be achieved in linear time. The whole process can be formulated:

$$x = \sum_{i=1}^{N} \alpha_i u_i \xrightarrow{smoothing} \tilde{x} = \sum_{i=1}^{n} \widetilde{\alpha_i} u_i, n \ll N$$

$$a_{p,q} = \frac{\left|\left(K_{p,:}, K_{q,:}\right)\right|^2}{{K_{p,:}}^2 {K_{q,:}}^2}$$

$$\left(K_{p,:}, K_{q,:}\right) = \sum_{t=1}^{k} \left(x_p^{(t)} \cdot x_q^{(t)}\right)$$

As for the third step, any unsupervised embedding methods can be applied to embed the coarsest graph. The last step is embedding refinement. Using Laplace smoothing to map the node representation to each node of the original graph, then embedding representation of the original graph node is obtained. The embedding results can be calculated as follows:

$$E_i = \left(\tilde{D}_i^{-\frac{1}{2}} \tilde{A}_i \tilde{D}_i^{-\frac{1}{2}}\right)^k \hat{E}_i = \left(\tilde{D}_i^{-\frac{1}{2}} \tilde{A}_i \tilde{D}_i^{-\frac{1}{2}}\right)^k H^i_{i+1} E_{i+1}$$

$$\tilde{A} = A + \sigma I, \tilde{D} = D + \sigma I$$

where A is the adjacency matrix, D is the degree matrix, H^i_{i+1} is the graph mapping operator between two coarsening levels i and i+1, and σ is a small value to ensure every node has its own self-loop.

RESULTS

Evaluation Criteria

To verifies the proposed method in the experiments, we followed the 5-fold cross-validation specification. To evaluate the proposed method more fairly, a range of performance evaluation measures were computed including accuracy (Acc.), sensitivity (Sen.), precision (Pre.), and the Matthews correlation coefficient (MCC), which can be defined respectively:

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where the TN, TP, FN, and FP denotes the number of correctly predicted positive and negative samples, wrongly predicted positive and negative samples, respectively. Furthermore, the Receiver Operating Characteristic (ROC) curve, which represents the results of multiple confusion matrices, using a false positive rate as its x-axis and true positive rate as its y-axis. The area under curve (AUC) of ROC, which follows a philosophy of the bigger the better, is also adopted to measure the performance of the proposed model.

Model Construction

We implemented our model on two data sets of different sizes. In order to maintain unity, we also integrated DeepWalk and Node2vec as the basic embedding methods into the proposed model, respectively. As for the hyperparameters in DeepWalk and Node2vec, we used 10 walks with a walk length of 80 and set the embedding dimension to 128. In addition, Node2vec has two parameters, return parameter p and in-out parameter q, which control the direction of the next step. We set them to 1.0 and 0.5, respectively. The effectiveness of these parameters is verified by other experiments. GraphZoom is used to enhance the graph and accelerate the graph embedding process. The hyperparameters used in GraphZoom are fusion parameter β and coarsening level l. β controls the proportion of attribute feature. Parameter l is used in the graph reduction process, which controls the size of the reconstructed graph. Coarsening level l represents the iteration that the original graph is to be reconstructed. With the increase of the coarsening level, the scale of the graph is smaller. We adapted 0.1 and 1 in the baseline model, respectively. After obtaining graph embedding representation, several classifiers were applied to predict protein pairs. It should note that all parameters used in classifiers were the default. The model overview is shown in **Figure 1**.

Performance on Two Large-Scale Datasets

We test the performance of our model on two benchmark datasets. To contextualize the empirical results on benchmarks, we construct a baseline model, which integrates DeepWalk (Perozzi et al., 2014) as the graph embedding method and LR (Hosmer et al., 2013) as a classifier. Five-fold cross-validation is used to test the baseline model. The results are shown in **Table 2**. In addition, we also compare the CPU time of two datasets, which is shown in **Figure 2**.

Our model achieves a highly predictive performance on both two datasets, which average accuracy is 0.99997 and 0.9979, respectively. Compared with GraphSAGE-PPI data et, our model achieves better results on the PPI network dataset, which demonstrates that our model has the ability to process large-scale dataset precisely. More specifically, the number of nodes and links of the PPI network dataset is three times that in the GraphSAGE-PPI dataset in fact; however, the time cost of two datasets is similar, which further demonstrates

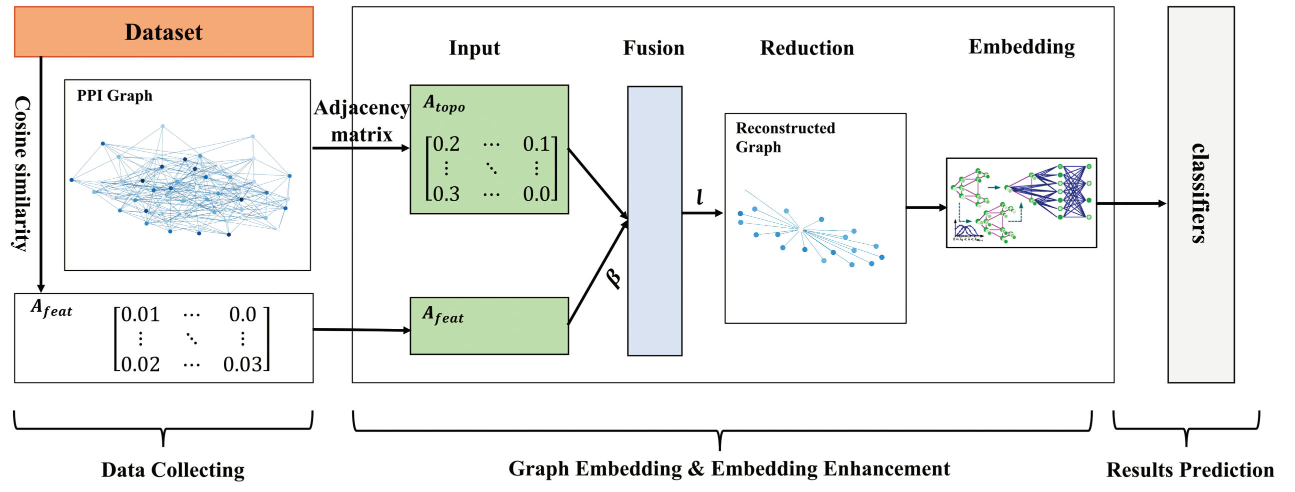

Dataset
PPI Graph
Cosine similarity
A_{feat}
$\begin{bmatrix} 0.01 & \cdots & 0.0 \\ \vdots & \ddots & \vdots \\ 0.02 & \cdots & 0.03 \end{bmatrix}$
Adjacency matrix
Input
A_{topo}
$\begin{bmatrix} 0.2 & \cdots & 0.1 \\ \vdots & \ddots & \vdots \\ 0.3 & \cdots & 0.0 \end{bmatrix}$
A_{feat}
Fusion
β
l
Reduction
Reconstructed Graph
Embedding
classifiers
Data Collecting
Graph Embedding & Embedding Enhancement
Results Prediction

FIGURE 1 | The overview of the proposed model.

TABLE 2 | Prediction results for two datasets. DW means Deepwalk, and LR represents Logistic Regression.

Baseline model	Fold	PPI network					GraphSAGE-PPI				
		Acc.	Pre.	Sen.	MCC	AUC	Acc.	Pre.	Sen.	MCC	AUC
LPPI (GZ-DW-LR)	0	0.99996	1.0	0.99992	0.99992	0.99996	0.9978	1.0	0.9956	0.9957	0.9978
	1	0.99997	1.0	0.99993	0.99993	0.99997	0.9979	1.0	0.9958	0.9958	0.9979
	2	0.99995	1.0	0.99991	0.99991	0.99996	0.9981	1.0	0.9961	0.9961	0.9981
	3	0.99997	1.0	0.99993	0.99993	0.99997	0.9980	1.0	0.9960	0.9959	0.9980
	4	0.99998	1.0	0.99996	0.99996	0.99998	0.9978	1.0	0.9957	0.9956	0.9978
Average		**0.99997**	**1.0**	**0.99993**	**0.99993**	**0.99996**	**0.9979**	**1.0**	**0.9958**	**0.9958**	**0.9979**

The bold values mean the best results achieved.

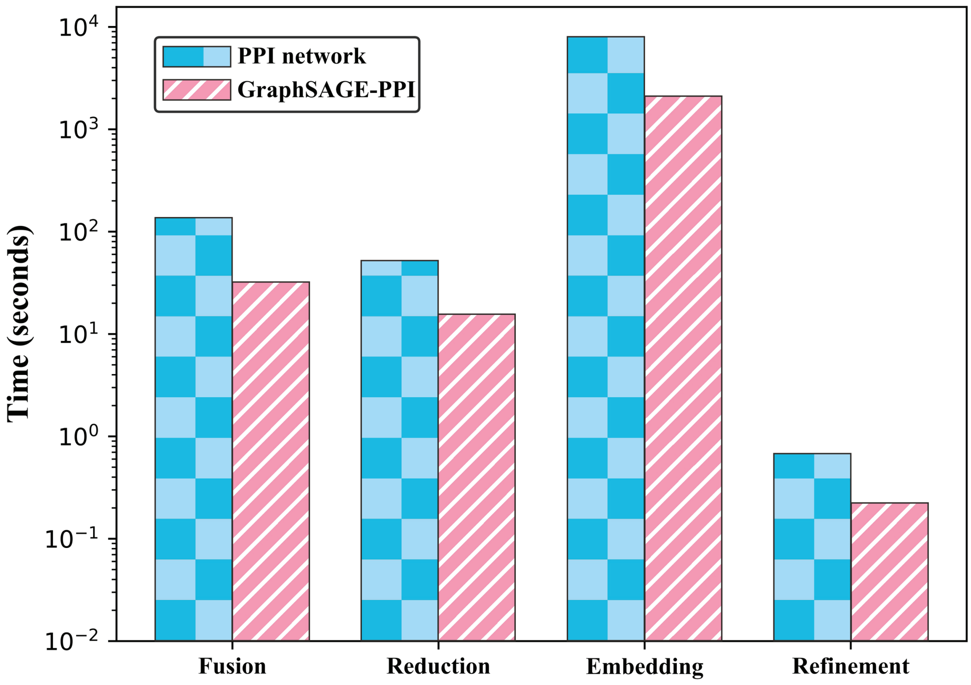


FIGURE 2 | Timing experiments of four parts on PPI network dataset and GraphSAGE-PPI dataset.

that our model can process the large-scale dataset efficiently. In conclusion, the proposed model has the ability to process high-density network both accurately and efficiently.

Comparing LPPI With Baseline Embedding Methods

In order to validate that the proposed model accelerates the embedding process without losing accuracy, we compared the proposed model with two baseline embedding methods, which are also integrated into the proposed model as part of the embedding. The results are shown in **Table 3** and **Figure 3**.

According to the results, firstly, the proposed model used less CPU time since LPPI had the graph reduction module, which can decrease the graph scale. In addition, it can be observed that the proposed model achieved higher accuracy than the other two baseline models on both the PPI network dataset and the GraphSAGE-PPI dataset. This is because LPPI contains more detailed information such as node attribute and concentrates more on the key part of the graph and eliminates noisy information. In conclusion, the proposed model performs better than baseline models mainly because (i) LPPI integrates node attribute information and node similarity as topology information into the model, which increase the accuracy of the proposed model, and (ii) LPPI reconstructs graph to reduce the graph scale, which is conducive to efficiency on embedding and noisy information eliminated.

Analysis on LPPI Kernels

There are two hyperparameters in LPPI model, which are fusion parameter β and coarsening level parameter l. In order to study the efficiency and accuracy of LPPI, we focused on two parameters. The results are shown in **Table 4** and **Figure 4**.

TABLE 3 | Summary of results in terms of mean classification accuracy (Acc.), AUC, and CPU time for different combinations in LPPI on the PPI network dataset and GraphSAGE-PPI dataset.

Method	PPI network			GraphSAGE-PPI		
	Acc.	AUC	Time(s)	Acc.	AUC	Time(s)
LPPI (GZ-DW-LR)	**0.99997**	0.99996	8131.417	0.9979	0.9979	2309.847
LPPI (GZ-NV-LR)	0.99993	**0.99997**	**5001.137**	**0.9984**	0.9983	**1232.644**
DeepWalk	0.99975	0.99990	12405.259	0.9544	0.9995	3633.228
Node2vec	0.99992	0.99995	7947.544	0.9879	**0.9999**	1580.749

The bold values mean the best results achieved.

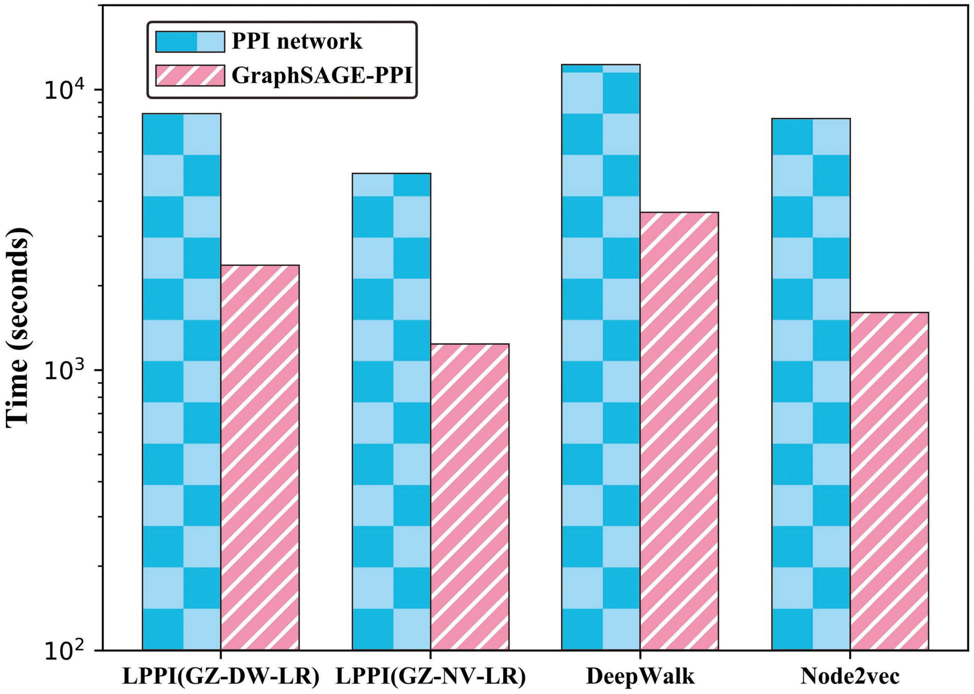


FIGURE 3 | Timing experiments of different embedding methods on PPI network dataset and GraphSAGE-PPI dataset.

Firstly, we discuss the influence of coarsening level. Coarsening level controls the size of the reconstructed graph. **Figure 5** shows that the bigger the coarsening level is, the smaller the reconstructed graph is. In our experiment, five different values are used. From the results, we can know that with the increase of the coarsening level, the accuracy of the two datasets is gradually decreased from 0.99997 to 0.99957 and from 0.9979 to 0.9858, respectively. Correspondingly, the CPU time is dramatically decreased from 8131.417 to 350.093 s and from 2309.847 to 69.745 s, respectively. When the coarsening level is 1, the model achieves the highest accuracy on the PPI network dataset, which is 0.99997, but it costs the most CPU time. The model with coarsening level 5 is the most efficient model as it costs the least CPU time, which is 350.093 s. More importantly, though the model using level 5 has the lowest accuracy, its accuracy is not much different from the model with level 1. As for the GraphSAGE-PPI dataset, LPPI achieves the best performance when the level is 2 with an accuracy of 0.9986 and AUC value of 0.9985. Overall, when the number of coarsening level is less than 5, the accuracy of LPPI is always higher than that of DeepWalk and LPPI improves the efficiency of DeepWalk by 17.8 times and 26.2 times on two datasets, respectively. Hence, experiment results prove that our model can accelerate the embedding process without losing accuracy.

Next, we discuss the fusion parameter β, which decides the proportion of attribute feature. In this part, we also try five different values for parameter β. According to our experiment results (**Figures 4C,D**), this parameter has a positive influence on the final result. With the increase of β, the accuracy is increase gradually. For the PPI network, the highest accuracy is 0.99997, which is achieved by 0.1, 0.8, and 1. As for the GraphSAGE-PPI, the highest accuracy is obtained when β is 0.8 and 1. This result indicates that combing the attribute feature with network embedding can improve the predictive performance. In addition, CPU time has not been affected by parameter β as this parameter has no influence on the scale of the reconstructed

TABLE 4 | Comparisons of different kernel parameters in GraphZoom in classification on the PPI network dataset and GraphSAGE-PPI dataset.

Method	PPI network			GraphSAGE-PPI		
	Acc.	AUC	Time(s)	Acc.	AUC	Time(s)
DeepWalk	0.99975	0.99990	12405.259	0.9544	0.9995	3633.228
LPPI(DW-LR,l = 1)	**0.99997**	**0.99996**	8131.417 (×1.5)	0.9979	0.9979	2309.847 (×1.6)
LPPI(DW-LR,l = 2)	0.99996	0.99996	4236.696 (×2.8)	**0.9986**	**0.9985**	1062.251 (×3.4)
LPPI(DW-LR,l = 3)	0.99996	0.99996	1810.727 (×6.9)	0.9971	0.9971	418.485 (×8.7)
LPPI(DW-LR,l = 4)	0.99987	0.99985	696.115 (×17.8)	0.9931	0.9931	138.625 (×26.2)
LPPI(DW-LR,l = 5)	0.99957	0.99957	**350.093 (×35.4)**	0.9858	0.9856	**69.745 (×52.1)**
LPPI(DW-LR,β = 0.1)	**0.99997**	**0.99996**	**8131.417 (×1.5)**	0.9979	0.9979	**2309.847 (×1.6)**
LPPI(DW-LR,β = 0.2)	0.99996	0.99980	8667.839 (×1.4)	0.9979	0.9979	2396.033 (×1.5)
LPPI(DW-LR,β = 0.4)	**0.99997**	**0.99996**	8606.011 (×1.4)	0.9980	0.9978	2318.294 (×1.6)
LPPI(DW-LR,β = 0.8)	0.99997	0.99997	8669.954 (×1.4)	**0.9982**	0.9982	2342.992 (×1.6)
LPPI(DW-LR,β = 1.0)	0.99997	0.99995	8836.558 (×1.4)	**0.9982**	0.9981	2384.745 (×1.5)

The bold values mean the best results achieved.

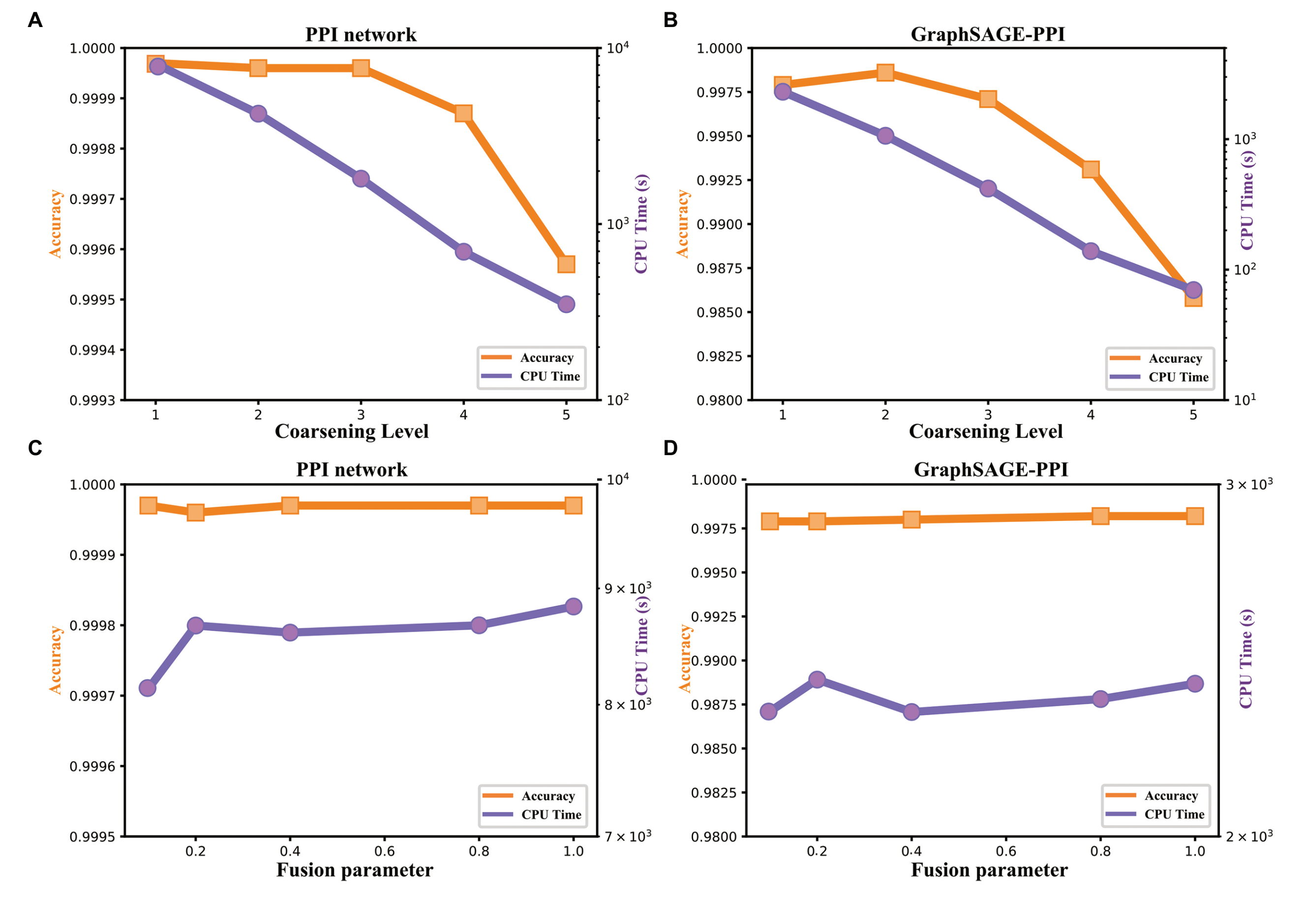


FIGURE 4 | Accuracy and timing experiments on two benchmark datasets. **(A)** Model performance with respect to the coarsening level on PPI network dataset. **(B)** Model performance with respect to the coarsening level on the GraphSAGE-PPI dataset. **(C)** Model performance about fusion parameter on PPI network dataset. **(D)** Model performance about fusion parameter on the GraphSAGE-PPI dataset.

graph. Even though, CPU time cost by LPPI with various fusion parameter is still less than that of DeepWalk.

In this part, we discuss two parameters used in our model. Parameter coarsening level l can accelerate the embedding process, parameter β can improve the accuracy value. These two parameters further demonstrate that our model has the ability to balance the performance of accuracy and efficiency.

TABLE 5 | Comparisons of different classifiers on the PPI network dataset and GraphSAGE-PPI dataset.

Method	PPI network			GraphSAGE-PPI		
	Acc.	AUC	Time (s)	Acc.	AUC	Time(s)
LPPI(GZ-DW-LR)	0.99997	0.99996	**8131.417**	0.9979	0.9979	**2309.847**
LPPI(GZ-DW-RF)	**0.99999**	**0.99998**	17783.854	**0.9999**	**0.9999**	2874.321
LPPI(GZ-DW-NB)	0.98799	0.99996	17673.404	0.9899	0.9956	2821.121

The bold values mean the best results achieved.

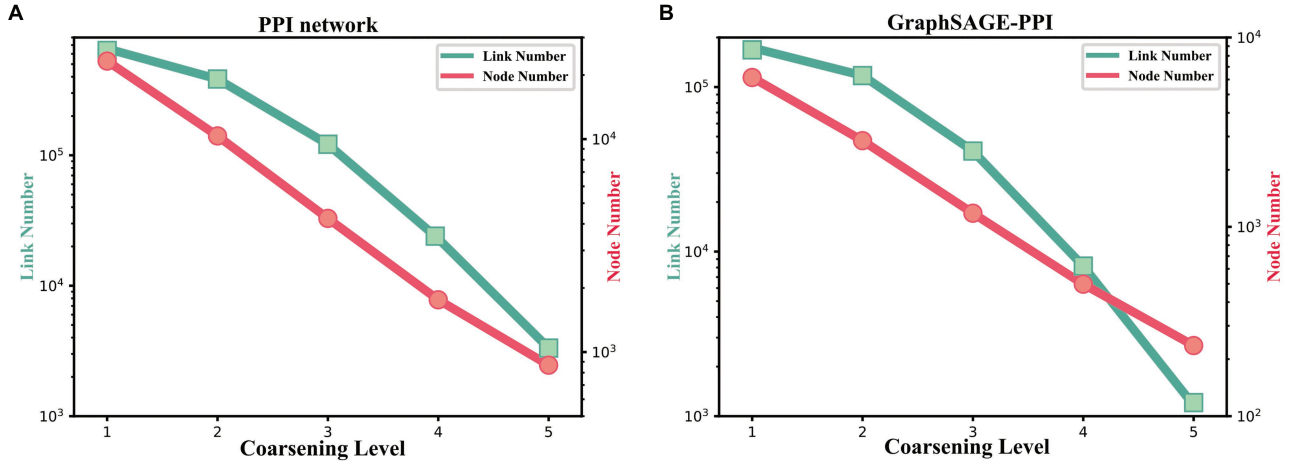


FIGURE 5 | (A) The change of link number and node number with the coarsening level increasing on the PPI network dataset. **(B)** The change of link number and node number with the coarsening level increasing on the GraphSAGE-PPI dataset.

Comparison of Different Classification Algorithms

After obtaining embedding features, classifiers are used to classify the protein pairs. In this section, we compare the results of different classifiers. Base on the baseline model, we compare three types of classifiers, including LR, RF, and NB (Rish, 2001; Liaw and Wiener, 2002; Hosmer et al., 2013) and the predictive performance is shown in **Table 5**. It should note that default parameters are used in different classifiers.

In our experiment, we test classifiers based on LPPI (GZ-DW). Among these three classifiers, LR is a linear model, RF belongs to an ensemble-based model, and NB is a generation model. From the results, it can be found that though RF achieves the best performances on both accuracy and AUC value for each dataset, it costs the longest time, which is not suitable for the sake of efficiency. On the other hand, LR has not only a promising performance with high accuracy and the AUC, but the least CPU time. As a result, LR is selected as the final classifier integrated into LPPI.

DISCUSSION

The proposed model has promising predictive performances on two large-scale datasets, the PPI network dataset and GraphSAGE-PPI dataset, which have 663,954 links and 186,421 links in total, respectively. Our model aims to address large-scale protein pairs prediction, efficiently and accurately. However, it is introductive to point out that there are still several limitations in our model. The current study constructs a multi-level framework for PPI prediction, containing four parts. In fact, classifiers as well as parameters affect results significantly, especially in classification tasks. Therefore, the performance of our model could still have a bias. Simultaneously, a multi-level framework is not convenient for a training model. In order to solve this problem, an end-to-end model is expected to be adapted. More specifically, we can replace classify layer with a forward neural network, which contributes to model training and CPU time. In addition, from the perspective of code implement, it is not efficient enough to link prediction tasks since the code is not parallelized, such as in the part of split data and 5-fold cross-validation.

Future efforts to improve the prediction of PPI based on the current study include (i) reducing the bias caused by classifiers, replacing the classify layer with a forward neural network, and (ii) improving efficiency through parallel computing, especially in the part of graph embedding.

CONCLUSION

In this study, we introduce a model LPPI, a multi-level framework to improve the accuracy and efficiency of large-scale protein-protein interactions prediction. The attribute feature is collected in LPPI firstly, which further is used to calculate the similarity between protein nodes to reconstruct a weighted graph. Then, a graph embedding method, such as DeepWalk and Node2vec, is applied to a new graph and generates topology features. Afterward, the classifier is used to test if protein pairs interact with each other. Experiments show that LPPI improves both classification accuracy and embedding speed on two benchmark datasets. Our work provides a new framework for large-scale protein-protein interactions prediction, which is beneficial to the detection of other biomedical molecule interactions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/Blair1213/LPPI.

AUTHOR CONTRIBUTIONS

X-RS and Z-HY designed the model and wrote the manuscript. X-RS, LH, Y-AH, YW, and H-CY conducted the experiments. Z-HY managed and directed the project. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (NSFC; No. 61572506), the National Outstanding Youth Science Fund Project of NSFC (No. 61722212), and the Pioneer Hundred Talents Program of the Chinese Academy of Sciences.

REFERENCES

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data. *Neural Comput.* 15, 1373–1396. doi: 10.1162/089976603321780317

Chen, K. -H., Wang, T. -F., and Hu, Y. -J. (2019). Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. *BMC Bioinformatics* 20:308. doi: 10.1186/s12859-019-2907-1

Deng, C., Zhao, Z., Wang, Y., Zhang, Z., and Feng, Z. (2019). 'GraphZoom: a multi-level spectral approach for accurate and scalable graph embedding.' Comput. Sci. [Preprint].

Gavin, A. -C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147. doi: 10.1038/415141a

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable feature learning for networks" in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge 1117 discovery and data mining (ACM)*; August 13–17, 2016; 855–864.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). "Inductive representation learning on large graphs."

He, T., Bai, L., and Ong, Y. S. (2019a). "Manifold regularized stochastic block model" in *31st International conference on tools with artificial intelligence (ICTAI'19);* November 4–6, 2019.

He, T., Liu, Y., Ko, T. H., Chan, K. C. C., and Ong, Y. S. (2019b). Contextual correlation preserving multi-view featured graph clustering. *IEEE Trans. Cybern.* 50, 4318–4331. doi: 10.1109/TCYB.2019.2926431

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. New Jersey: John Wiley & Sons.

Kipf, T. N., and Welling, M. (2016). "Semi-supervised classification with graph convolutional networks."

Liaw, A., and Wiener, M. (2002). "Classification and regression by randomForest." R News 2, 18–22.

Luo, P., Tian, L. -P., Ruan, J., and Wu, F. -X. (2019). Disease gene prediction by integrating PPI networks, clinical RNA-seq data and OMIM data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 222–232. doi: 10.1109/TCBB.2017.2770120

Ma, D. -C., Diao, Y. -B., Guo, Y. -Z., Li, Y. -Z., Zhang, Y. -Q., Wu, J., et al. (2011). A novel method to predict protein-protein interactions based on the information of protein-protein interaction networks and protein sequence. *Protein Pept. Lett.* 18, 906–911. doi: 10.2174/092986611796011482

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality" in *Proceedings of the 26th international conference on neural information processing systems-Volume 2*; December 12–15, 2013; Lake Tahoe, Nevada: Curran Associates Inc., 3111–3119.

Mudita, S., and Resat, H. (2007). A domain-based approach to predict protein-protein interactions. *BMC Bioinformatics* 8:199. doi: 10.1186/1471-2105-8-199

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "DeepWalk: online learning of social representations" in *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining;* August 24–27, 2014; New York, USA: Association for Computing Machinery, 701–710.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *J. Univ. Comput. Sci.* 1:127.

Romero-Molina, S., Ruiz-Blanco, Y. B., Harms, M., Münch, J., and Sanchez-Garcia, E. (2019). PPI-detect: a support vector machine model for sequence-based prediction of protein–protein interactions. *J. Comput. Chem.* 40, 1233–1242. doi: 10.1002/jcc.25780

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A.* 104, 4337–4341. doi: 10.1073/pnas.0607879104

Su, C., Tong, J., Zhu, Y., Cui, P., and Wang, F. (2018). Network embedding in biomedical data science. *Brief. Bioinform.* 21, 182–197. doi: 10.1093/bib/bby117

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Velikovi, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2017). "Graph attention networks."

Wang, D., Cui, P., and Zhu, W. (2016). "Structural deep network embedding" in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (ACM)*; August 13–17, 2016; 1225–1234.

Wang, Y., You, Z., Li, X., Chen, X., Jiang, T., and Zhang, J. (2017c). PCVMZM: using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein–protein interactions from protein sequences. *Int. J. Mol. Sci.* 18:1029. doi: 10.3390/ijms18051029

Wang, Y. -B., You, Z. -H., Li, L. -P., Huang, Y. -A., and Yi, H. -C. (2017a). Detection of interactions between proteins by using legendre moments descriptor to extract discriminatory information embedded in pssm. *Molecules* 22:1366. doi: 10.3390/molecules22081366

Wang, Y. -B., You, Z. -H., Li, X., Jiang, T. -H., Chen, X., Zhou, X., et al. (2017b). Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol. Biosyst.* 13, 1336–1344. doi: 10.1039/c7mb00188f

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2019). Graph embedding on biomedical networks: methods, applications, and evaluations. *Bioinformatics* 36, 1241–1251. doi: 10.1093/bioinformatics/btz718

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 30 March 2021
doi: 10.3389/fgene.2021.631505

Feature Selection Using Approximate Conditional Entropy Based on Fuzzy Information Granule for Gene Expression Data Classification

*Hengyi Zhang**

College of Animal Science and Technology, Northwest A&F University, Yangling, China

Classification is widely used in gene expression data analysis. Feature selection is usually performed before classification because of the large number of genes and the small sample size in gene expression data. In this article, a novel feature selection algorithm using approximate conditional entropy based on fuzzy information granule is proposed, and the correctness of the method is proved by the monotonicity of entropy. Firstly, the fuzzy relation matrix is established by Laplacian kernel. Secondly, the approximately equal relation on fuzzy sets is defined. And then, the approximate conditional entropy based on fuzzy information granule and the importance of internal attributes are defined. Approximate conditional entropy can measure the uncertainty of knowledge from two different perspectives of information and algebra theory. Finally, the greedy algorithm based on the approximate conditional entropy is designed for feature selection. Experimental results for six large-scale gene datasets show that our algorithm not only greatly reduces the dimension of the gene datasets, but also is superior to five state-of-the-art algorithms in terms of classification accuracy.

Keywords: feature selection, Laplacian kernel, fuzzy information granule, fuzzy relation matrix, approximate conditional entropy

Edited by:
Wilson Wen Bin Goh,
Nanyang Technological University,
Singapore

Reviewed by:
Guosheng Han,
Xiangtan University, China
Yusen Zhang,
Shandong University, China

****Correspondence:***
Hengyi Zhang
zhanghengyi2000@163.com

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *20 November 2020*
Accepted: *12 March 2021*
Published: *30 March 2021*

Citation:
Zhang H (2021) Feature Selection Using Approximate Conditional Entropy Based on Fuzzy Information Granule for Gene Expression Data Classification.
Front. Genet. 12:631505.
doi: 10.3389/fgene.2021.631505

INTRODUCTION

The development of DNA microarray technology has brought about a large number of gene expression data. It is a hot topic in bioinformatics to analyze and mine the knowledge behind these data (Sun et al., 2019b). As the most basic data mining method, classification is widely used in the analysis of gene expression data. Due to the small sample size and high dimensionality of gene expression data, the traditional classification methods are often ineffective when applied to gene expression data directly (Fu and Wang, 2003; Mitra et al., 2011; Phan et al., 2012; Konstantina et al., 2015). It has become a consensus in the academic community to reduce the dimensionality before classification. Feature selection is the most widely used dimensionality reduction method in gene expression data because it can maintain the biological significance of each feature. Feature selection can not only reduce the time and space complexity of classification learning algorithm, avoid dimensionality disaster, and improve the prediction accuracy of classification, but also help to explain biological phenomena.

Feature selection methods are generally divided into three categories: filter, wrapper, and embedded method (Hu et al., 2018). The filter method obtains the optimal subset of features by judging the similarity between the features and the objective function based on the statistical characteristics of data. The wrapper method uses a specific model to carry out multiple rounds of training. After each round of training, several features are removed according to the score of the objective function, and then the next round of training is carried out based on the new feature set. In this way, recursion is repeated until the number of remaining features reaches the required number. The embedded method uses machine learning algorithm to get the weight coefficient of each feature in the first place, and then selects the feature according to the weight coefficient from large to small. Wrapper and embedded methods have heavy computational burden and are not suitable for large-scale gene data sets. Our feature selection method belongs to the filter method, in which a heuristic search algorithm is used to find an optimal subset of features using approximate conditional entropy based on fuzzy information granule for gene expression data classification.

Attribute reduction is a fundamental research topic and an important application of granular computing (Dong et al., 2018; Wang et al., 2019). Attribute reduction can be used for feature selection. Granular computing is a new concept and new computing paradigm of information processing, which is mainly used to deal with fuzzy and uncertain information (Qian et al., 2011).

Pawlak (1982) proposed the rough set theory. Rough set theory is a new mathematical tool to deal with fuzziness and uncertainty. Granular computing is one of the important research contents of rough set theory. On the basis of equivalence relation, rough set theory is only suitable for dealing with discrete data widely existing in real life. When dealing with attribute reduction problem of continuous data in classical rough set theory, discretization method is often used to convert continuous data into discrete data, but the discretization will inevitably lead to information loss (Dai and Xu, 2012). To overcome this drawback, Hu et al. proposed a neighborhood rough set model (Hu et al., 2008, 2011). Using neighborhood rough set model to select attribute of decision table containing continuous data can keep classification ability well and need not discretize it. The existing neighborhood rough set attribute reduction methods are based on the perspective of algebra or information theory. The definition of attribute significance based on algebra theory only describes the influence of attributes on the definite classification subset contained in the universe. The definition of attribute significance based on information theory only describes the influence of attributes on uncertain classification subsets contained in the universe. A single perspective is not comprehensive (Jiang et al., 2015).

Zadeh (1979) proposed the concept of information granulation based on fuzzy sets theory. Objects in the universe are granulated into a set of fuzzy information granules by a fuzzy-binary relation (Tsang et al., 2008; Jensen and Shen, 2009).

In this article, a heuristic feature selection algorithm based on fuzzy information granules and approximate conditional entropy is designed to improve the classification performance of gene expression data sets. The experimental results for several gene expression data sets show that the proposed algorithm can find optimal reduction sets with few genes and high classification accuracy.

The remainder of this article is organized as follows. Section "Materials and Methods" gives the gene expression datasets for the experiment and our feature selection algorithm. Section "Experimental Results and Analysis" shows and analyzes the experimental results. Section "Conclusion and Discussion" summarizes this study and discusses future research focus.

MATERIALS AND METHODS

Gene Expression Data Sets

The following six gene expression datasets are used in this article.

(1) Leukemia1 dataset consists of 7129 genes and 72 samples with two subtypes: patients and healthy people (Sun et al., 2019a).
(2) Leukemia2 dataset consists of 5327 genes and 72 samples with three subtypes: ALL-T (acute lymphoblastic leukemia, T-cell), ALL-B (acute lymphoblastic leukemia, B-cell), and AML (acute myeloid leukemia) (Dong et al., 2018).
(3) Brain Tumor dataset consists of 10,367 genes and 50 samples with four subtypes (Huang et al., 2017).
(4) 9_Tumors dataset consists of 5726 genes and 60 samples with nine subtypes: non-small cell lung cancer, colon cancer, breast cancer, ovarian cancer, leukemia, kidney cancer, melanoma, prostate cancer, and central nervous system cancer (Ye et al., 2019).
(5) Robert dataset consists of 23,416 genes and 194 samples with two subtypes: Musculus CD8+T-cells and L1210 cells (Kimmerling et al., 2016).
(6) Ting dataset consists of 21,583 genes and 187 samples with seven subtypes: GMP cells, MEF cells, MP cells, nb508 cells, TuGMP cells, TuMP cells, and WBC cells (Ting et al., 2014).

The six gene expression datasets are summarized in **Table 1**.

Fuzzy Sets and Fuzzy-Binary Relation

Let $U = \{x_1, x_2, \ldots, x_n\}$ be a nonempty finite set and denote a universe, $I = [0, 1]$, I^U denotes all fuzzy sets on U.

Fuzzy sets are regarded as the extensions of classical sets (Zadeh, 1965).

F is a fuzzy set on U, i.e., $F : U \rightarrow I$, then $F(x_i)$ is the membership degree of x_i to F.

The cardinality of $F \in I^U$ is $|F| = \sum_{i=1}^{n} F(x_i)$.

TABLE 1 | Description of six experimental datasets.

No.	Datasets	Genes	Samples	Classes
1	Leukemia1	7129	72	2 (47/25)
2	Leukemia2	5327	72	3 (9/38/25)
3	Brain_Tumor	10,367	50	4 (14/7/14/15)
4	9_Tumors	5726	60	9 (9/7/8/6/6/8/8/2/6)
5	Robert	23,416	194	2 (88/106)
6	Ting	21,583	187	7 (18/12/75/16/20/34/12)

Fuzzy-binary relation are fuzzy sets on two universes. $I^{U\times U}$ denotes all fuzzy-binary relations on $U \times U$.

Fuzzy-binary relation R can be represented by

$$M_R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \tag{1}$$

where $r_{ij} = R(x_i, x_j) \in I$ is the similarity of x_i and x_j .

Information Systems and Rough Sets

Definition 2.1 (Li et al., 2017). Let Ube a set of objects and A a set of attributes. Suppose that U and A are finite sets. If each attribute $a \in A$ determines an information function $a : U \to V_a$, where V_a is the set of function values of attribute a, then the pair (U, A) is called an information system.

Moreover, if $A = C \bigcup D$, C is a condition attribute set and D is a decision attribute set, then the pair (U, A) is called a decision information system.

If (U, A) is an information system and $P \subseteq A$, then an equivalence relation (or indiscernibility relation) $ind(P)$ can be defined by $(x, y) \in ind(P) \Leftrightarrow \forall a \in P, a(x) = a(y)$.

Obviously, $ind(P) = \bigcap\limits_{a\in P} ind(\{a\})$.

For $P \subseteq A$ and $x \in U$, denote $[x]_{ind(P)} = \{y \,|(x, y) \in ind(P)\}$ and $U/ind(P) = \{[x]_{ind(P)} \,|x \in U \}$.

Usually, $[x]_{ind(P)}$ and $U/ind(P)$ are briefly denoted by $[x]_P$ and U/P, respectively.

According to the rough set theory, for $P \subseteq A$, $X \subseteq U$ is characterized by $\bar{P}(X)$ and $\underline{P}(X)$, where $\underline{P}(X) = \bigcup\{Y \mid Y \in U/P, Y \subseteq X\}$ and $\bar{P}(X) = \bigcup\{Y \mid Y \in U/P, Y \bigcap X \neq \phi \}$.

$\underline{P}(X)$ and $\bar{P}(X)$ are referred to as the lower and upper approximations of X, respectively.

X is crisp if $\bar{P}(X) = \underline{P}(X)$ and X is rough if $\bar{P}(X) \neq \underline{P}(X)$.

The Approximately Equal Relation on Fuzzy Sets

Given $F, G \in I^U$. For $x \in U$, $F(x)$ and $G(x)$ are the membership degrees of x belonging to fuzzy sets F and G, respectively. $F(x)$ and $G(x) \in [0, 1]$. Actually, it is very difficult to ensure that the equation $F(x) = G(x)$ holds. For this reason, we propose the following approximately equal relation of fuzzy sets.

Definition 2.2 Given $A, B \in I^U$. If there exists $k \in N$ $(k \geq 2)$ such that for any $x \in U$, $A(x), B(x) \in [0, 1/k)$ or $A(x), B(x) \in [1/k, 2/k)\ldots$ or $A(x), B(x) \in [(k-1)/k, 1]$, then we say that A is approximately equal to B, and denote it by $A \overset{k}{\approx} B$, where k is regarded as a threshold value.

Definition 2.3 For each $a \in U$, define $x^R : U \to [0, 1]$, $x^R(a) = R(x, a)$ $(x \in U)$, x^R is referred to as a fuzzy set that means the membership degree of a to x.

Definition 2.4 $[x]_R = \{y \,\Big|\, x^R(a) \overset{k}{\approx} y^R(a) \ , y \in U\}$, $[x]_R$is referred to as the fuzzy equal class of x induced by the fuzzy relation R on U.

Definition 2.5 $[x_i]_R (i = 1, 2, \ldots, |U|)$ is named as the fuzzy information granule induced by the fuzzy relation R on U.

Definition 2.6$G(R) = \{[x_1]_R, [x_2]_R, \ldots, [x_n]_R\}$ is referred to as the fuzzy-binary granular structure of the universe U induced by R.

It is easy to prove: $\underline{P}(X) = \{x \mid [x]_R \subseteq X, [x]_R \in G(R)\}$, $\overline{P}(X) = \{x \mid [x]_R \bigcap X \neq \phi, [x]_R \in G(R)\}$.

Fuzzy-Binary Relation Based on Laplacian Kernel

Hu et al. (2010) found that there are some relationships between rough sets and Gaussian kernel method, so Gaussian kernel is used to obtain fuzzy relations. Compared with Gaussian kernel, Laplacian kernel has higher peak, faster reduction and smoother tail. Therefore, Laplacian kernel is better than Gaussian kernel in describing the similarity between objects. In this article, we use Laplacian kernel $k(x_i, x_j) = \exp(-\frac{||x_i - x_j||}{\sigma})$ to extract the similarity between two objects from decision information system, where $||x_i - x_j||$ is the Euclidean distance between two objects x_i and x_j. In general, σ is a given positive value.

Obviously, $k(x_i, x_j)$ satisfies:

(1) $k(x_i, x_j) \in (0, 1]$.
(2) $k(x_i, x_j) = k(x_j, x_i)$.
(3) $k(x_i, x_i) = 1$.

Let $R = (k(x_i, x_j))_{n\times n}$, then R is called the fuzzy relation matrix induced by Laplacian kernel.

Feature Selection Using Approximate Conditional Entropy Based on Fuzzy Information Granule

Approximate Accuracy and Approximate Conditional Entropy

Definition 2.7 Given a decision information system $(U, C \bigcup D)$, $\forall X \subseteq U$, $X \neq \phi$ (ϕ is an empty set), then the approximate accuracy of X is defined as

$$a(X) = \frac{\left|\underline{P}(X)\right|}{\left|\bar{P}(X)\right|} \tag{2}$$

where $|.|$ denotes the cardinality of set. Obviously, $0 \leq a(X) \leq 1$.

Definition 2.8 Given a decision information system $(U, C \bigcup D)$, $\forall B \subseteq C$, the fuzzy information granule of object x under B is $[x]_{R_B}$, the partition of U derived from D is $\{X_1, X_2, \ldots, X_k\}$, then the conditional entropy of D relative to B is defined as

$$H(D/B) = -\sum_{j=1}^{k}\sum_{i=1}^{|U|} \frac{\left|[x_i]_{R_B} \bigcap X_j\right|}{|U|} \log \frac{\left|[x_i]_{R_B} \bigcap X_j\right|}{\left|[x_i]_{R_B}\right|} \tag{3}$$

where R_B denotes the fuzzy relation based on attribute set B and log is a base-2 logarithm.

The approximate accuracy can effectively measure the imprecision of the set caused by the boundary region, while the conditional entropy can effectively measure the knowledge uncertainty caused by the information granularity. We combine the two to propose approximate conditional entropy.

Definition 2.9 Let $(U, C\bigcup D)$ be a decision information system, $\forall B\subseteq C$, the fuzzy information granule of object x under B is $[x]_{R_B}$, the partition of U derived from D is $\{X_1, X_2, ..., X_k\}$, $a_B(X_i)$ is the approximate accuracy of X_i under R_B, then the approximate conditional entropy of D relative to B is defined as

$$H_{ace}(D/B) = -\sum_{j=1}^{k}\sum_{i=1}^{|U|}\log(2-a_B(X_j))\frac{\left|[x_i]_{R_B}\bigcap X_j\right|}{|U|}$$

$$\log\frac{\left|[x_i]_{R_B}\bigcap X_j\right|}{\left|[x_i]_{R_B}\right|} \quad (4)$$

Theorem 2.1 Let $(U, C\bigcup D)$ be a decision information system, $\forall B\subseteq C$, the fuzzy information granule of object x under B is $[x]_{R_B}$, the partition of U derived from D is $\{X_1, X_2, ..., X_k\}$.

(1) $H_{ace}(D/B)$ gets the maximum value $|U|\log|U|$ if and only if $[x_i]_{R_B} = U(i = 1, 2, ..., n)$ and $\left|X_j\right| = 1(j = 1, 2, ..., k = n)$.

(2) $H_{ace}(D/B)$gets the minimum value 0 if and only if $[x_i]_{R_B}\subseteq [x_i]_{R_D}(i = 1, 2, ..., n)$.

Proof. (1) Due to $[x_i]_{R_B} = U(i = 1, 2, ..., n)$ and $\left|X_j\right| = 1(j = 1, 2, ..., k)$, we have $a_B(X_j) = 0(\mathrm{j} = 1, 2, ..., \mathrm{k})$ according to Definition 2.7.

Thus, $\log(2 - a_B(X_j)) = 1(j = 1, 2, ..., k)$.

Clearly, $\frac{\left|[x_i]_{R_B}\bigcap X_j\right|}{|U|}\log\frac{\left|[x_i]_{R_B}\bigcap X_j\right|}{\left|[x_i]_{R_B}\right|} = \frac{1}{|U|}\log\frac{1}{|U|}$.

By Definition 2.9, we have $H_{ace}(D/B) = |U|\log|U|$.

The converse is also true.

(2) Due to $[x_i]_{R_B}\subseteq [x_i]_{R_D}(i = 1, 2, ..., n)$, we have $a_B(X_j) = 1(\mathrm{j} = 1, 2, ..., \mathrm{k})$ according to Definition 2.7. Thus $\log(2 - a_B(X_j)) = 0(\mathrm{j} = 1, 2, .., \mathrm{k})$. Obviously, $H_{ace}(D/B) = 0$ according to Definition 2.9.

The converse is also true.

Theorem 2.2 Let $(U, C\bigcup D)$ be a decision information system, $\forall L, M\subseteq C$, if $M\subseteq L$, then $H_{ace}(D/M)\geq\ H_{ace}\ (D/L)$.

Proof. Due to $M\subseteq L\subseteq C$, we have $\underline{P_M}(X)\subseteq\underline{P_L}(X)$ and $\overline{P_M}(X)\supseteq\overline{P_L}\ (X)$.

Then $a_M(X)\leq a_L(X)$ according to Definition 2.7.

By $M\subseteq L$ and $U/D = \{X_1, X_2, ..., X_k\}$, we have

$$-\frac{\left|[x_i]_{R_M}\bigcap X_j\right|}{|U|}\log\frac{\left|[x_i]_{R_M}\bigcap X_j\right|}{\left|[x_i]_{R_M}\right|}$$

$$\geq -\frac{\left|[x_i]_{R_L}\bigcap X_j\right|}{|U|}\log\frac{\left|[x_i]_{R_L}\bigcap X_j\right|}{\left|[x_i]_{R_L}\right|}\geq 0 \quad (5)$$

Consequently, $H_{ace}(D/M)\geq H_{ace}(D/L)$ according to Definition 2.9.

Theorem 2.2 shows that $H_{ace}(D/B)$ decreases monotonically with the increase of the number of attributes in B, which is very important for constructing forward greedy algorithm of attributes reduction.

Definition 2.10 Let $(U, C\bigcup D)$ be a decision information system and $B\subseteq C$, if $H_{ace}(D/B) = H_{ace}(D/C)$ and $H_{ace}(D/(B - \{b\})) > H_{ace}(D/C)(\forall b\in B)$, then B is called a reduction of C relative to D.

The first condition guarantees that the selected attribute subset has the same amount of information as the whole attribute set. The second condition guarantees that there is no redundancy in the attribute reduction set.

Definition 2.11 Assume that $(U, C\bigcup D)$ be a decision information system, $\forall c\in C$, define the following indicator,

$$IIA(c, C, D) = H_{ace}(D/(C - \{c\})) - H_{ace}(D/C) \quad (6)$$

then $IIA(c, C, D)$ is called the importance of internal attribute of c in C relative to D.

Definition 2.12 Assume that $(U, C\bigcup D)$ be a decision information system, $\forall c\in C$, if $IIA(c, C, D) > 0$, then attribute c is called a core attribute of C relative to D.

Definition 2.13 Assume that $(U, C\bigcup D)$ be a decision information system, $B\subseteq C$, $\forall d\in C{-}B$, define the following indicator,

$$IEA(d, B, C, D) = H_{ace}(D/B) - H_{ace}(D/(B\bigcup\{d\})) \quad (7)$$

then $IEA(d, B, C, D)$ is called the importance of external attribute of d to B relative to D.

$IEA(d, B, C, D)$ shows the change of approximate conditional entropy after adding attribute d. The larger $IEA(d, B, C, D)$ is, the more important d is to B relative to D.

Feature Selection Algorithm Using Approximate Conditional Entropy

In this article, a novel feature selection algorithm using approximate conditional entropy (FSACE) is proposed and described as follows.

```
Input: A decision information system $(U, C\bigcup D)$ and $\sigma$.
Output:A selected gene subset $B$.
Step 1. Initialize $B = \phi$.
Step 2. Compute $H_{ace}(D/C)$.
Step 3.$\forall c\in C$, compute $IIA(c, C, D)$, if $IIA(c, C, D) > 0$, then $B = B\bigcup\{c\}$.
Step 4. If $B = \phi$, then turn to step 5. If $B\neq\phi$, compute $H_{ace}(D/B)$. If
    $H_{ace}(D/B) = H_{ace}(D/C)$, then turn to step 6; otherwise, turn to step 5.
Step 5. Let $M = C - B$, select a attribute $m\in M$ so that it satisfies
    $IEA(m, B, C, D) = \max_{x\in M} IEA(x, B, C, D)$. Let $B = B\bigcup\{m\}$, compute $H_{ace}(D/B)$. If
    $H_{ace}(D/B) = H_{ace}(D/C)$, then turn to step 6; otherwise, turn to step 5.
Step 6. The feature selection subset $B$ is obtained, and the algorithm ends.
```

EXPERIMENTAL RESULTS AND ANALYSIS

All experiments are performed on a personal computer running Windows 10 with an Intel(R) Core(TM) i7-4790 CPU operating at 3.60 GHz with 8 GB memory using MATLAB R2019a. The classifiers (KNN, CART, and SVM) are selected to verify the classification accuracy, where the parameter $k = 3$ in KNN and Gaussian kernel function is selected in SVM. Other parameters of the three algorithms are the default values of the software.

Influence of Different Values of σ on Classification Performance

In this part, the classification accuracy of different Laplacian kernel parameters values of σ is tested. For gene expression data, feature selection aims to improve classification accuracy by eliminating redundant genes. The different values of σ influence the size of granulated gene data, which affects the classification accuracy of selected genes. Therefore, the different values of σ should be set in the process of feature selection of gene expression data sets. Moreover, the different values of σ also affect the composition of the selected gene subset. To obtain a suitable σ and a good gene subset, the classification accuracy of the selected gene subset for different values of σ should be discussed in detail.

The corresponding experiments are performed to graphically illustrate the classification accuracy of FSACE under different values of σ. The results are shown in **Figure 1**, where the horizontal axis denotes σ ∈ [0.05, 1] at intervals of 0.05, and the vertical axis represents the classification accuracy.

Figure 1 shows that σ greatly influences the classification performance of FSACE. σ is usually set to make the classification accuracy highest. Thus, the appropriate parameter values of σ can be obtained for each data set from **Figure 1**. In **Figure 1A**, for Leukemia1 data set, when σ is 0.95, the classification accuracy is the highest. In **Figure 1B**, for Leukemia2 data set, when σ is 0.55, the classification accuracy is the highest. In **Figure 1C**, for Brain tumor data set, when σ is 0.80, the classification accuracy is the highest. In **Figure 1D**, for 9-tumors data set, when σ is 0.75, the classification accuracy is the highest. In **Figure 1E**, for Robert data set, when σ is 0.60, the classification accuracy is the highest. In **Figure 1F**, for Ting data set, when σ is 0.75, the classification accuracy is the highest. Therefore, the appropriate values of σ for different data sets are determined.

The Feature Selection Results and Classification Performance of FSACE

The classification results obtained from the three classifiers (KNN, CART, and SVM) with 10-fold cross-validation are shown in **Table 2** on the test data by FSACE.

Table 2 shows that FSACE not only greatly reduces the dimensionality of all six gene expression data sets, but also improves the classification accuracy.

The results of feature genes selection from six gene expression data sets are shown in **Table 3** using FSACE.

Comparison of the Classification Performance of Several Entropy-Based Feature Selection Algorithms

To evaluate the performance of FSACE in terms of classification accuracy, FSACE algorithm is compared with several state-of-the-art feature selection algorithms, including EGGS (Chen et al., 2017), EGGS-FS (Yang et al., 2016), MEAR (Xu et al., 2009), Fisher (Saqlain et al., 2019), and Lasso (Tibshirani, 1996). According to the change trend of Fisher scores of six gene datasets, we select the top-200 genes as the reduction set for Fisher algorithm.

Tables 4–9 show the experimental results of six gene expression data sets using six different feature selection methods.

As shown in **Tables 4**, **5**, FSACE has the highest average classification accuracy for Leukemia1 and Leukemia2, and

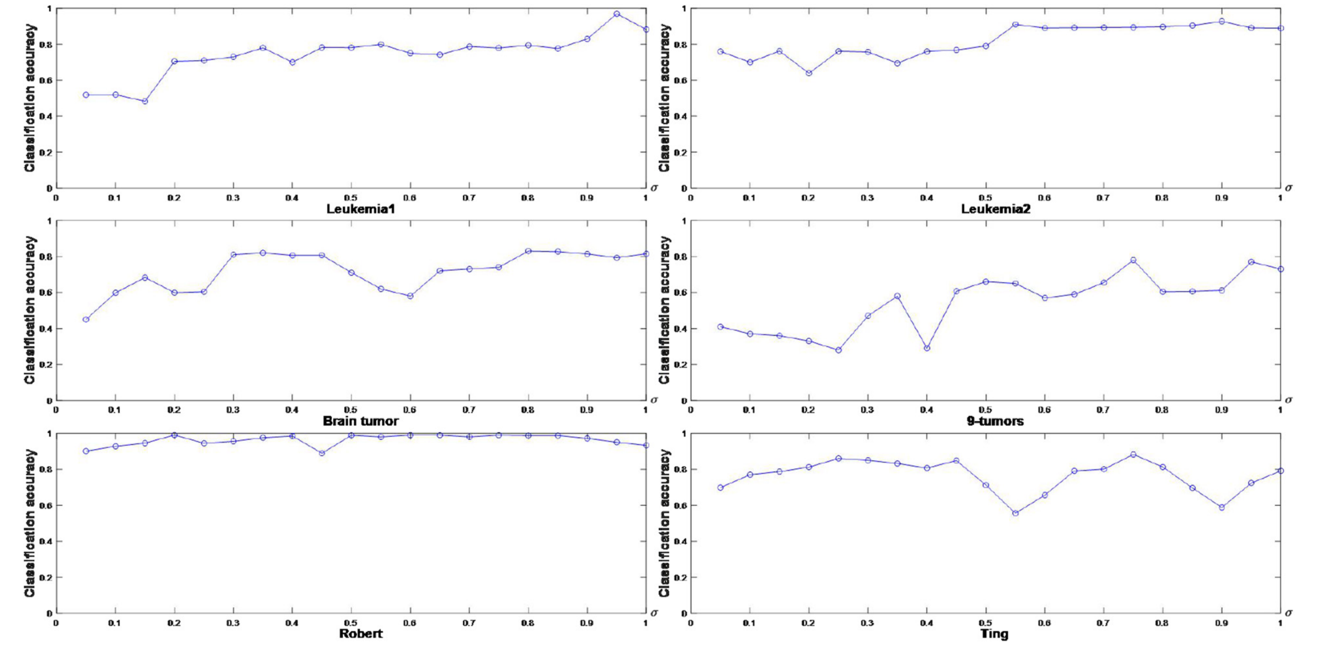


FIGURE 1 | Classification accuracy for six gene expression data sets with different values of **σ**.

TABLE 2 | Classification results of six gene expression data sets.

Data sets	Original data				Feature selection data using FSACE			
	Genes	CART	KNN	SVM	Genes	CART	KNN	SVM
Leukemia1	7129	0.822	0.839	0.917	9	0.911	0.947	0.931
Leukemia2	5327	0.849	0.820	0.834	9	0.891	0.894	0.878
Brain tumor	10,367	0.571	0.604	0.737	5	0.743	0.631	0.614
9-tumors	5726	0.273	0.349	0.334	2	0.318	0.359	0.355
Robert	23,416	0.947	0.928	0.933	14	0.985	0.974	0.990
Ting	21,583	0.864	0.826	0.841	17	0.873	0.847	0.882
Average	12,258	0.721	0.728	0.766	9.333	0.787	0.775	0.775

TABLE 3 | The selected feature genes on six gene expression data sets using FSACE.

Data sets	The selected feature gene subsets
Leukemia1	(758,1144,1630,2659,3897,4196,5552,6471,6584)
Leukemia2	(568,848,861,1610,2197,3256,3358,4688,5032)
Brain tumor	(642,7169,7844,9413,9794)
9-tumors	(1677,2590)
Robert	(12883,1600,9892,16398,8720,4510,18137,2320,14931, 14679,10352,12481,18034,406)
Ting	(4754,5676,2503,5379,3304,4752,6015,2193,15687,641, 7938,2629,6837,4653,19016,8621,4267)

TABLE 4 | Classification accuracy of Leukemia1 using six different feature selection algorithms.

Feature selection method	Genes	CART	KNN	SVM	Average
ECGS (Li et al., 2017)	8	0.744	0.619	0.813	0.725
EGGS-FS (Hu et al., 2010)	5	0.821	0.794	0.701	0.772
MEAR (Chen et al., 2017)	3	0.939	0.919	0.925	0.928
Fisher (Saqlain et al., 2019)	200	0.639	0.857	0.778	0.758
Lasso (Tibshirani, 1996)	52	0.857	0.960	0.972	0.929
FSACE	9	0.911	0.947	0.931	0.930

TABLE 5 | Classification accuracy of Leukemia2 using six different feature selection algorithms.

Feature selection method	Genes	CART	KNN	SVM	Average
ECGS (Li et al., 2017)	3	0.571	0.509	0.557	0.546
EGGS-FS (Hu et al., 2010)	2	0.907	0.871	0.874	0.884
MEAR (Chen et al., 2017)	5	0.903	0.829	0.872	0.868
Fisher (Saqlain et al., 2019)	200	0.726	0.803	0.846	0.792
Lasso (Tibshirani, 1996)	37	0.817	0.914	0.909	0.880
FSACE	9	0.891	0.894	0.878	0.888

exhibits better classification performance than the other five algorithms.

As shown in **Tables 6**, **7**, MEAR cannot work on Brain Tumor data set and 9-tumors data set, its results are denoted by the sign –. FSACE obtains the highest average classification accuracy among the five feature selection algorithms for Brain Tumor data set and 9-tumors data set.

TABLE 6 | Classification accuracy of Brain tumor using six different feature selection algorithms.

Feature selection method	Genes	CART	KNN	SVM	Average
ECGS (Li et al., 2017)	9	0.515	0.491	0.544	0.517
EGGS-FS (Hu et al., 2010)	5	0.388	0.490	0.531	0.470
MEAR (Chen et al., 2017)	–	–	–	–	–
Fisher (Saqlain et al., 2019)	200	0.630	0.704	0.617	0.650
Lasso (Tibshirani, 1996)	–	–	–	–	–
FSACE	5	0.743	0.631	0.614	0.663

TABLE 7 | Classification accuracy of 9-tumors using six different feature selection algorithms.

Feature selection method	Genes	CART	KNN	SVM	Average
ECGS (Li et al., 2017)	1	0.177	0.102	0.672	0.317
EGGS-FS (Hu et al., 2010)	1	0.224	0.203	0.393	0.273
MEAR (Chen et al., 2017)	–	–	–	–	–
Fisher (Saqlain et al., 2019)	200	0.249	0.335	0.414	0.333
Lasso (Tibshirani, 1996)	27	0.199	0.361	0.322	0.294
FSACE	2	0.318	0.359	0.355	0.344

TABLE 8 | Classification accuracy of Robert using six different feature selection algorithms.

Feature selection method	Genes	CART	KNN	SVM	Average
ECGS (Li et al., 2017)	11	0.948	0.937	0.964	0.950
EGGS-FS (Hu et al., 2010)	6	0.957	0.954	0.975	0.962
MEAR (Chen et al., 2017)	–	–	–	–	–
Fisher (Saqlain et al., 2019)	200	0.976	0.990	0.989	0.985
Lasso (Tibshirani, 1996)	21	0.984	0.991	0.989	0.988
FSACE	14	0.993	0.991	0.985	0.990

TABLE 9 | Classification accuracy of Ting using six different feature selection algorithms.

Feature selection method	Genes	CART	KNN	SVM	Average
ECGS (Li et al., 2017)	12	0.793	0.781	0.651	0.742
EGGS-FS (Hu et al., 2010)	9	0.745	0.717	0.626	0.696
MEAR (Chen et al., 2017)	–	–	–	–	–
Fisher (Saqlain et al., 2019)	200	0.833	0.779	0.770	0.794
Lasso (Tibshirani, 1996)	56	0.833	0.833	0.845	0.837
FSACE	17	0.833	0.833	0.872	0.846

Tables 8, **9** shows that MEAR still can not work on Robert data set and Ting data set, which indicates that the algorithm is not stable. Our algorithm still has the highest classification accuracy among all the algorithms. Although the classification accuracy of our algorithm is only a little higher than lasso algorithm, the number of attributes reduced by our algorithm is much less than lasso algorithm.

Tables 4–9 show that the average number of attributes reduced by our algorithm is slightly more than that of MEAR, ECGS, and EGGS-FS, but the average classification accuracy is much higher than that of these three algorithms.

Therefore, FSACE can not only effectively remove noise and redundant data from the original data, but also improve the classification accuracy of gene expression data sets.

CONCLUSION AND DISCUSSION

Firstly, the concept of approximate conditional entropy is given and its monotonicity is proved in this article. Approximate conditional entropy can describe the uncertainty of knowledge from two aspects of boundary and information granule. And then, a novel feature selection algorithm FSACE is proposed based on the approximate conditional entropy. Finally, the effectiveness of the proposed algorithm is verified on several gene expression data sets. Experimental results show that compared with several state-of-the-art feature selection algorithms, the proposed feature selection algorithm not only can obtain compact features, but also improve classification performance. The time complexity of FSACE is $O(|U|^2\,|C|^2)$. Because the gene expression data sets usually contain a large number of genes, the time complexity of FSACE is high. In addition, FSACE does not consider the interaction between attributes. Therefore, reducing the time complexity of FSACE and seeking more efficient feature selection algorithm considering interaction between attributes are two issues that we will study in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi (cancer Program Legacy Publication Resources).

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Chen, Y., Zhang, Z., Zheng, J., Ying, M., and Yu, X. (2017). Gene selection for tumor classification using neighborhood rough sets and entropy measures. *J. Biomed. Inform*. 67, 59–68. doi: 10.1016/j.jbi.2017.02.007

Dai, J., and Xu, Q. (2012). Approximations and uncertainty measures in incomplete information systems. *Inf. Sci*. 198, 62–80. doi: 10.1016/j.ins.2012.02.032

Dong, H., Li, T., Ding, R., and Sun, J. (2018). A novel hybrid genetic algorithm with granular information for feature selection and optimization. *Appl. Soft Comput*. 65, 33–46. doi: 10.1016/j.asoc.2017.12.048

Fu, X., and Wang, L. (2003). Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. *IEEE Trans. Syst. Man Cybern. Part B Cybern*. 33, 399–409. doi: 10.1109/tsmcb.2003.810911

Hu, L., Gao, W., Zhao, K., Zhang, P., and Wang, F. (2018). Feature selection considering two types of feature relevancy and feature interdependency. *Expert Syst. Appl*. 93, 423–434. doi: 10.1016/j.eswa.2017.10.016

Hu, Q., Yu, D., Liu, J., and Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Inf. Sci*. 178, 3577–3594. doi: 10.1016/j.ins.2008.05.024

Hu, Q., Zhang, L., Chen, D., Witold, P., and Daren, Y. (2010). Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications. *Int. J. Approx. Reason*. 51, 453–471. doi: 10.1016/j.ijar.2010.01.004

Hu, Q., Zhang, L., Zhang, D., Wei, P., Shuang, A., and Witold, P. (2011). Measuring relevance between discrete and continuous features based on neighborhood mutual information. *Expert Syst. Appl*. 38, 10737–10750. doi: 10.1016/j.eswa.2011.01.023

Huang, X., Zhang, L., Wang, B., Li, F. Z., and Zhang, Z. (2017). Feature clustering based support vector machine recursive feature elimination for gene selection. *Appl. Intell*. 48, 1–14.

Jensen, R., and Shen, Q. (2009). New approaches to fuzzy-rough feature selection. *IEEE Trans. Fuzzy Syst*. 17, 824–838. doi: 10.1109/tfuzz.2008.924209

Jiang, F., Wang, S., Du, J., and Sui, Y. F. (2015). Attribute reduction based on approximation decision entropy. *Control and Decis*. 30, 65–70. doi: 10.3390/e20010065

Kimmerling, R., Szeto, G., Li, J., Alex, S. G., Samuel, W. K., Kristofor, R. P., et al. (2016). A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nat. Commun*. 7:10220.

Konstantina, K., Themis, P., Konstantinos, P. E., Michalis, V. K., and Dimitrios, I. F. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J*. 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Li, Z., Liu, X., Zhang, G., Xie, N., and Wang, S. (2017). A multi-granulation decision-theoretic rough set method for distributed fc-decision information systems: an application in medical diagnosis. *Appl. Soft Comput*. 56, 233–244. doi: 10.1016/j.asoc.2017.02.033

Mitra, S., Das, R., and Hayashi, Y. (2011). Genetic networks and soft computing. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 8, 94–107.

Pawlak, Z. (1982). Rough sets. *Int. J. Comput. Inf. Sci*. 11, 341–356.

Phan, J., Quo, C., and Wang, M. (2012). Cardiovascular genomics: a biomarker identification pipeline. *IEEE Trans. Inf. Technol. Biomed*. 16, 809–822. doi: 10.1109/titb.2012.2199570

Qian, Y., Liang, J., Wu, W., and Dang, C. (2011). Information granularity in fuzzy binary GrC model. *IEEE Trans. Fuzzy Syst*. 19, 253–264. doi: 10.1109/tfuzz.2010.2095461

Saqlain, S. M., Sher, M., Shah, F. A., Khan, I., Ashraf, M. U., Awais, M., et al. (2019). Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines[J]. *Knowl. Inf. Syst*. 58, 139–167. doi: 10.1007/s10115-018-1185-y

Sun, L., Wang, L., Xu, J., and Zhang, S. (2019a). A neighborhood rough sets-based attribute reduction method using Lebesgue and entropy measures. *Entropy* 21, 1–26.

Sun, L., Zhang, X., Qian, Y., Xu, J., and Zhang, S. (2019b). Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. *Inf. Sci*. 502, 18–41. doi: 10.1016/j.ins.2019.05.072

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol*. 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Ting, D., Wittner, B., Ligorio, M., Brian, W. B., Ajay, M. S., Xega, K., et al. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep*. 8, 1905–1918. doi: 10.1016/j.celrep.2014.08.029

Tsang, E., Chen, D., Yeung, D., Wang, X. Z., and Lee, J. W. T. (2008). Attributes reduction using fuzzy rough sets. *IEEE Trans. Fuzzy Syst*. 16, 1130–1141. doi: 10.1109/tfuzz.2006.889960

Wang, C., Shi, Y., Fan, X., and Shao, M. W. (2019). Attribute reduction based on k-nearest neighborhood rough sets. *Int. J. Approx. Reason*. 106, 18–31. doi: 10.1016/j.ijar.2018.12.013

Xu, F., Miao, D., and Wei, L. (2009). Fuzzy-rough attribute reduction via mutual information with an application to cancer classification. *Comput. Math. Appl*. 57, 1010–1017. doi: 10.1016/j.camwa.2008.10.027

Yang, J., Liu, Y., Feng, C., and Zhu, G. Q. (2016). Applying the fisher score to identify Alzheimer's disease-related genes. *Genet. Mol. Res.* 15, 1–9.

Ye, C., Pan, J., and Jin, Q. (2019). An improved SSO algorithm for cyber-enabled tumor risk analysis based on gene selection. *Future Gener. Comput. Syst.* 92, 407–418. doi: 10.1016/j.future.2018.10.008

Zadeh, L. (1965). Fuzzy sets. *Inf. Control* 8, 338–353.

Zadeh, L. (1979). *Fuzzy Sets and Information Granularity, Advance in Fuzzy Set Theory & Application*. Amsterdam: North Holland Publishing, 3–18.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 13 April 2021
doi: 10.3389/fgene.2021.660275

Predicting Metabolite–Disease Associations Based on LightGBM Model

Cheng Zhang, Xiujuan Lei and Lian Liu*

School of Computer Science, Shaanxi Normal University, Xi'an, China

Metabolites have been shown to be closely related to the occurrence and development of many complex human diseases by a large number of biological experiments; investigating their correlation mechanisms is thus an important topic, which attracts many researchers. In this work, we propose a computational method named LGBMMDA, which is based on the Light Gradient Boosting Machine (LightGBM) to predict potential metabolite–disease associations. This method extracts the features from statistical measures, graph theoretical measures, and matrix factorization results, utilizing the principal component analysis (PCA) process to remove noise or redundancy. We evaluated our method compared with other used methods and demonstrated the better areas under the curve (AUCs) of LGBMMDA. Additionally, three case studies deeply confirmed that LGBMMDA has obvious superiority in predicting metabolite–disease pairs and represents a powerful bioinformatics tool.

Keywords: metabolite-disease associations, light gradient boosting machine, features, computational method, performance evaluation

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Qi Zhao,
University of Science and Technology Liaoning, China
Wei Peng,
Kunming University of Science and Technology, China

****Correspondence:***
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *29 January 2021*
Accepted: *05 March 2021*
Published: *13 April 2021*

Citation:
Zhang C, Lei X and Liu L (2021) Predicting Metabolite–Disease Associations Based on LightGBM Model. Front. Genet. 12:660275. doi: 10.3389/fgene.2021.660275

INTRODUCTION

Metabolism is a series of ordered chemical reactions, which has a significant influence on biological life maintenance, such as the organism's growth, reproduction, and reaction to the external environment (Dunn and Ellis, 2005). Metabolic processes are usually divided into two types. The first type is decomposing large molecules to acquire energy, such as cell respiration, while the other type is utilizing energy for the synthesis of each part inside the cells, such as nucleic acids and proteins (Cheng et al., 2017). In unhealthy conditions, relevant products in metabolism (metabolites) will be abnormal, which indicates that finding more disease-related metabolites is beneficial to disease prevention and treatment (Boja et al., 2014). Consequently, it is of high importance to identify the relationship among metabolites and diseases.

Although some traditional techniques of metabolomics has prompted their analysis and development, such as nuclear magnetic resonance (NMR) spectroscopy or liquid/gas chromatography-mass spectrometry (LC/GC-MS) (Xianlin et al., 2011; Tang et al., 2014), the proportion of undiscovered associations between metabolites and diseases is still high. Moreover, some limitations exist, such as the time and funds required to mine disease-related metabolites in biological experiments. Therefore, effective computational methods for predicting disease-related metabolites are attracting more and more attention, which is beneficial to promoting the

Abbreviations: AUC, area under the curve; GIP, gaussian interaction profile; LOOCV, leave-one-out cross-validation; ROC, receiver operating characteristic.

development to discover potential metabolite–disease associations. The idea of Random Walk with Restart for MiRNA-Disease Association (RWRMDA) (Hu et al., 2018) is to construct a metabolite–metabolite functional similarity network and implement RWR from known disease-related metabolite seed nodes to prioritize potential disease-related ones. However, this method uses less information for diseases or metabolites when calculating similarities, and its predictive performance needs to be improved.

In this article, we present a computational method, LGBMMDA, based on Light Gradient Boosting Machine (LightGBM) (Ke et al., 2017), to identify metabolite–disease associations (**Figure 1**). Firstly, we extract the data of metabolite-related pathways as part of the integrated similarity network. Secondly, features are selected from the relevant similarity network and known metabolite–disease associations using the statistical measures, graph theoretical measures, and matrix factorization measures. Furthermore, the principal component analysis (PCA) (Deutsch, 2004) algorithm, which is a technique for analyzing and simplifying datasets, is utilized to extract deep features. Thirdly, the LightGBM classifier is utilized to obtain the potential association scores. In addition, the LOOCV and fivefold cross-validation are adopted to evaluate the performance of LGBMMDA, which achieves 0.9738 and 0.9715, respectively. Besides, three types of case studies for common diseases are carried out to evaluate the ability of the method to predict metabolites. These aforementioned experiments show that LGBMMDA is a reliable and excellent model to infer unknown metabolites–diseases associations.

MATERIALS AND METHODS

Human Metabolite–Disease Associations

We extracted the experimentally confirmed human metabolite–disease associations from the last updated database (HMDB) (Wishart et al., 2017). Then, we performed the following steps on these associations: Firstly, the disease-related symptoms from the human symptom–disease network (HSDN) (Zhou et al., 2014; Ma et al., 2016) are used to calculate disease similarity after repeated associations; thus, the diseases that do not exist in the HSDN are removed. Secondly, the metabolite-related pathways from HMDB become part of the metabolite similarities, such that we keep the metabolites that are relevant to the diseases we selected. Finally, we obtain 127 diseases and 794 metabolites, which have 1,908 experimentally human metabolite–disease associations (see **Figure 2**). The parameters *nm* and *nd* represent the number of metabolites and diseases, respectively. Matrix *M* represents the adjacency matrix of metabolite–disease associations, such that the entity $M(i,j)$ in row *i* and column *j* is 1 if disease *i* is associated with metabolite *j* and 0 otherwise.

Metabolite Functional Similarity

According to the hypothesis that metabolites with similar functions have a higher probability of possessing similar pathways, we utilize the Hamming similarity (Charikar, 2002) to measure the functional similarity of two metabolites by considering their related pathways. The metabolite functional similarity matrix is defined as $MHS_{(nm \times nm)}$, such that the element value is calculated as follows (Zhang et al., 2020)

$$MHS\left(m_i, m_j\right) = 1 - \frac{\sum_{k=1}^{np} MpV\left(MP\left(k, i\right), MP\left(k, j\right)\right)}{ns} \tag{1}$$

$$MpV\left(MP\left(k, i\right), MP\left(k, j\right)\right) = \begin{cases} 1, & \textit{if the values of } MP\left(k, i\right) \textit{ and } MP\left(k, j\right) \textit{ are different} \\ 0, & \textit{if the values of } MP\left(k, i\right) \textit{ and } MP\left(k, j\right) \textit{ are same} \end{cases} \tag{2}$$

where $MHS\left(m_i, m_j\right)$ represents the Hamming similarity between metabolite node m_i and m_j; *np* denotes the number of pathways. If there are existing associations between the metabolite *i* and pathway *k*, $MP\left(k, i\right)$ is set to 1 in metabolite-pathway association network (*MP*).

Disease Functional Similarity

Considering the assumption that two diseases with similar functions usually have similar symptoms, the values of two disease-related symptom sets are used to obtain their functional similarities. Let the sets $S_d{}^a = \{S_d{}^a(1), S_d{}^a(2), S_d{}^a(as)\}$ and sets $S_d{}^b = \left\{S_d{}^b(1), S_d{}^b(2), S_d{}^b\left(bs\right)\right\}$ describe the symptom sets of diseases *a* and *b*, where *as* and *bs* denote the number of symptoms associated with diseases *a* and *b*, respectively. Firstly, we calculate the information entropy of $S_d{}^a$ as follows (Gu et al., 2017)

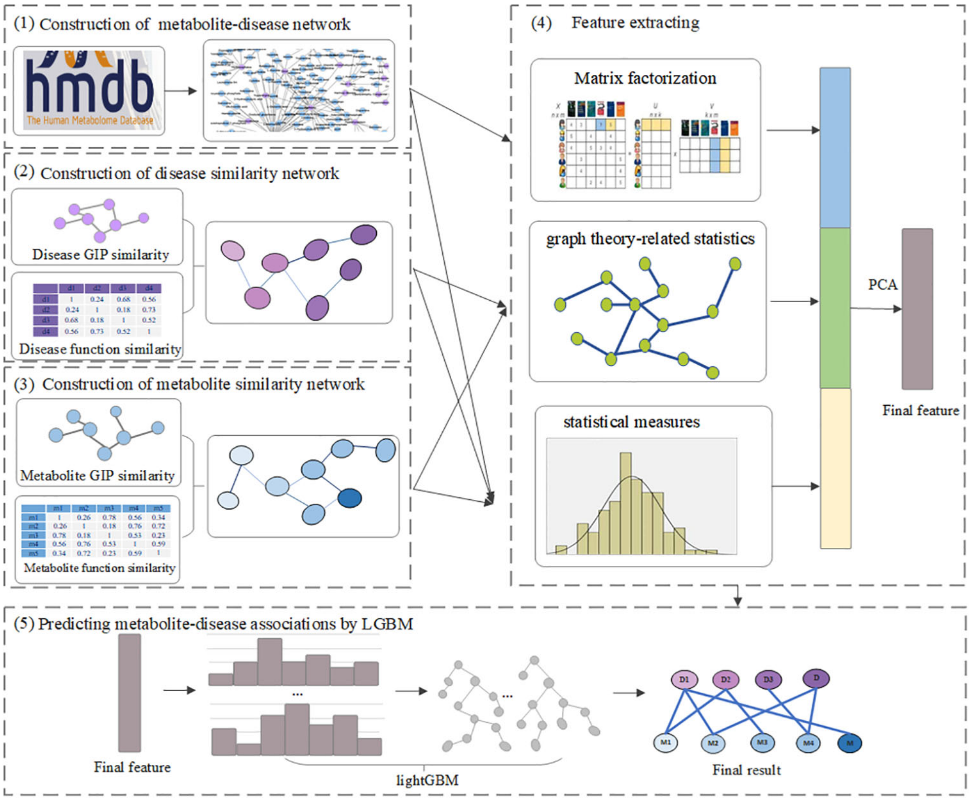
$$H\left(S_d{}^a\right) = -\sum_{i=1}^{ns} p\left(S_d{}^a(i)\right)\{\log_2 p\left(S_d{}^a(i)\right)\} \tag{3}$$

$$p\left(S_d{}^a(i)\right) = \frac{n(S_d{}^a(i))}{Tn} \tag{4}$$

where *Tn* denotes the number of disease-symptom associations, $n(S_d{}^a(i))$ is the number of the *i*th symptom related with disease *a* in the disease-symptom set, $p\left(S_d{}^a(i)\right)$ represents the frequency about the *i*th symptom associated with disease *a*, and $H\left(S_d{}^a\right)$ is the information entropy of $S_d{}^a$. The normalized mutual information (NMI) of $S_d{}^a$ and $S_d{}^b$ is used to measure the functional similarity between disease *a* and *b* as follows:

$$DNF\left(d_a, d_b\right) = \frac{2H\left(S_d{}^a \bigcap S_d{}^b\right)}{H\left(S_d{}^a\right) + H\left(S_d{}^b\right)} \tag{5}$$

where matrix *DNF* represents the functional similarity matrix; $S_d{}^a$, $S_d{}^b$, and $H\left(S_d{}^a \bigcap S_d{}^b\right)$ denote the information entropy of $S_d{}^a$, $S_d{}^b$ and the intersection set of $S_d{}^a$ and $S_d{}^b$, respectively.


(1) Construction of metabolite-disease network
hmdb
The Human Metabolome Database
(2) Construction of disease similarity network
Disease GIP similarity
Disease function similarity
(3) Construction of metabolite similarity network
Metabolite GIP similarity
Metabolite function similarity
(4) Feature extracting
Matrix factorization
graph theory-related statistics
statistical measures
PCA
Final feature
(5) Predicting metabolite-disease associations by LGBM
Final feature
lightGBM
Final result
D1
D2
D3
D
M1
M2
M3
M4
M

	d1	d2	d3	d4
d1	1	0.24	0.68	0.56
d2	0.24	1	0.18	0.73
d3	0.68	0.18	1	0.52
d4	0.56	0.73	0.52	1

	m1	m2	m3	m4	m5
m1	1	0.26	0.78	0.56	0.34
m2	0.26	1	0.18	0.76	0.72
m3	0.78	0.18	1	0.53	0.23
m4	0.56	0.76	0.53	1	0.59
m5	0.34	0.72	0.23	0.59	1

FIGURE 1 | The flowchart of LGBMMDA.

Gaussian Interaction Profile Kernel Similarity

Following literature (Gu et al., 2017) the GIP kernel for the similarities about diseases and metabolites captures the key features of the metabolite–disease association data. Calculating such kind of similarities is based on the assumption that similar diseases are more likely to contain functionally similar metabolites, and vice versa. Let the binary vector $V(d_i)$, which is the row vector of the matrix M where the disease d_i is located, represent the interaction profiles of disease d_i. Then, the relevant similarities for diseases $DGS(d_i, d_j)$ between the diseases d_i and d_j can be shown as follows:

$$DGS(d_i, d_j) = exp\left(-\omega_d ||V(d_i) - V(d_i)||^2\right) \tag{6}$$

$$\omega_d = \omega'{}_d/(\frac{1}{nd}\sum_{i=1}^{nd} ||V(d_i)||^2) \tag{7}$$

where ω_d is a parameter that controls the kernel bandwidth, acquired by normalizing the new bandwidth parameter $\omega'{}_d$. Similarly, the GIP kernel of the similarities$MGS\left(m_i, m_j\right)$ between metabolites m_i and m_j is defined as follows:

$$MGS(m_i, m_j) = exp(-\omega_d ||V(m_i) - V(m_j)||^2) \tag{8}$$

$$\omega_m = \omega'_m/(\frac{1}{nm}\sum_{i=1}^{nm} ||V(m_i)||^2) \tag{9}$$

where ω_m is a parameter that controls the kernel bandwidth, acquired by normalizing the new bandwidth parameter $\omega\,'_m$.

Integrated Similarity for Metabolites and Diseases

In order to ensure that similarity information exists for every pair in metabolites or diseases, we integrated the disease functional similarities with GIP kernel similarities, which is shown as follows:

$$IDS(d_i, d_j) = \begin{cases} DNS\left(d_i, d_j\right) & if\ DNS\left(d_i, d_j\right) \neq 0 \\ \quad DGS(d_i, d_j) & otherwise \end{cases} \tag{10}$$

where $IDS(d_i, d_j)$ represents the integrated disease similarities. Similarly, the integrated metabolite similarity matrix (IMS) is given as follows:

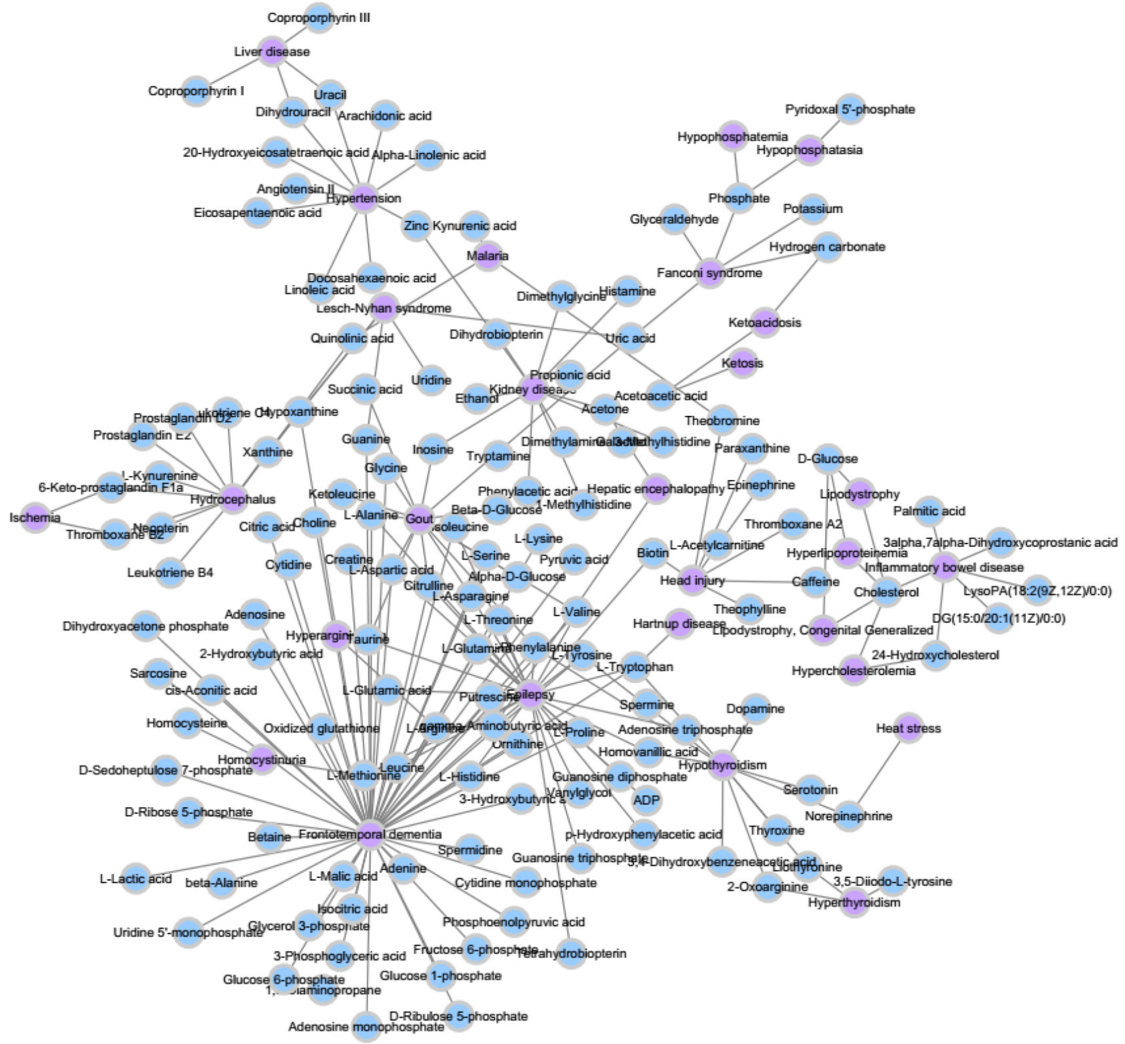


FIGURE 2 | A part of known metabolite–disease association network.

$$IMS(m_i, m_j) = \begin{cases} FHS\left(m_i, m_j\right) \text{ } if \text{ } FHS\left(m_i, m_j\right) \neq 0 \\ MGS(m_i, m_j) \text{ } otherwise \end{cases} \tag{11}$$

Feature Extraction

Firstly, type 1 features (*F1*), which consist of the values of the sum, mean, and histogram distributions of metabolite/disease similarities, are calculated using the statistical measures for each disease/metabolite. We start by calculating the number of known associations in the relevant *i*th row/*j*th column of *M*. Then, the average of all similarity scores is computed according to the *i*th/*j*th row of *IDS*/*IMS*. Simultaneously, the similarity scores that ranges at [0, 1] are split into *n* parts (*n* = 5 in this work), and the proportion of similarity scores for *d(j)*/*m(i)* that fell into each part are counted as the histogram feature.

Secondly, type 2 features (*F2*) are calculated, which include the information about graph theory-related statistics. Before obtaining this type of features, we construct the unweighted graph, in which two nodes have an edge if their similarity score is beyond the mean value of all entities in *IDS*/*IMS*. Then, we extract the relevant neighbors' information, betweenness, closeness, eigenvector centrality, and PageRank (Franceschet, 2010) scores of the disease/metabolite similarity network in an unweighted graph.

Thirdly, type 3 features (*F3*) are calculated. These features consist of the information about metabolite–disease pairs based on matrix factorization of *M*. The nonnegative matrix factorization (NMF) (Lee and Seung, 1999; Akbar et al., 2020), which was proposed by Lee and Seung, 1999, can help to solve the matrix sparsity problem. Thus, the metabolite–disease association matrix *M* can be factorized into two low-rank feature matrices A ∈ R^{nm*k} and B ∈ R^{k*nd}, where *k* denotes the dimension of the metabolite and disease features in the low-rank spaces (*k* = 20).

ALGORITHM 1 | Greedy bundling.

```
Input: $F_t$: features, Max_c:: max conflict count
Construct graph G
searchOrder← G.sortByDegree()
bundles ←{}, bundlesConflict ←{}
for ι in searchOrder do
      needNew ← True
      for j=1 to len(bundles) do
      cnt ← ConflictCnt(bundles[j],$F_t$[i])
         if cnt + bundlesConflict[i] ≤Max_c then
            bundles[j].add($F_t$[i]), needNew ← False
            break
      if needNew then
         Add $F_t$[i] as a new bundle to βυνδλεσ
Output: bundles
```

ALGORITHM 2 | Merge exclusive features.

```
Input: nD: number of data
Input: F: One bundle of exclusive features
binRanges ← {0}, totalBin ← 0
for f in F do
      totalBin +=f.numBin
      binRanges.append(totalBin)
newBin ← new Bin(numData)
for ι=1 to nD do
      newBin[i] ← 0
      for i=1 to len(F) do
         if Φ[j].bin[i] 0 then
            newBin[i] ← F[j].bin[i] + binRanges[j]
Output: newBin, binRanges
```

Finally, the feature sets $F(i,j) = [F1, F2, F3]$ for disease i and metabolite j is obtained. Meanwhile, PCA is applied to extract the more useful features.

LIGHT GRADIENT BOOSTING MACHINE

Some boosting algorithms, such as the Gradient Boosting Decision Tree (GBDT) and eXtreme Gradient Boosting (XGBoost), have a common weakness that all the sample points for every feature are scanned when obtaining the best segmentation point; this is very time-consuming and computationally expensive to meet current needs. In order to reduce the cost of the experiment, we use LightGBM as the classifier (Friedman, 2001; Ke et al., 2017). LightGBM includes two main algorithms: Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

In the GOSS algorithm, the training instances are firstly ranked according to the absolute values of their gradients in descending order. Then, the top-a × 100% instances with the larger gradients are kept and combined into an instance subset A. Besides, the (1 − a) × 100% instances with the smaller gradients are integrated in the remaining set A^c, and a further subset B with the size b × $|A^C|$ is randomly sampled. Finally, the instances are split according to the estimated variance gain $V_j'(d)$ over the subset A $\bigcup$ B. The variance gain of splitting feature j at point d is shown as follows (Ke et al., 2017)

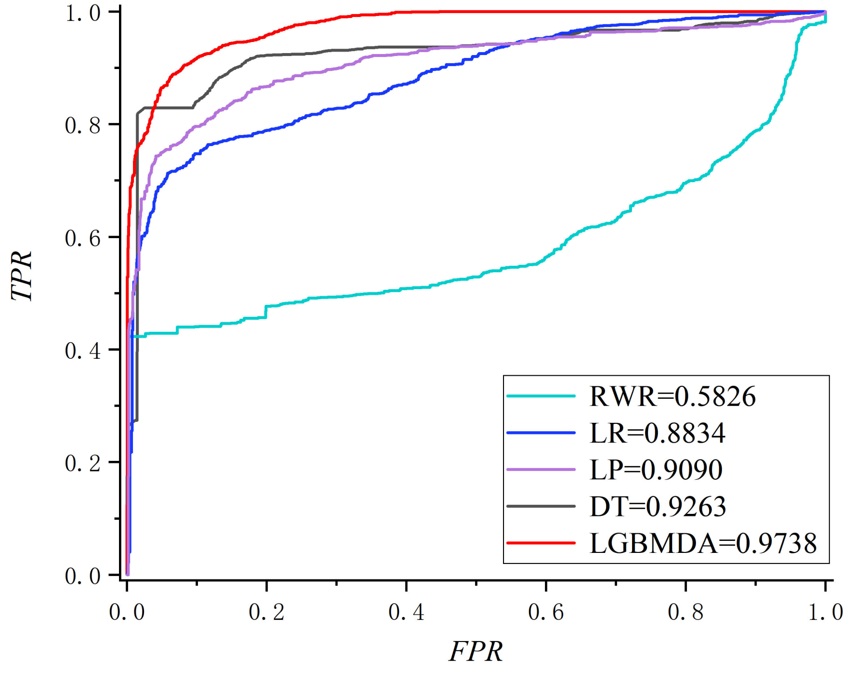


FIGURE 3 | The ROC about LOOCV.

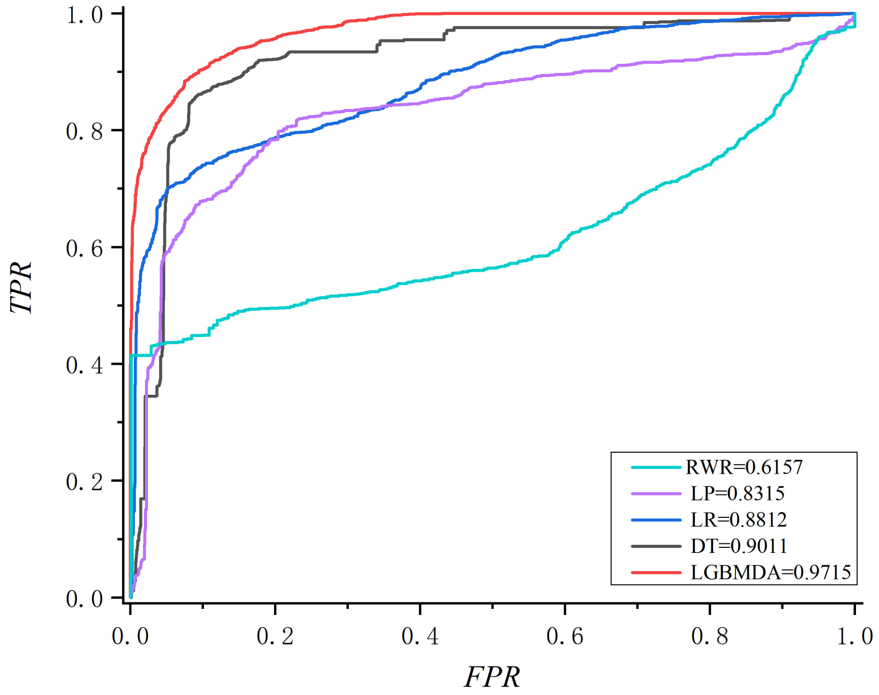


FIGURE 4 | The ROC about fivefold cross validation.

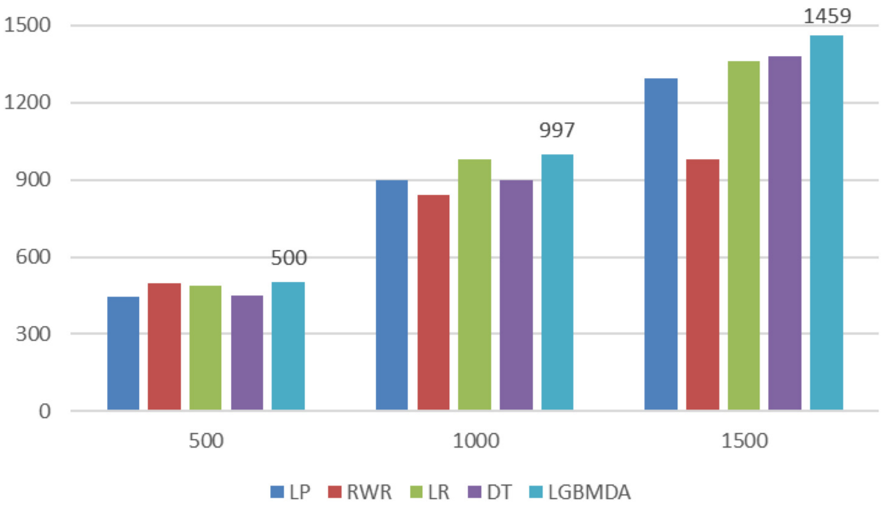


FIGURE 5 | Comparison of the top *k* ranks with different methods.

$$V_j'(d) = \frac{1}{n}(\frac{(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in A_l} g_i)^2}{n_l{}^j(d)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i)^2}{n_r{}^j(d)}) \qquad (12)$$

where$A_l = \{x_i \in \mathrm{A} : x_{ij} \le \mathrm{d}\}$, $A_r = \{x_i \in \mathrm{A} : x_{ij} > \mathrm{d}\}$, $B_l = \{x_i \in \mathrm{B} : x_{ij} \le \mathrm{d}\}$, $B_r = \{x_i \in B: x_{ij} > d\}$, and $\frac{1-a}{b}$ is used to normalize the sum of the gradients over B back to the size of A^c. Each x_i is a vector with the dimension s in space X^S. In every gradient boosting iteration, the negative gradients of the loss function with respect to the output of the model are defined as $\{g_1, \cdots, g_n\}$, where n is the number of vectors in space X^S.

In the EFB algorithm, unnecessary computation for zero feature values is avoided by binding mutually exclusive features together in a histogram to form a feature. There are two main ideas for EFB. In algorithm 1, the function is to consider which features should be bundled together, while

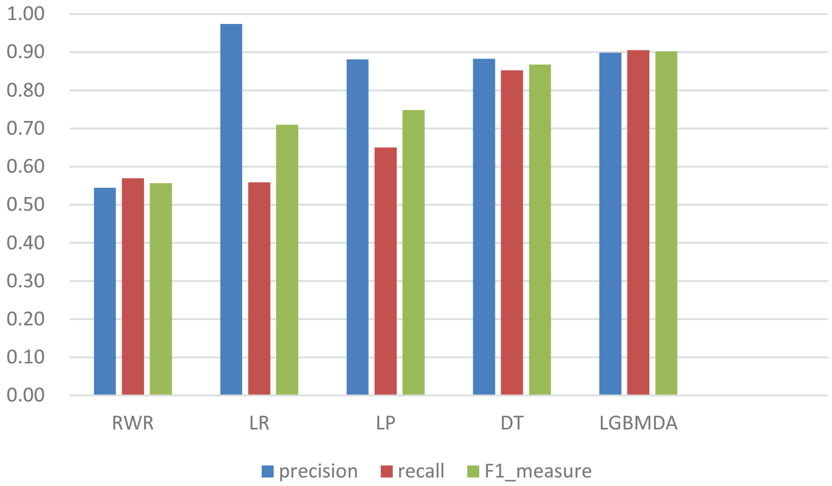


FIGURE 6 | Comparison of the precision, recall, and F1_measure with different methods.

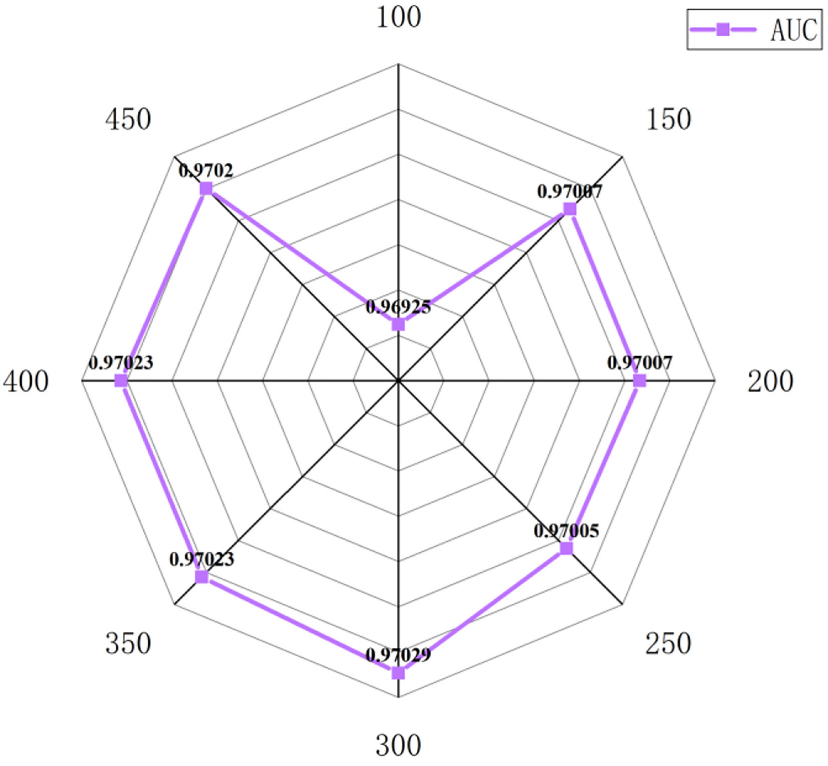


FIGURE 7 | The AUC value of different *n_estimators*.

algorithm 2 determines how to construct the bundle as follows (Ke et al., 2017):

RESULTS

In this section, we utilize LOOCV and fivefold cross-validation to evaluate the performance of LGBMMDA. In LOOCV, each confirmed metabolite–disease pair is treated as the test set in turn, while the other confirmed pairs are regarded as training sets. Besides, the unconfirmed associations are regarded as potential candidates for true associations. We plot the ROCs curves and use the area under the ROC curve (AUC) as the evaluating indicator. Furthermore, we also use fivefold cross-validation as an evaluation tool to verify the performance of our method. In this method, the known information about

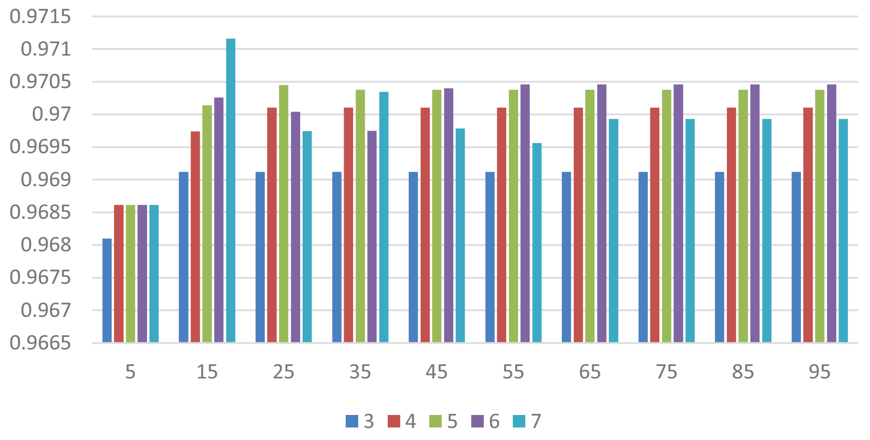


FIGURE 8 | The AUC value of different *max_depth* and *num_leaves*. Different color represents different values of *max_depth.* The *X* axis represents the different values of *num_leaves*, and the *Y* axis represents relevant AUCs.

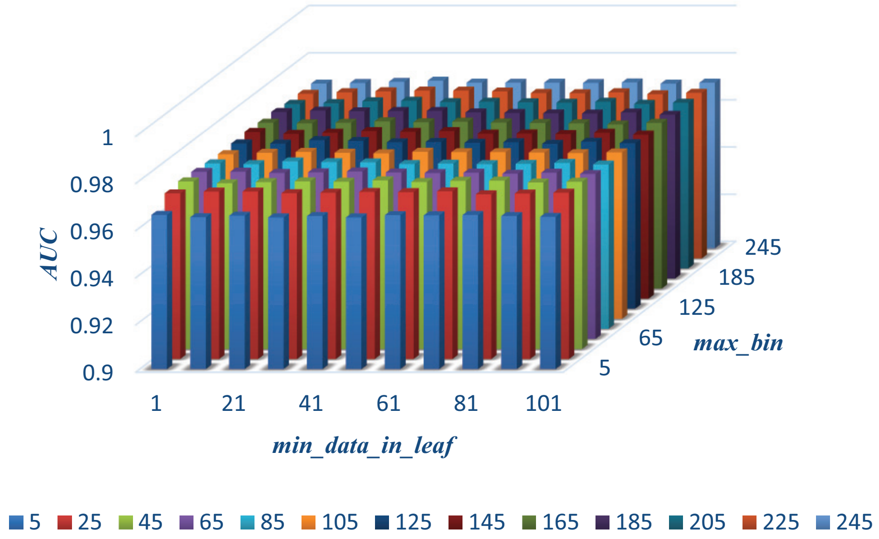


FIGURE 9 | The AUC values of different max_bin and min_data_in_leaf.

metabolites and diseases is randomly divided into five equal parts. Then, each part is used as the test set in turn, while the other four parts represent the training set. This helps to avoid having the test and training data overlapping with each other and ensures unbiased comparisons. In this study, we compare our method with some state-of-the-art methods, including the label propagation algorithm (LP), which is a semi-supervised learning method based on graph (and its basic idea is to predict the label information of unlabeled nodes by using the label information of labeled nodes); random walk (RWR), which is close to Brownian motion and is the ideal mathematical state of Brownian motion; logistic regression (LR), which is a machine learning method solving binary (0 or 1) problems and estimating the possibility of something; and decision tree (DT), which is the process of classifying data through a series of rules. The results show that LGBMMDA achieved AUC values of 0.9738 and 0.9715 in LOOCV and fivefold cross-validation, respectively (see **Figures 3**, **4**). In addition, we analyze the scores of known associations about LOOCV and count the number of known associations correctly identified by each algorithm (see **Figure 5**). It can be seen from **Figure 6** that our proposed method is superior to other methods in terms of precision, recall, and F1-measure (0.898596, 0.90566, and 0.9021, respectively). Although the precision of LR is higher than our method, the recall of LR is significantly lower. Our method is steadier than LR.

PARAMETER ANALYSIS

In this section, we select some significant parameters to be adjusted in LightGBM. Firstly, we set the parameter *n_estimators*, which is related to the number of residual trees, from 100 to 500,

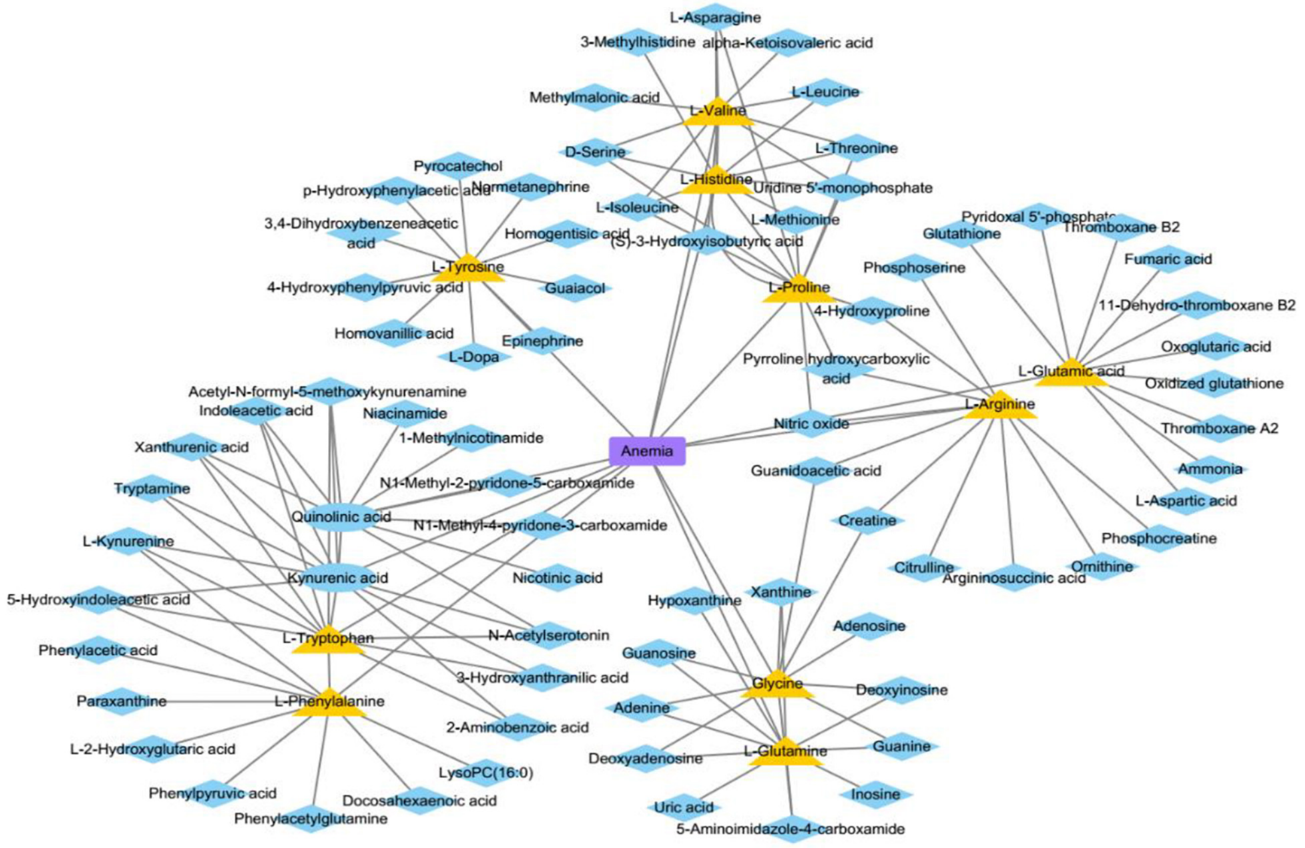


FIGURE 10 | The associations between anemia and some metabolites. The blue ellipses represent the known metabolites about anemia in this study. The yellow triangles represent the top 10 predicted metabolites relevant to anemia. The blue diamonds represent the top 10 neighbors about predicted or known metabolites.

while other important parameters are set to default. **Figure 1** shows that we get better results when *n_estimators* is set to 300 (see **Figure 7**). In order to improve the accuracy, the values of the parameter *max_depth*, which limits the maximum depth of the tree model, is set from 3 to 8, and *num_leaves*, which controls the number of leaf nodes, is set from 5 to 100. As a result, *max_depth* = 7 and *num_leaves* = 15 achieve better performance (see **Figure 8**). Finally, the range of *max_bin*, which has an effect on overfitting, is set from 5 to 256, and *min_data_in_leaf*, which is the minimum number of samples contained on a leaf node, is set from 1 to 100. The results show that *max_bin* = 45 and *min_data_in_leaf* = 51 are better than other values (see **Figure 9**).

CASE STUDY

In this section, we analyze three kinds of diseases, anemia, uremia, and asthma, in case studies to discover their pathogenic mechanisms from the perspective of metabolites. There are 10, 9, and 7 metabolites of these diseases that could be verified out of the top 10 predicted metabolites, respectively. **Figure 10** shows anemia and its relevant metabolites.

Anemia is caused by the inability of the body to produce enough hemoglobin, which is a protein that carries oxygen to blood cells and tissues. This disease has common symptoms, such as fatigue and dizziness. We conduct our method on a case study of anemia (see **Table 1**) to select the top 10 most likely associated metabolites, and all of them are associated with anemia according to literature in NCBI. For instance, L-histidine (Peterson et al., 1998) acts as a semi-essential amino acid, which is medically used in the treatment of anemia (Wang et al., 2020).

Table 1 | Candidate metabolites of anemia.

	Anemia	
Rank	**Metabolite name**	**Evidences**
1	L-Histidine	PMID: 32498848
2	L-Proline	PMID: 26821380
3	Glycine	PMID: 30853991
4	L-Arginine	PMID: 31355573
5	L-Valine	PMID: 30860750
6	L-Tryptophan	PMID: 32153576
7	L-Glutamine	PMID: 32350885
8	L-Tyrosine	PMID: 32764239
9	L-Glutamic acid	PMID: 30628549
10	L-Phenylalanine	PMID: 26956768

Table 2 | Candidate metabolites of asthma.

Asthma		
Rank	**Metabolite name**	**Evidences**
1	L-Histidine	PMID: 31206804
2	L-Proline	PMID: 29059088
3	L-Tryptophan	PMID: 31951781
4	L-Glutamic acid	–
5	3-Hydroxybutyric acid	PMID: 32213896
6	Succinic acid	PMID: 14846625
7	L-Methionine	PMID: 32778730
8	1-Methylhistidine	PMID: 24783928
9	L-Threonine	–
10	PC(18:1(11Z)/22:1(13Z))	–

Table 3 | Candidate metabolites of uremia.

Uremia		
Rank	**Metabolite name**	**Evidences**
1	L-Histidine	PMID: 8676800
2	L-Proline	PMID: 20355181
3	3-Hydroxybutyric acid	
4	Biotin	PMID: 6322032
5	Xanthine	PMID: 19379356
6	L-Tryptophan	PMID: 935125
7	Inosine	PMID: 9607216
8	Succinic acid	PMID: 13837895
9	L-Glutamic acid	PMID: 6508956
10	gamma-Aminobutyric acid	PMID: 16797388

Asthma is a common and frequent disease, which has the main symptoms of paroxysmal wheezing, chest tightness, and cough. The field of metabolomics has been used to explore the metabolic signatures of asthma, both for biomarker identification and pathophysiologic mechanisms research. We perform our method on a case study of asthma, and 7 of the top 10 predicted metabolites that are interrelated with asthma are verified to be correlative (see **Table 2**). For example, L-proline (Nadler et al., 1988) is one of metabolic characteristics of asthma, which is supported by experimental asthma models and clinical studies in children and adults (Pite et al., 2018). Another example is L-tryptophan (Hartzema et al., 1991), which has long been suggested to be relevant to the pathophysiology of asthma (Hu et al., 2020).

Uremia is a serious kidney disease that is caused by a disorder in the internal biochemical process after renal function loss. We conduct our calculation method on a case study of uremia. As illustrated in **Table 3**, 9 of the top 10 predicted metabolites that are interrelated with uremia are verified to be correlative. For example, L-histidine is found to be significantly enhanced in the brain in uremia patients (Schmid et al., 1996). The L-proline in body fluids is a biological parameter for patients with renal insufficiency and chronic uremia (Hanwen, Sun et al., 2009).

DISCUSSION

Uncovering complex disease-related metabolites is a vital research topic in metabolomics. To this end, we proposed a computational model called LGBMMDA under the framework of LightGBM. The experimental results by cross-validation have proven that our method outperforms previously used methods. Furthermore, three case studies indicate that the metabolite–disease correlations predicted in our method can be effectively demonstrated by relevant experiments. The LGBMMDA method is expected to be a useful biomedical research tool for predicting potential metabolite–disease associations.

There are three factors that contribute to the ideal predictive performance of LGBMMDA. Our method makes the following contributions for uncovering metabolite–disease associations: Firstly, the data of the metabolite–pathway associations are selected as metabolite functional similarities, which is a novel way to calculate similarities between metabolites. Secondly, three features are extracted by different angles, which keeps the diversity of features and contributes to a reliable performance. Thirdly, our method utilizes the reliable classifier of LightGBM, which ensures an effectively predictive accuracy.

However, there are several limitations in our prediction model. On the one hand, many parameters of GBM need to be adjusted. In this work, parameter adjustment is only carried out by some experiments. In future work, some algorithms might be used to adjust those parameters. On the other hand, more useful methods for calculating relevant similarities could be beneficial to enhancing the performance of our model. In the future, more biologically relevant information is expected to be available, which can be used to refine the similarities.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data about metabolites can be found here: https://hmdb.ca/.

AUTHOR CONTRIBUTIONS

CZ carried out the method IBNPLNSMDA to predict the potential associations of metabolites and diseases, participated in its design, and drafted the manuscript. XL and LL helped to draft the manuscript. All authors read and approved the final manuscript.

FUNDING

We are grateful for the financial support comes from the National Natural Science Foundation of China (61672334, 61972451, and 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

ACKNOWLEDGMENTS

We thank EditSprings (https://www.editsprings.com/) for the expert linguistic services provided.

REFERENCES

Akbar, J. A., Kusalik, A., and Wu, F. X. (2020). MDIPA: A microRNA-drug interaction prediction approach based on nonnegative matrix factorization. *Bioinformatics* 36, 5061–5067. doi: 10.1093/bioinformatics/btaa577

Boja, E. S., Fehniger, T. E., Baker, M. S., Marko-Varga, G., and Rodriguez, H. (2014). "Analytical validation considerations of multiplex mass-spectrometry-based proteomic platforms for measuring protein biomarkers. *J. Proteome Res.* 13, 5325–5332. doi: 10.1021/pr500753r

Charikar, M. (2002). "Similarity estimation techniques from rounding algorithms," in *Proceedings on 34th Annual ACM Symposium on Theory of Computing*, (New York, NY), 380–388.

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2017). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief. Bioinform.* 20, 203–209. doi: 10.1093/bib/bbx103

Deutsch, H. P. (2004). *Principle Component Analysis*. London: Palgrave Macmillan.

Dunn, W. B., and Ellis, D. I. (2005). Metabolomics: current analytical platforms and methodologies. *Trends Anal. Chem.* 24, 285–294. doi: 10.1016/j.trac.2004.11.021

Franceschet, M. (2010). PageRank: Standing on the shoulders of giants. *arXiv[preprint]* arXiv:1002.2858,

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 1189–1232.

Gu, C., Bo, L., Xiaoying, L., Lijun, C., Haowen, C., Keqin, L., et al. (2017). "Network-based collaborative filtering recommendation model for inferring novel disease-related miRNAs. *RSC Advances* 7, 44961–44971. doi: 10.1039/c7ra09229f

Hartzema, A. G., Porta, M. S., Tilson, H. H., Milburn, D. S., and Myers, C. W. (1991). Tryptophan toxicity: a pharmacoepidemiologic review of eosinophilia-myalgia syndrome. 25, 1259–1262. doi: 10.1177/106002809102501116

Hu, Q., Jin, L., Zeng, J., Wang, J., Zhong, S., Fan, W., et al. (2020). Tryptophan metabolite-regulated Treg responses contribute to attenuation of airway inflammation during specific immunotherapy in a mouse asthma model. *Hum. Vaccin. Immunother.* 16, 1891–1899. doi: 10.1080/21645515.2019.1698900

Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., and Cheng, L. (2018). Identifying diseases-related metabolites using random walk. *BMC Bioinformatics* 19(Suppl 5):116.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (Long Beach, CA), 3149–3157.

Lee, D. D., and Seung, H. S. J. N. (1999). "Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2016). An analysis of human microbe–disease associations. *Brief. Bioinform.* 18, 85–97.

Nadler, J. V., Wang, A., and Hakim, A. (1988). "Toxicity of L-proline toward rat hippocampal neurons. *Brain Res.* 456, 168–172. doi: 10.1016/0006-8993(88)90358-7

Peterson, J. W., Boldogh, I., Popov, V. L., Saini, S. S., and Chopra, A. K. (1998). "Anti-inflammatory and antisecretory potential of histidine in *Salmonella*-challenged mouse small intestine. *Lab. Invest.* 78, 523–534.

Pite, H., Morais-Almeida, M., and Rocha, S. M. (2018). "Metabolomics in asthma: where do we stand? *Curr. Opin. Pulm. Med.* 24, 94–103. doi: 10.1097/mcp.0000000000000437

Schmid, G., Bahner, U., Peschkes, J., and Heidland, A. (1996). "Neurotransmitter and monoaminergic amino acid precursor levels in rat brain: effects of chronic renal failure and of malnutrition. *Miner. Electrolyte Metab.* 22, 115–118.

Sun, H., Li, L., and Wu, Y. (2009). Capillary electrophoresis with electrochemiluminescence detection for simultaneous determination of proline and fleroxacin in human urine. *Drug Test. Anal.* 1, 87–92. doi: 10.1002/dta.22

Tang, X., Lin, C. C., Spasojevic, I., Iversen, E. S., Chi, J. T., Marks, J. R., et al. (2014). A joint analysis of metabolomics and genetics of breast cancer. *Breast Cancer Res.* 16, 415.

Wang, X. Z., Zhang, Z. Q., Guo, R., Zhang, Y. Y., Zhu, N. J., Wang, K., et al. (2020). Dual-emission CdTe quantum dot@ZIF-365 ratiometric fluorescent sensor and application for highly sensitive detection of l-histidine and Cu2. *Talanta* 217, 121010. doi: 10.1016/j.talanta.2020.121010

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., et al. (2017). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* 46, D608–D661.

Xianlin, H., Rozen, S., Boyle, S. H., Hellegers, C., Cheng, H., Burke, J. R., et al. (2011). "Metabolomics in early Alzheimer's disease: identification of altered plasma sphingolipidome using shotgun lipidomics. *PLoS One* 6:e21643. doi: 10.1371/journal.pone.0021643

Zhang, Y., Chen, M., Cheng, X., and Wei, H. (2020). MSFSP: a novel mirna–disease association prediction model by federating multiple-similarities fusion and space projection. *Front. Genet.* 11:389.

Zhou, X., Menche, J., Barabási, A. L., and Sharma, A. (2014). Human symptoms–disease network. *Nat. Commun.* 5, 4212.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 May 2021
doi: 10.3389/fgene.2021.666575

Predicting Drug-Disease Association Based on Ensemble Strategy

Jianlin Wang[1], Wenxiu Wang[1], Chaokun Yan[1], Junwei Luo[2]* and Ge Zhang[1]*

[1] School of Computer and Information Engineering, Henan University, Kaifeng, China, [2] College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

Drug repositioning is used to find new uses for existing drugs, effectively shortening the drug research and development cycle and reducing costs and risks. A new model of drug repositioning based on ensemble learning is proposed. This work develops a novel computational drug repositioning approach called CMAF to discover potential drug-disease associations. First, for new drugs and diseases or unknown drug-disease pairs, based on their known neighbor information, an association probability can be obtained by implementing the weighted K nearest known neighbors (WKNKN) method and improving the drug-disease association information. Then, a new drug similarity network and new disease similarity network can be constructed. Three prediction models are applied and ensembled to enable the final association of drug-disease pairs based on improved drug-disease association information and the constructed similarity network. The experimental results demonstrate that the developed approach outperforms recent state-of-the-art prediction models. Case studies further confirm the predictive ability of the proposed method. Our proposed method can effectively improve the prediction results.

Keywords: drug repositioning, ensemble strategy, similarity measure, matrix completion, drug-disease association

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Bolin Chen,
Northwestern Polytechnical University, China
Wei Peng,
Kunming University of Science and Technology, China

****Correspondence:***
Chaokun Yan
ckyan@henu.edu.cn
Junwei Luo
luojunwei@hpu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *10 February 2021*
Accepted: *23 March 2021*
Published: *03 May 2021*

Citation:
Wang J, Wang W, Yan C, Luo J and Zhang G (2021) Predicting Drug-Disease Association Based on Ensemble Strategy. Front. Genet. 12:666575. doi: 10.3389/fgene.2021.666575

1. INTRODUCTION

Traditional drug discovery is a high-risk, high-investment, and long-term process (Li et al., 2015). It is well-known that it usually takes more than 10 years and more than $800 million to bring a new drug to market (Adams and Brantner, 2006). Additionally, the probability of drug approval success is below 10% (Ashburn and Thor, 2004). Considering the challenges of traditional drug discovery, the drug repositioning method is rising in popularity (Cano et al., 2017) and has attracted increasing interest from the research community and pharmaceutical industry (Shameer et al., 2015). Some successful repositioning drugs, such as duloxetine, sildenafil, and thalidomide, have generated high revenues in the history of their patent holders or companies (Ashburn and Thor, 2004).

The purpose of drug repositioning is to discover new indications for old drugs. Recently, many computational drug repositioning techniques, such as machine learning-based models, have been used to identify potential drug-disease interactions (Li et al., 2015). For example, Napolitano et al. (2013) melded drug-related features into a single information layer, which was used to train a multi-class support vector machine classifier whose output was a therapeutic class for a given drug. Chen and Li (2017) proposed the flexible and robust multiple-source learning (FRMSL) method to integrate multiple heterogeneous data sources to obtain drug-drug similarity and disease-disease similarity, and used the Kronecker regularized least squares (KronRLS) approach to solve the prediction problem. Liang et al. (2017) used Laplacian regularized sparse subspace learning to find novel drug indications, integrating multiple pieces of information. Most machine learning-based models using negative samples are generated randomly from unknown associations, among which some false negatives may be included, resulting in a biased decision boundary (Liu et al., 2016a).

In recent years, with the rapid advance of high-throughput biology, huge amounts of multi-omic data have been yielded and several databases have been developed to store these valuable data (Chen et al., 2019; Luo et al., 2020). With the development of publicly available drug-related or disease-related databases, the network-based method is widely used in drug repositioning. The network-based method discovered potential drug–disease associations by propagating information in a heterogeneous biological network containing some information about diseases, drugs, or targets (Luo et al., 2018). For example, Yu et al. (2015) used drugs, protein complexes, and diseases to construct a tripartite network, which inferred the association probabilities of drug-disease pairs. Martìnez et al. (2015) developed DrugNet, a model for drug-disease and disease-drug prioritization; a network of interconnected drugs, proteins, and diseases was built, and DrugNet was used for drug repositioning. Luo et al. (2016) utilized drug- and disease-related properties to compute comprehensive similarity measures and the utility bi-random walk (BiRW) algorithm to find new uses for existing drugs. In recent years, the matrix factorization-based method has been successfully applied to biological association prediction, such as lncRNA-disease (Fu et al., 2017; Lan et al., 2020), drug-target (Liu et al., 2016b; Shi et al., 2018), and drug-disease (Zhang et al., 2018). The method can integrate prior information flexibly and integrate much information and many features into the framework to improve the accuracy of prediction. Zhang et al. (2018) developed a similarity-constrained matrix factorization approach (SCMFDD), which utilizes known drug-disease interactions, drug features, and disease features to predict potential drug-disease associations. Gönen and Kaski (2014) developed a new probabilistic method KBMF2MKL, which extended kernelized matrix factorization by incorporating multiple kernel learning. However, association prediction with matrix factorization has some limitations on the accuracy and prediction performance, especially for new diseases or drugs, which are called cold start problems. So, given different prediction approaches, an ensemble method is a promising way to combine their capacity in predicting the associations between drugs and diseases.

In this work, we develop a new drug repositioning model, CMAF, which integrates three methods (matrix factorization-based, label propagation-based, and network consistency projection-based methods) to obtain the final prediction result. To assess the performance of the developed approach, 10-fold cross-validation was implemented, and from the experimental results, we can see that ensemble models can combine different information to achieve high-accuracy performance. The experimental results demonstrate that CMAF obtained better results than the other four recent models in predicting potential drug-disease associations.

2. MATERIALS AND METHODS

In this section, we first introduce the gold standard dataset used in this study. Then, a proposed drug repositioning method named CMAF is presented to discover new uses for existing drugs. The overall flowchart of CMAF is shown in **Figure 1**, which contains the following three steps. First, the WKNKN algorithm is used as a preconditioning step to compute the temporary association score for new drugs and diseases or unknown drug-disease pairs. Second, a new drug-drug similarity network and a new disease-disease similarity network can be established. Third, three classical models are used to predict potential drug-disease associations separately, and their prediction results are ensembled to obtain the final association possibility of drug-disease pairs.

2.1. Dataset

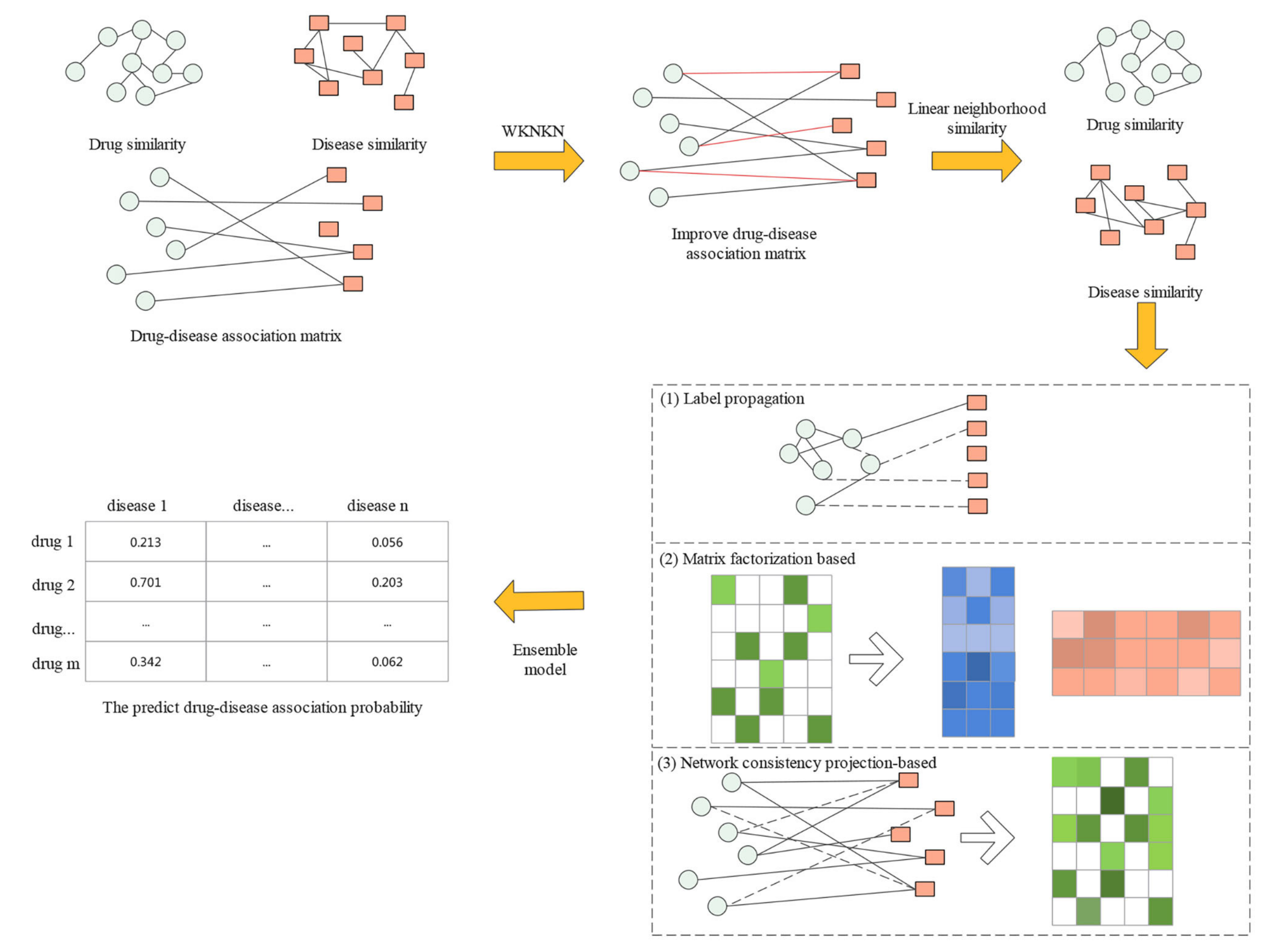
The dataset used in this paper is curated manually from multiple biological datasets (Gottlieb et al., 2011). The dataset has 593 drugs and 313 diseases involving 1,933 validated drug-disease pairs. The drugs are collected from DrugBank (Wishart et al., 2006), and the diseases are extracted from Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2002).

The drug similarity is computed by the Chemical Development Kit (CDK) (Steinbeck et al., 2006) in terms of SMILES (Weininger, 1988) chemical structures, and the similarity between drug pairs is denoted as the Tanimoto score (Tanimoto, 1958) of their 2D chemical fingerprints. The disease similarity is computed using MimMiner (van Driel et al., 2006), which measures the similarity of two diseases by calculating the similarity between the MeSH terms (Lipscomb, 2000) present in the medical description information from the OMIM database.

2.2. Improved Drug-disease Association

A known drug-disease association Y can be modeled as a two-dimensional matrix, which has m drug rows and n disease columns, where each entry is denoted by Y_{ij}. The i-th row vector of the adjacency matrix Y, $Y(r_i) = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$, is the interaction profile for drug r_i. Similarly, the j-th column vector of the adjacency matrix Y, $Y(d_j) = (Y_{1j}, Y_{2j}, \ldots, Ymj)$, is the interaction profile for disease d_j.

It should be noted that the interaction profiles of new drugs or new diseases are all zero values. Additionally, many of the non-associations in Y are unobserved situations that could have potential interactions (i.e., false negatives). Therefore, we used WKNKN (Ezzat et al., 2017) to obtain the interaction likelihood value for non-associated drug-disease pairs in terms of their K nearest known neighbors [the K nearest known neighbors can be obtained by the K nearest neighbors (KNN) function according to their drug or disease similarity]. Here, we set K = 5. For every drug r_i, the similarity of its chemical structure with the K known drugs nearest to it and their corresponding values in the interaction profiles are utilized to obtain the interaction


Drug similarity
Disease similarity
Drug-disease association matrix
WKNKN
Improve drug-disease association matrix
Linear neighborhood similarity
Drug similarity
Disease similarity
(1) Label propagation
(2) Matrix factorization based
(3) Network consistency projection-based
Ensemble model
The predict drug-disease association probability

	disease 1	disease...	disease n
drug 1	0.213	...	0.056
drug 2	0.701	...	0.203
drug...	...	...	...
drug m	0.342	...	0.062

FIGURE 1 | Flowchart of CMAF.

likelihood profile of the drug r_i as follows:

$$Y_r(p) = \left(\sum_{i=1}^{K} w_i Y\left(r_i\right) \right) / Q_r \tag{1}$$

where r_i to r_k represent the K known nearest neighbors of drug r_p; the weight coefficient is $w_i = T^{i-1} S^r(r_i, r_p)$ where $T \leq 1$ is the decay term, and here, we set T to 0.5; and $S^r(r_i, r_p)$ is the similarity between r_i and r_p. Moreover, $Q_r = \sum_{i=1}^{K} S^r\left(r_i, r_p\right)$ is the normalization term. For the same reason, the interaction likelihood profile of disease d_j is as follows:

$$Y_d(q) = \left(\sum_{j=1}^{K} w_j Y\left(d_j\right) \right) / Q_d \tag{2}$$

where d_1 to d_k represent the K known nearest neighbors of disease d_q, the weight coefficient is $w_j = T^{j-1} S^d(d_j, d_q)$, the decay term T is 0.5, $S^d(d_j, d_q)$ is the similarity between d_j and d_q, and the normalization term is $Q_d = \sum_{j=1}^{K} S^d\left(d_j, d_q\right)$.

Then, we fuse Y_r and Y_d to replace $Y_{ij} = 0$ by taking the average of the two values mentioned above and denote it as Y_{rd}; we can then obtain a new adjacency matrix Y.

$$Y = max(Y, Y_{rd}) \tag{3}$$

where, $Y_{rd} = (Y_r + Y_d)/2$.

2.3. Improved Similarity of Drugs and Diseases

Similarity-based methods are widely used to find similar drugs (Vilar and Hripcsak, 2017). Some studies have shown that the use of similarity measures in drug repositioning often shows high predictive power (Azad et al., 2020). Therefore, similarity measurement is always regarded as an important step in drug repositioning research. The improvement of similarity can improve the prediction performance (Wang and Kurgan, 2019), reduce the computation cost, and make the similarity-based method more attractive and promising (Ding et al., 2014).

Relevant studies found that each data point can be linearly reconstructed from its neighborhood (Wang and Zhang, 2008), we can calculate the pairwise drug similarity and pairwise disease similarity, which is the same method as in previous works (Zhang et al., 2017).

Here, we use drug data points as an example. Let x_i represent the feature vector of the i-th drug. The optimization problem is expressed as:

where $N(x_i)$ denotes the set of $K(0 < K < n)$ nearest neighbors. Here, we set K to 100.

$$\begin{aligned} \min_{\omega_i} \varepsilon_i = & \left\| x_i - \textstyle\sum_{i_j : x_i \in N(x_i)} \omega_{i,i_j} x_{i_j} \right\|^2 \\ & = \textstyle\sum_{i_j, i_k : x_{i_j}, x_{i_k} \in N(x_i)} \omega_{i,i_j} G^i_{i_j,i_k} \omega_{i,i_k} = \omega_i^T G^i \omega_i \\ \text{s.t. } & \textstyle\sum_{i_j : x_{i_j} \in N(x_i)} \omega_{i,i_j} = 1, \omega_{i,i_j} \geq 0, j = 1, 2, \ldots, K \end{aligned} \tag{4}$$

$G^i_{i_j,i_l} = \left(x_i - x_{i_j}\right)^T \left(x_i - x_{i_l}\right)$. ω_{i,i_j} are the weights x_{i_j} for rebuilding x_i and can be seen as the similarity of x_i and x_{i_j}.

To avoid over-fitting, we add the regularization term for the rebuilt weight w_i and the objective function can be transformed as follows:

$$\begin{aligned} & \min_{\omega_i} \varepsilon_i = \omega_i^T G^i \omega_i + \lambda \left\| \omega_i \right\|^2 = \omega_i^T \left(G^i + \lambda I\right) \omega_i \\ & \text{s.t. } \textstyle\sum_{i_j : x_{i_j} \in N(x_i)} \omega_{i,i_j} = 1, \omega_{i,i_j} \geq 0, j = 1, 2, \ldots, K \end{aligned} \tag{5}$$

where λ denotes the regularization parameter. Here, we set $\lambda = 1$.

We adopt standard quadratic programming to solve Equation (5), and its solution is called the *linear neighborhood similarity*. Here, a weight matrix W can be obtained, which we regard as the drug linear neighborhood similarity S^{r^*}.

Likewise, we can obtain the disease linear neighborhood similarity S^{d^*}.

2.4. Prediction Method

In this section, we use the drug linear neighborhood similarity and disease linear neighborhood similarity S^{d^*} to carry out three classical approaches to predict unobserved drug-disease interactions separately and ensemble their prediction results to obtain the final association possibility of drug-disease pairs.

2.4.1. Label Propagation

Label propagation (LP) methods perform the following task: given a weighted network, in which a small part of the nodes are labeled (with labels, such as positive), calculate the labels of the remaining unlabeled nodes (Zhang et al., 2015).

We formulate S^{d^*} as a directed graph, where drugs are nodes and the edge between drug r_i and drug r_j is weighted by the linear neighborhood similarity between the two drugs.

After constructing the graph, we utilize a label propagation approach to predict the unknown drug-disease pair association score (LPRIA). The known drug-disease associations are considered the initial node label information, and then the label information is updated. In each step, each drug node absorbs its neighbor's label information with probability α and maintains the initial state with probability $1 - \alpha$. Here, we set α as 0.5. The updated process can be written as:

$$Y_j^{t+1} = \alpha S^{r^*} Y_j^t + (1 - \alpha) Y_j^0 \tag{6}$$

where, Y_j^0denotes the j-th column of the initial drug-disease interaction matrix Y (i.e., the initial states of all drugs for disease d_j). Furthermore, taking all diseases into account, the update process can be formulated in matrix form as:

$$Y^{t+1} = \alpha S^{r^*} Y^t + (1 - \alpha) Y^0 \tag{7}$$

Equation (7) will be used to update the label matrix until it converges, and Equation (7) will converge to:

$$Y^{r^*} = (1 - \alpha) \left(I - \alpha S^{r^*}\right)^{-1} Y^0 \tag{8}$$

where I represents the identity matrix and Y^{r^*} represents the predicted drug-disease pair probability from the drug side. For the convergence analysis of this update process, please refer to Wang and Zhang (2008).

Likewise, we constructed the label propagation approach from the disease side to obtain the predicted drug-disease interaction score matrix Y^{d^*}. The final association score Y^* is obtained according to the average of Y^{r^*} and Y^{d^*}.

2.4.2. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is an unsupervised model (Fujita et al., 2018). Its goal is to obtain two non-negative matrices and take their product as the optimal approximation to the original matrix. From the perspective of drug repositioning, the drug-disease association matrix $Y \in R^{m \times n}$ is factorized into two non-negative matrices, $W \in R^{m \times k}$ and $H \in R^{n \times k}$ ($k \ll \min(m, n)$), here, we set k to 100, and $Y \approx WH^T$.

To avoid over-fitting and increase the learning performance, Tikhonov and graph regularization terms are added to the standard NMF model to predict novel drug-disease pairs (NMFRIA). NMFRIA's objective function is as follows:

$$\begin{aligned} \min_{W,H} & \left\| Y - WH^T \right\|_F^2 + \lambda_l \left(\|W\|_F^2 + \|H\|_F^2 \right) + \lambda_r \operatorname{Tr}\left(W^T L_r W\right) \\ & + \lambda_d \operatorname{Tr}\left(H^T L_d H\right) \\ \text{s.t.} & \quad W \geq 0, H \geq 0 \end{aligned} \tag{9}$$

where λ_l, λ_r, and λ_d represent the regularization coefficients; $Tr(\cdot)$ denotes the trace of a matrix, $L_r = D_r - S^{r^*}$ is the graph Laplacian matrix for the drug similarity matrices, S^{r^*} and $L_d = D_d - S^{d^*}$ are the graph Laplacian matrices for the disease similarity matrices S^{d^*} (Liu et al., 2014); and D_r and D_d represent the diagonal matrices whose entries are the row sums of S^{r^*} and S^{d^*}, respectively.

The method proposed by Xiao et al. (2018) is adopted to solve the minimization problem, and W and H are updated with an iterative equation. Here, the updating rules can be defined as:

$$w_{ik} \leftarrow w_{ik} \frac{\left(YH + \lambda_r S^{r^*} W\right)_{ik}}{\left(WH^T H + \lambda_l W + \lambda_r D_r W\right)_{ik}} \tag{10}$$

$$h_{jk} \leftarrow h_{jk} \frac{\left(Y^T W + \lambda_d S^{d^*} H\right)_{jk}}{\left(HW^T W + \lambda_l H + \lambda_d D_d H\right)_{jk}} \tag{11}$$

where w_{ik} represents the i-th row and the k-th column of non-negative matrix W, and h_{jk} represents the j-th row and the k-th column of non-negative matrix H.

According to Equations (10) and (11) the two non-negative matrices W and H are updated until convergence, and then we can obtain the predicted drug-disease interaction matrix as $Y^{**} = WH^T$. Here, we set λ_l to 2, and $\lambda_r = \lambda_d = 0.0001$.

2.4.3. Network Consistency Projection

Network consistency projection (NCP) considers drugs r_i that have a higher similarity to other drugs in the drug similarity matrix; the more drugs are associated with disease d_j, the higher the spatial similarity of drug r_i with disease d_j (and vice versa). Here, we use the NCP approach (Gu et al., 2016) for drug-disease association (NCPRIA) to obtain the predicted association scores between unknown drug-disease pairs.

NCPRIA computes the association probability between drug r_i and disease d_j by fusing two network consistency projection scores (the drug and disease space projection scores). Considering that unknown drug-disease pairs are not confirmed by experiment, which cannot prove that they are unrelated, and to prevent 0 from being the denominator, we replace 0 in the matrix Y with 10–30.

The drug space projection is the projection of the drug similarity network S^{r^*} on the drug-disease interaction network Y, which can be described as follows:

$$NCP\_R(i,j) = \frac{S^{r^*}(i,:)^*Y(:,j)}{|Y(:,j)|} \tag{12}$$

where $S^{r^*}(i,:)$ denotes the similarities between drug r_i and all other drugs in the i-th row of matrix S^{r^*} and $Y(:,j)$ denotes the associations between disease d_j and all drugs. $|Y(:,j)|$ represents the length of the vector $Y(:,j)$. $NCP\_R(i,j)$ represents the network consistency projection score of $S^{r^*}(i,:)$ on $Y(:,j)$. It is worth noting that the smaller the angle is between $S^{r^*}(i,:)$ and $Y(:,j)$, the more drugs are related to disease *j* and the more similar drugs there are to drug *i*, the larger the network consistency projection score $NCP\_R(i,j)$.

Similarly, we can obtain the disease space projection score as follows:

$$NCP\_D(i,j) = \frac{Y(i,:)^*S^{d^*}(:,j)}{|Y(i,:)|} \tag{13}$$

where $S^{d^*}(:,j)$ denotes the j-th column of matrix S^{d^*} and $Y(i,:)$ denotes the i-th row of drug-disease association Y. $NCP\_D(i,j)$ represents the network consistency projection score of $S^{d^*}(:,j)$ on $Y(i,:)$.

Finally, the projection score for the drug space and disease space are fused and normalized as follows:

$$Y^{***}(i,j) = \frac{NCP\_R(i,j) + NCP\_D(i,j)}{\left|S^{r^*}(i,:)\right| + \left|S^{d^*}(:,j)\right|} \tag{14}$$

where Y^{***} represents the predicted drug-disease association matrix and $Y^{***}(i,j)$ is the final predicted score of drug r_i and disease d_j.

2.4.4. Integrating the Prediction Results

According to the three aforementioned computational drug repositioning methods, to obtain better performance, a fusion model is adopted to integrate their predicted results, and the final prediction score between drugs and diseases is computed as follows:

$$Rt = 1 - \left(1 - Y^{*}\right)\left(1 - Y^{**}\right)\left(1 - Y^{***}\right) \tag{15}$$

In particular, Y^* is the predicted drug-disease association probability of the LPRIA method, Y^{**} is the predicted association probability of the NMFRIA method, Y^{***} is the predicted association probability of the NCPRIA method, and *Rt* stands for the final predicted drug-disease association probability.

3. EXPERIMENTS AND RESULTS

In this section, the performance of our approach, CMAF, is systematically evaluated. First, we describe the evaluation metrics. Based on a gold standard dataset, we compare our approach with several recent prediction algorithms and present the results in this section. In addition, the effectiveness of the developed method is further confirmed by case studies.

3.1. Evaluation Metrics

To evaluate the prediction performance of the proposed CMAF method, 10-fold cross-validation was conducted on the gold standard dataset. In each round of 10-fold cross-validation, all the recorded drug-disease pairs were randomly divided into 10 equal-sized parts. Each part was taken as a test set in turn, while the remaining nine parts of the data were merged as the training set, thus generating 10 pairs of training sets and test sets. To obtain convincing results, 10-fold cross-validation was repeated 10 times, and the average value of 10-folds was taken as the final result. After performing association prediction based on the training set, we can obtain the prediction values for each association. Then, for each drug, the test drug-disease associations are ranked together with all unconfirmed drug-disease pairs (candidate associations) in descending order according to the predicted values. For each specific ranking threshold, four metrics: true positive (TP), false negative (FN), false positive (FP), and true negative (TN), can be obtained based on the ranking results. If a test association has a higher rank value than the given threshold, it is considered as a correctly identified positive sample. Likewise, a candidate association is considered a correctly identified negative sample if it has a lower rank than the given threshold.

To provide an intuitive explanation of the evaluation metrics, a confusion matrix is first defined, which is built by comparing actual values with predicted outcomes. The two classes are constructed with positives and negatives, as shown in **Table 1**.

Next, the evaluation metrics of the true positive rate (TPR) and false positive rate (FPR) can be defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

Where TP and FP represent the numbers of correctly and wrongly identified positive samples and TN and FN represent the numbers of correctly and wrongly identified negative samples; TPR and FPR are calculated based on these four metrics. Furthermore, TPR is the ratio of known drug-disease pairs that are correctly predicted, and FPR is the proportion of unconfirmed drug-disease pairs that are predicted.

After that, the receiver operating characteristic (ROC) curve can be drawn based on TPR and FPR at different thresholds. Meanwhile, the area under ROC (AUC) can be calculated to evaluate the prediction performance. The larger the value of the AUC, the better the prediction performance. For instance, if the value of the AUC is equal to 1, it means the best performance.

3.2. Comparison With Other Methods

In this section, to evaluate the ability of the proposed approach, we compare CMAF with four other recently proposed computational drug repositioning approaches: NBI (Cheng et al., 2012), BNNR (Yang et al., 2019), HGBI (Wang et al., 2013), and NGRHMDA (Huang et al., 2017). NBI is based on a bipartite network and constructs a two-step diffusion model for drug repositioning (Cheng et al., 2012). BNNR was developed to utilize a bounded nuclear norm regularization approach to construct the drug-disease matrix under the low-rank assumption (Yang et al., 2019). HGBI was proposed according to the guilt-by-association principle and an intuitive interpretation of information flow on a heterogeneous graph (Wang et al., 2013). NGRHMDA uses neighbor-based collaborative filtering and a graph-based scoring method to obtain the association score (Huang et al., 2017). Although HGBI and NBI were originally used to predict potential drug-target associations and NGRHMDA was originally used to predict new microbe-disease associations, they can also be used to predict new drug-disease associations. The parameter values used in NBI, BNNR, HGBI, and NGRHMDA are set based on their corresponding literature.

The predictive ability of all drug repositioning approaches is evaluated in terms of the AUC specified in section 3.1. As shown in **Figure 2**, the results demonstrate that our developed approach, CMAF, is superior to the other four drug repositioning approaches. In detail, CMAF obtains an AUC value of 0.941, while BNNR, HGBI, NBI, and NGRHMDA achieve inferior results of 0.931, 0.832, 0.583, and 0.503, respectively.

TABLE 1 | Confusion matrix.

		Actual value	
		Positive	**Negative**
Predicted value	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

3.3. Comparison of the Three Methods With Their Combined Model

The effectiveness of the fusion method is evaluated in this section. We performed drug-disease association prediction on

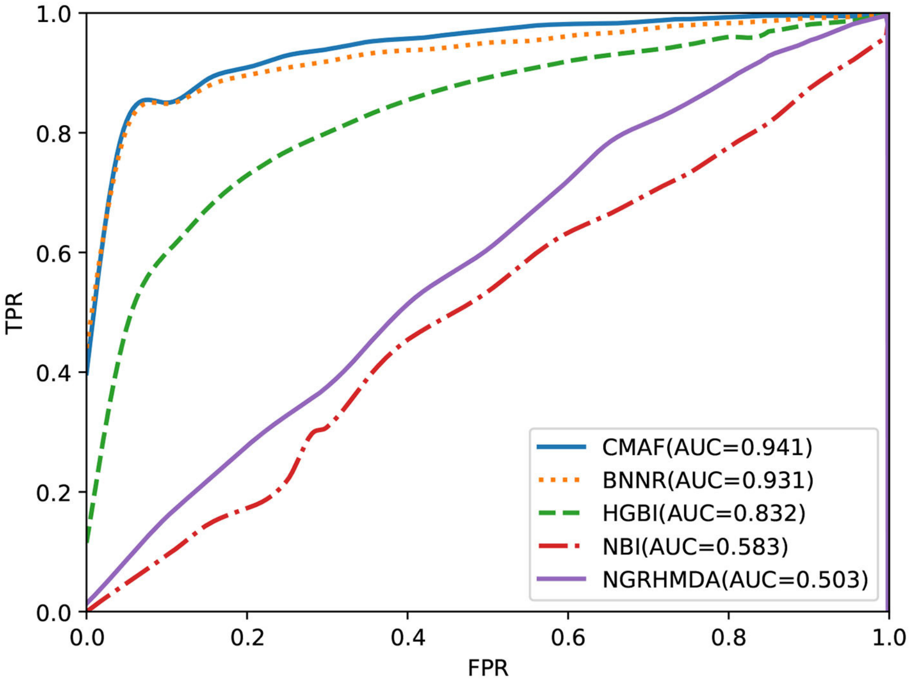


FIGURE 2 | Prediction results of various methods according to ROC curve analysis.

the gold standard dataset by using three methods (i.e., the LPRIA, NMFRIA, and NCPRIA methods) and their combined method. As shown in **Figure 3**, the AUC values of the three methods LPRIA, NMFRIA, and NCPRIA were 0.927, 0.923, and 0.920, respectively; however, the fusion method CMAF obtained an AUC value of 0.941. The experimental results

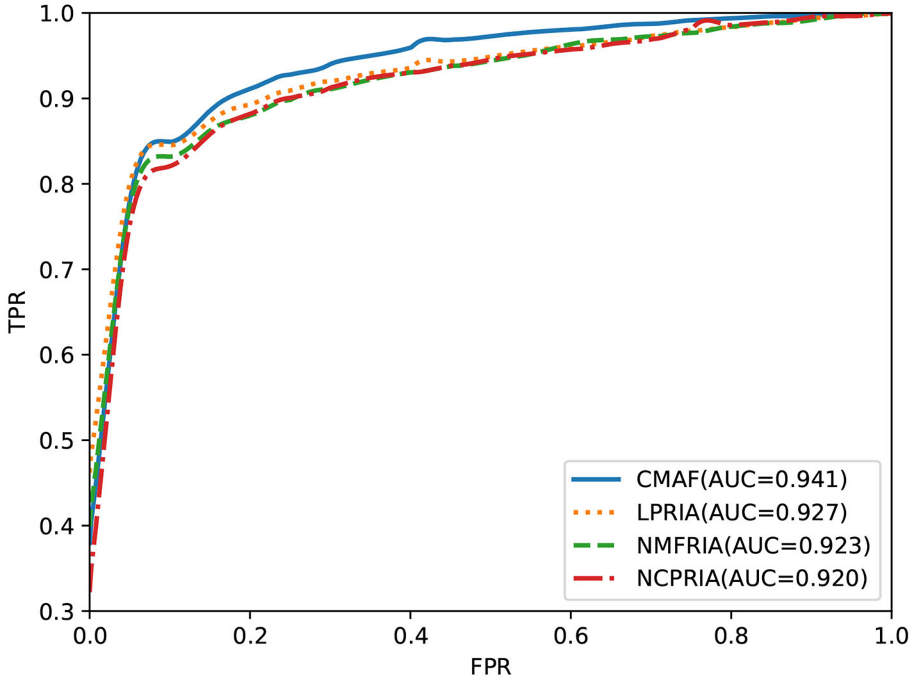


FIGURE 3 | Prediction performance of CMAF and the three individual methods according to the ROC curve.

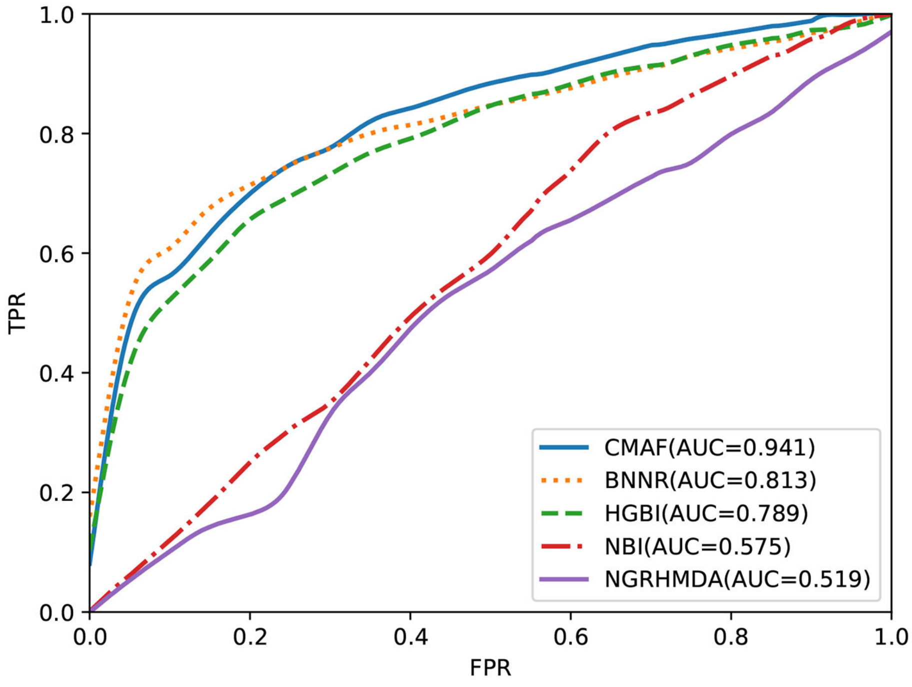


FIGURE 4 | Prediction performance of CMAF and the other four methods in predicting drug-disease associations for new drugs according to the ROC curves.

TABLE 2 | Case studies of four chosen drugs: levodopa, flecainide, zoledronic acid, and amantadine.

Drug (DrugBank IDs)	Top 5 candidate diseases (OMIM IDs)	Evidence
DB01235	168600	KEGG/DB/CTD
Levodopa	125320	DB/CTD
	165199	
	254770	
	190400	
DB01195	608583	CTD
Flecainide	194200	KEGG/CTD
	115000	DB/CTD
	157300	
	608622	CTD
DB00399	166710	KEGG/CTD
Zoledronic acid	102400	
	144700	CTD
	166300	
	114480	CTD
DB00915	168600	KEGG/DB/CTD
Amantadine	125320	DB/CTD
	605055	
	104300	CTD
	607225	

For each drug, the top five ranked predicted drugs are listed below.

illustrated the effectiveness of our fusion approach. Specifically, the CMAF method obtained the best performance among these four methods.

3.4. Prediction for New Drugs

To test the predictive performance of CMAF for new drugs, a de novo prediction test was executed. In de novo drug validation, for each of the drugs, we deleted all of its known associations, and they were used for testing samples in turn; the other known drug-disease association was used as the training sample. The rankings of the removed drug-disease associations relative to the drug candidate associations were obtained by *de novo* testing, which was used to assess the predictive performance. To compare the predictive ability of different methods in *de novo* testing of new drugs, the other four prediction methods also underwent de novo prediction tests. The experimental results are shown in **Figure 4**, and the graph demonstrates that our CMAF is superior to the other approaches. In detail, CMAF obtains an AUC value of 0.941, while the results of BNNR, HGBI, NBI, and NGRHMDA are 0.813, 0.789, 0.575, and 0.519, respectively.

4. CASE STUDIES

After verifying the predicted performance of CMAF in terms of 10-fold cross-validation, the ability of our proposed model to identify new indications for a given drug is further validated here. To predict new drug-disease interactions, all known drug-disease pairs are considered as the training set, and the remaining unknown drug-disease pairs form the candidate set. By applying our CMAF method, we can obtain all the candidates' set prediction scores. According to the prediction scores, for every drug, all the candidate diseases are ranked.

As an example, we selected some drugs and the corresponding top five candidate diseases as verified information, and then we found that some of them were confirmed in the KEGG (Kanehisa et al., 2013), DrugBank and CTD (Davis et al., 2014) databases, as shown in **Table 2**. For example, the effectiveness of levodopa in treating Parkinson's disease (PD) due to its ability to cross the blood-brain barrier can be retrieved from the KEGG, DrugBank, and CTD databases. In addition, relevant literature has shown that levodopa-treated patients have gained improvement in most Parkinsonian features in the past half-century (Lewitt, 2015). Flecainide is helpful for treating atrial fibrillation, as can be retrieved from CTD, and there is literature to prove that in clinical trials and real-world use, flecainide is more effective than other antiarrhythmic drugs (AADs) for the acute termination of recent-onset atrial fibrillation (Echt and Ruskin, 2020). From KEGG and CTD, zoledronic acid can be found to treat and prevent multiple forms of osteoporosis. There is also literature to prove that zoledronic acid administered once yearly for up to 3 years improved bone mineral density (BMD) at several skeletal sites, reduced fracture risk and bone turnover, and/or preserved bone structure and mass relative to placebo in clinical studies in patients with primary or secondary osteoporosis (Dhillon, 2016). Amantadine is an antiviral that can be used to cure PD and can be retrieved from KEGG, DB, and CTD. Relevant literature suggests that amantadine is an old antiviral compound that moderately ameliorates impaired motor behavior in Parkinson's disease (Müller et al., 2019).

5. CONCLUSION

This work proposed a new computational drug repositioning model named CMAF to find new uses for existing drugs. First, the number of known drug-disease interactions is far less than that of unknown drug-disease interactions in practice, which leads to the problem of data sparseness for drug repositioning. Therefore, we used the WKNKN method as a pre-processing step to compute the temporary association scores for these unknown drug-disease interactions in terms of their known neighbors, and then we computed the linear neighborhood similarity for drugs and diseases. After that, the LPRIA, NMFRIA, and NCPRIA methods were adopted to obtain three predictive association possibilities. Finally, we adopted an ensemble strategy to fuse these three prediction models to obtain the hopefully final prediction result. Compared with several recent computational drug repositioning models, our proposed CMAF approach achieves better predictive performance.

Even though our proposed method obtains promising results, it still has some limitations. First, we plan to consider integrating more predictive methods into the ensemble strategy. Second, CMAF utilizes only single drug-drug similarity and disease-disease similarity to construct prediction methods. In the future, we will compute multiple drug-drug similarities and disease-disease similarities and combine diverse similarities to further improve the predictive performance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CY and JW conceived and designed the approach. WW performed the experiments. JL analyzed the data. GZ and WW wrote the manuscript. CY and GZ supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 61802113, 61802114, and 61972134), Science and Technology Development Plan Project of Henan Province (Nos. 202102210173 and 212102210091), China Post-doctoral Science Foundation (No. 2020M672212), and Henan Province Post-doctoral Research Project Founding.

ACKNOWLEDGMENTS

This paper was recommended by the 5th Computational Bioinformatics Conference.

REFERENCES

Adams, C. P., and Brantner, V. V. (2006). Estimating the cost of new drug development: is it really $802 million? *Health Affairs* 25, 420–428. doi: 10.1377/hlthaff.25.2.420

Ashburn, T. T., and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov*. 3, 673–683. doi: 10.1038/nrd1468

Azad, A. K. M., Dinarvand, M., Nematollahi, A., Swift, J., Lutze-Mann, L., and Vafaee, F. (2020). A comprehensive integrated drug similarity resource for *in-silico* drug repositioning and beyond. *Brief. Bioinform*. doi: 10.26434/chemrxiv.12376505.v1. [Epub ahead of print].

Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., et al. (2017). Automatic selection of molecular descriptors using random forest: application to drug discovery. *Expert Syst. Appl*. 72, 151–159. doi: 10.1016/j.eswa.2016.12.008

Chen, H., and Li, J. (2017). "A flexible and robust multi-source learning algorithm for drug repositioning," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, ACM-BCB '17* (New York, NY: Association for Computing Machinery), 510–515. doi: 10.1145/3107411.3107473

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y. P. P., et al. (2019). ILDMSF: Inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2019.2936476. [Epub ahead of print].

Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., et al. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol*. 8:e1002503. doi: 10.1371/journal.pcbi.1002503

Davis, A. P., Grondin, C., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B., et al. (2014). The comparative toxicogenomics database's 10th year anniversary: Update 2015. *Nucleic Acids Res*. 43, D914–D920. doi: 10.1093/nar/gku935

Dhillon, S. (2016). Zoledronic acid (Reclast(®), Aclasta(®)): a review in osteoporosis. *Drugs* 76, 1683–1697. doi: 10.1007/s40265-016-0662-4

Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. (2014). Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Brief. Bioinform*. 15, 734–747. doi: 10.1093/bib/bbt056

Echt, D., and Ruskin, J. (2020). Use of flecainide for the treatment of atrial fibrillation. *Am. J. Cardiol*. 125, 1123–1133. doi: 10.1016/j.amjcard.2019.12.041

Ezzat, A., Zhao, P., Wu, M., Li, X. L., and Kwoh, C. K. (2017). Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 14, 646–656. doi: 10.1109/TCBB.2016.2530062

Fu, G., Wang, J., Domeniconi, C., and Yu, G. X. (2017). Matrix factorization based data fusion for the prediction of lncrna-disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Fujita, N., Mizuarai, S., Murakami, K., and Nakai, K. (2018). Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Sci. Rep*. 8:9743. doi: 10.1038/s41598-018-28066-w

Gönen, M., and Kaski, S. (2014). Kernelized bayesian matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell*. 36, 2047–2060. doi: 10.1109/TPAMI.2014.2313125

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). Predict: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol*. 7:496. doi: 10.1038/msb.2011.26

Gu, C., Liao, B., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. *Sci. Rep*. 6:36054. doi: 10.1038/srep36054

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. A. (2002). Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 30, 52–55. doi: 10.1093/nar/30.1.52

Huang, Y. A., You, Z. H., Huang, Z. A., Zhang, S., and Yan, G. (2017). Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med*. 15:209. doi: 10.1186/s12967-017-1304-7

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 42, D199–D205. doi: 10.1093/nar/gkt1076

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). LDICDL: lncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2020.3034910. [Epub ahead of print].

Lewitt, P. (2015). Levodopa therapy for Parkinson's disease: pharmacokinetics and pharmacodynamics. *Mov. Disord*. 30, 64–72. doi: 10.1002/mds.26082

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2015). A survey of current trends in computational drug repositioning. *Brief. Bioinform*. 17, 2–12. doi: 10.1093/bib/bbv020

Liang, X., Zhang, P., Yan, L., Fu, Y., Peng, F., Qu, L., et al. (2017). LRSSL: predict and interpret drug-isease associations based on data integration using sparse subspace learning. *Bioinformatics* 33, 1187–1196. doi: 10.1093/bioinformatics/btw770

Lipscomb, C. E. (2000). Medical subject headings (MESH). *Bull. Med. Libr. Assoc*. 88, 265–266.

Liu, H., Song, Y., Guan, J., Luo, L., and Zhuang, Z. (2016a). Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinformatics* 17:539. doi: 10.1186/s12859-016-1336-7

Liu, X., Zhai, D., Zhao, D., Zhai, G., and Gao, W. (2014). Progressive image denoising through hybrid graph laplacian regularization: a unified framework. *IEEE Trans. Image Process*. 23, 1491–1503. doi: 10.1109/TIP.2014.2303638

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X. (2016b). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol*. 12:e1004760. doi: 10.1371/journal.pcbi.1004760

Luo, H., Li, M., Wang, S., Liu, Q., Li, Y., and Wang, J. (2018). Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 34, 1904–1912. doi: 10.1093/bioinformatics/bty013

Luo, H., Li, M., Yang, M., Wu, F. X., Li, Y., and Wang, J. (2020). Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief. Bioinform*. 22, 1604–1619. doi: 10.1093/bib/bbz176

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F. X., et al. (2016). Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* 32, 2664–2671. doi: 10.1093/bioinformatics/btw228

Martìnez, V., Navarro, C., Cano, C., Fajardo, W., and Blanco, A. (2015). Drugnet: network-based drug-disease prioritization by integrating heterogeneous data. *Artif. Intell. Med*. 63, 41–49. doi: 10.1016/j.artmed.2014.11.003

Müller, T., Kuhn, W., and Möhr, J. D. (2019). Evaluating ADS5102 (amantadine) for the treatment of Parkinson's disease patients with dyskinesia. *Expert Opin. Pharmacother*. 20, 1181–1187. doi: 10.1080/14656566.2019.1612365

Napolitano, F., Zhao, Y., M Moreira, V., and Tagliaferri, R. (2013). Drug repositioning: a machine-learning approach through data integration. *J. Cheminform*. 5:30. doi: 10.1186/1758-2946-5-30

Shameer, K., Readhead, B., and Dudley, J. T. (2015). Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Curr. Top. Med. Chem*. 15, 5–20. doi: 10.2174/1568026615666150112103510

Shi, J. Y., Zhang, A. Q., Zhang, S. W., Mao, K. T., and Yiu, S. M. (2018). A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization. *BMC Syst. Biol*. 12:136. doi: 10.1186/s12918-018-0663-x

Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., and Willighagen, E. L. (2006). Recent developments of the chemistry development kit (CDK)–an open-source Java library for chemo-and bioinformatics. *Curr. Pharm. Des*. 12, 2111–2120. doi: 10.2174/138161206777585274

Tanimoto, T. T. (1958). *An Elementary Mathematical Theory of Classification and Prediction*. New York, NY: International Business Machines Corporation.

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*. 14, 535–542. doi: 10.1038/sj.ejhg.5201585

Vilar, S., and Hripcsak, G. (2017). The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. *Brief. Bioinform*. 18, 670–681. doi: 10.1093/bib/bbw048

Wang, C., and Kurgan, L. (2019). Review and comparative assessment of similarity-based methods for prediction of drug-protein interactions in the druggable human proteome. *Brief. Bioinform*. 20, 2066–2087. doi: 10.1093/bib/bby069

Wang, F., and Zhang, C. (2008). Label propagation through linear neighborhoods. *IEEE Trans. Knowl. Data Eng*. 20, 55–67. doi: 10.1109/TKDE.2007.190672

Wang, W., Yang, S., and Li, J. (2013). Drug target predictions based on heterogeneous graph inference. *Pac. Symp. Biocomput*. 18, 53–64. doi: 10.1142/9789814447973_0006

Weininger, D. (1988). Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci*. 28, 31–36. doi: 10.1021/ci00057a005

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). Drugbank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res*. 34, D668–D672. doi: 10.1093/nar/gkj067

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microrna-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35, i455–i463. doi: 10.1093/bioinformatics/btz331

Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8:S2. doi: 10.1186/1755-8794-8-S2-S2

Zhang, P., Wang, F., Hu, J., and Sorrentino, R. (2015). Label propagation prediction of drug-drug interactions based on clinical side effects. *Sci. Rep*. 5:12339. doi: 10.1038/srep12339

Zhang, W., Chen, Y., and Li, D. (2017). Drug-target interaction prediction through label propagation with linear neighborhood information. *Molecules* 22:2056. doi: 10.3390/molecules22122056

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* 19:233. doi: 10.1186/s12859-018-2220-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 27 May 2021
doi: 10.3389/fgene.2021.670852

m6AGE: A Predictor for N6-Methyladenosine Sites Identification Utilizing Sequence Characteristics and Graph Embedding-Based Geometrical Information

Yan Wang[1,2], Rui Guo[1], Lan Huang[1], Sen Yang[1]*, Xuemei Hu[1] and Kai He[1]

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, and College of Computer Science and Technology, Jilin University, Changchun, China, [2] School of Artificial Intelligence, Jilin University, Changchun, China

OPEN ACCESS

Edited by:
Wilson Wen Bin Goh,
Nanyang Technological University,
Singapore

Reviewed by:
Panagiotis Alexiou,
Central European Institute
of Technology (CEITEC), Czechia
Jia Meng,
Xi'an Jiaotong-Liverpool University,
China

***Correspondence:**
Sen Yang
ystop2020@gmail.com

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 22 February 2021
Accepted: 29 April 2021
Published: 27 May 2021

Citation:
Wang Y, Guo R, Huang L, Yang S,
Hu X and He K (2021) m6AGE:
A Predictor for N6-Methyladenosine
Sites Identification Utilizing Sequence
Characteristics and Graph
Embedding-Based Geometrical
Information. Front. Genet. 12:670852.
doi: 10.3389/fgene.2021.670852

N^6-methyladenosine (m^{6}A) is one of the most prevalent RNA post-transcriptional modifications and is involved in various vital biological processes such as mRNA splicing, exporting, stability, and so on. Identifying m^{6}A sites contributes to understanding the functional mechanism and biological significance of m^{6}A. The existing biological experimental methods for identifying m^{6}A sites are time-consuming and costly. Thus, developing a high confidence computational method is significant to explore m^{6}A intrinsic characters. In this study, we propose a predictor called m6AGE which utilizes sequence-derived and graph embedding features. To the best of our knowledge, our predictor is the first to combine sequence-derived features and graph embeddings for m^{6}A site prediction. Comparison results show that our proposed predictor achieved the best performance compared with other predictors on four public datasets across three species. On the *A101* dataset, our predictor outperformed 1.34% (accuracy), 0.0227 (Matthew's correlation coefficient), 5.63% (specificity), and 0.0081 (AUC) than comparing predictors, which indicates that m6AGE is a useful tool for m^{6}A site prediction. The source code of m6AGE is available at https://github.com/bokunoBike/m6AGE.

Keywords: m^{6}A, machine learning, graph embedding, feature fusion, CatBoost

INTRODUCTION

N^6-methyladenosine (m^{6}A) is one of the most prevalent RNA post-transcriptional modifications. It was first found in mammalian RNA in 1974 (Desrosiers et al., 1974). Subsequently, m^{6}A modification was observed in various species, such as Saccharomyces cerevisiae (Schwartz et al., 2013), Arabidopsis (Luo et al., 2014), humans and mouse (Dominissini et al., 2012). Research shows that m^{6}A sites are enriched in long internal exons and 3′UTRs around stop codons rather than randomly distributed in the genome (Dominissini et al., 2012; Meyer et al., 2012; Wan et al., 2015). It has been reported that m^6A modification is associated with many biological processes, including but not limited to protein translation and localization (Meyer and Jaffrey, 2014), mRNA splicing and stability (Nilsen, 2014), RNA localization and degradation (Meyer and Jaffrey, 2014). Therefore, precisely identifying m^6A sites contributes to understanding the regulatory mechanism and biological significance of m^6A modification.

High-throughput techniques have enabled locating the m^6A sites in genomes. MeRIP-Seq (or m6A-Seq), a combination of immunoprecipitation and next-generation sequencing technology, has successfully mapped m^6A in several species genomes (Dominissini et al., 2012; Schwartz et al., 2013; Wan et al., 2015). In 2015, Chenet al. developed photo-crosslinking-assisted m^6A-sequencing (PA-m^6A-seq) which provided a high-resolution (about 23nt) mammalian map (Chen et al., 2015a). MeRIP-Seq and PA-m6A-seq can only locate the high methylation regions of m^6A rather than the exact positions. In the same year, Linder produced a single-nucleotide resolution map of m^6A sites using a new technology termed miCLIP (Linder et al., 2015). However, the current experimental methods face a lot of limitations and expensive costs. With the rapid development of computational methods, it is possible to use machine learning algorithms to predict m^6A. Hence, building advanced models to predict m^6A sites is significant for the following research of m^6A.

Chen et al. (2015b) proposed the first predictor named iRNA-Methyl for m^6A sites in Saccharomyces cerevisiae, using three physical-chemical properties of dinucleotide and SVM classifier. WHISTLE (Chen et al., 2019) integrates genomic features besides the sequence features to train a predictor with SVM classifier. Liu and Chen (2020) developed a computational method called iMRM for detecting different RNA modifications simultaneously with XGBoost classifier. Recently, deep learning methods show better performance trend in bioinformatics problems. DeepM6ASeq (Zhang and Hamada, 2018), BERMP (Huang et al., 2018), Gene2vec (Zou et al., 2019), DeepPromise (Chen et al., 2020), and im6A-TS-CNN (Liu et al., 2020) establish deep learning frameworks by using convolutional neural network (CNN) layers and gated recurrent unit (GRU) to seek the m^6A sites on DNA/RNA sequence level on the same dataset as SRAMP (Zhou et al., 2016). In this study, seven kinds of sequence-derived features are employed to encode RNA sequences, including CTD (Tong and Liu, 2019), Pseudo k-tuple Composition (PseKNC) (Guo et al., 2014), nucleotide pair spectrum (NPS) (Zhou et al., 2016), nucleotide pair position specificity (NPPS) (Xing et al., 2017), nucleotide chemical properties and density (NCP-ND) (Golam Bari et al., 2013), electron-ion interaction pseudopotentials (EIIP) (Nair and Sreenadhan, 2006), and bi-profile Bayes (BPB) (Shao et al., 2009). Besides, graph embedding methods are innovatively introduced to distill the potential structure information. Firstly, a network is constructed by mapping each sample of the dataset to a node. Secondly, the three graph embedding methods SocDim (Tang and Liu, 2009), Node2Vec (Grover and Leskovec, 2016), and GraRep (Cao et al., 2015) are used to learn the distributed representation of the sample in an unsupervised manner. At last, all the feature vectors are merged as the input of model. The predictive results show that m6AGE improves the performance of identifying m^6A sites.

MATERIALS AND METHODS

Datasets

The m^6A sites of different species share different consensus motifs. The adenosines lying within the consensus motif are considered to be the potential methylation sites. The samples in the dataset are RNA sequence segments with the potential methylation sites at their center. The samples with the m^6A sites experimentally annotated are put into the positive dataset, whereas the other samples are put into the negative dataset.

There have been many datasets across multiple species for training m^6A site predictors. We have collected four datasets that involve three species: Saccharomyces cerevisiae, Arabidopsis thaliana, and human. The following are details of these datasets.

A101. Wang extracted *A.thaliana* m^6A sites from the m^6A peak data of Luo et al. (2014) and Wan et al. (2015). The dataset (Wang and Yan, 2018) Wang built contains 2,518 positive samples and 2,518 negative samples. Every sample in the dataset is a 101nt RNA sequence segment.

A25. Luo obtained 4,317 m^6A peaks detected both in Can-0 and Hen-16 strains. After removing the sequences with more than 60% sequence similarity, Chen et al. (2016) obtained 394 positive samples. The same number of negative samples were selected randomly from sequences without the m^6A site. The length of every sample is 25nt.

S21. Chen further constructed this dataset (Chen et al., 2015c) based on the previous work (Chen et al., 2015b). They selected 832 RNA sequence segments as the positive samples in the training set whose distances to the m^6A-seq peaks are less than 10nt. Then, 832 of 33,280 RNA sequence segments with non-methylated adenines were selected randomly as negative samples in the training set. The rest 475 RNA sequences with methylated adenine and 4750 of 33,280 RNA sequences with non-methylated adenine constitute the independent testing dataset. The length of every sample is 21nt.

H41. Chen obtained the m^6A-containing sequences in *Homo sapiens* from RMBase (Chen et al., 2017). All the m^6A sites in these sequences conform to the RRACH motif. The dataset contains 1,130 positive samples and 1,130 negative samples. The length of every sample is 41nt.

Construction of Input Feature

Conventional machine learning models require numerical vectors as input features. The feature extraction methods selected have an important impact on the performance of the model. To fully characterize the context of m^6A sites, seven sequence-derived features were used. In addition, we build a network based on the whole dataset, by mapping each sample to node and the similarity between samples to edges in the network, and then use graph embedding (neighborhood-based node embedding) methods to extract features in an unsupervised manner. The

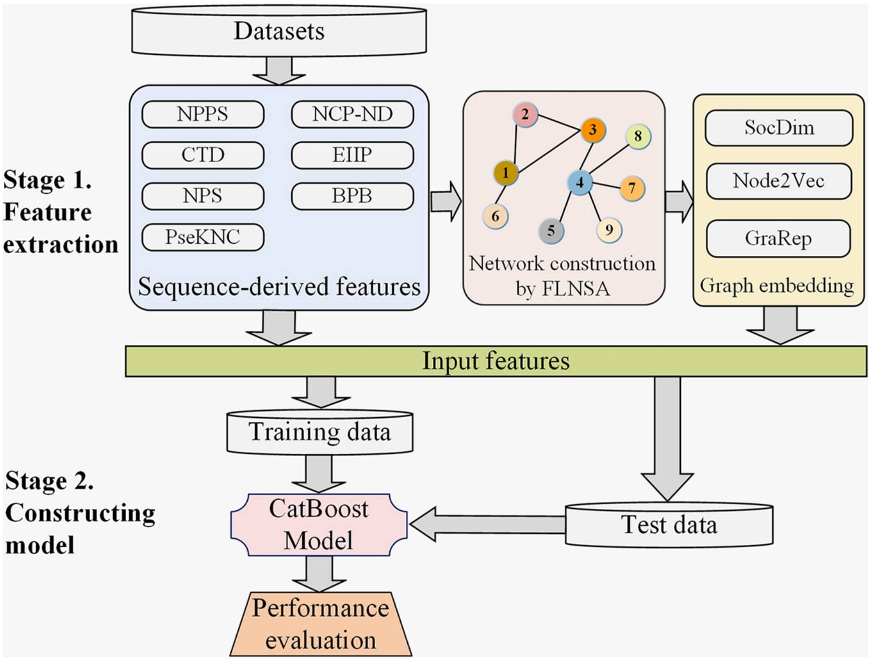


FIGURE 1 | The computational framework of our predictor m6AGE. There are two main stages in the construction of m6AGE. Stage 1. Sequence-derived features are extracted, and graph embeddings are learned. Sequence-derived feature encoding methods directly encode RNA sequences into numerical vectors, including CTD, NPS, PseKNC, NPPS, NCP-ND, EIIP, and BPB feature encoding method. All the sequences are mapped to nodes of a network, and then their graph embeddings (SocDim, Node2Vec, and GraRep) are learned in an unsupervised manner. At last, the sequence features and graph embeddings are merged as input features. Stage 2. We divide the data into training data and test data with a ratio of 4:1. The training data is used to train a CatBoost model. The test data is used to evaluate the performance of our predictor.

computational framework of our predictor is illustrated in **Figure 1**. In the following, we will introduce the sequence-derived features and the graph embeddings, respectively.

Sequence-Derived Features

CTD Feature

CTD (Tong and Liu, 2019) is one of the global sequence descriptors. The first descriptor C (nucleotide composition) describes the percentage composition of each nucleotide in the sequence. The second descriptor T (nucleotide transition) describes the frequency of four different nucleotides present in adjacent positions. The third descriptor D (nucleotide distribution) describes five relative positions of each nucleotide along the RNA sequence which are the first one, 25%, 50%, 75%, and the last one.

PseKNC Feature

With the successful application of the pseudo component method in peptide sequence processing, its idea has been further extended to the study of DNA and RNA sequences feature representation. The Pseudo k-tuple Composition (PseKNC) combines the local and global sequence information of RNA (Guo et al., 2014) and transforms an RNA sequence into the following vector:

$$\mathbf{D}_{\text{PseKNC}}=\left[d_1, d_2, \ldots, d_{4^k}, d_{4^k+1}, \ldots, d_{4^k+\lambda}\right]^T \tag{1}$$

where,

$$d_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j}\left(1 \le u \le 4^k\right) \\ \frac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j}\left(4^k < u \le 4^k+\lambda\right) \end{cases} \tag{2}$$

where $d_u\left(u = 1, 2, ..., 4^k\right)$ is the occurrence frequency of the u-th k-nucleotide in this RNA sequence; the parameter w is the weight factor; the parameter λ is the number of totals counted tiers of the correlations along an RNA sequence. The j-tier correlation factor θ_j is defined as follows:

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Theta\left(R_iR_{i+1}, R_{i+j}R_{i+j+1}\right),$$

$$\left(j = 1, 2, \ldots, \lambda; \lambda < L\right) \tag{3}$$

The correlation function $\Theta\,(,)$ is calculated by the following formula:

$$\Theta\left(R_iR_{i+1}, R_{i+j}R_{i+j+1}\right)=\frac{1}{\mu}\sum_{\nu=1}^{\mu}\left[P_\nu\left(R_iR_{i+1}\right)-P_\nu\left(R_{i+j}R_{i+j+1}\right)\right]^2 \tag{4}$$

where μ is the number of RNA physicochemical properties used. R_iR_{i+1} is the dinucleotide at position i of this RNA.

$P_\nu\ (R_iR_{i+1})$ is the standardized numerical value of the ν-th RNA physicochemical properties for dinucleotide R_iR_{i+1}.

Six RNA physicochemical properties are considered: “Rise”, “Roll”, “Shift”, “Slide”, “Tilt”, “Twist”.

NPS Feature

The nucleotide pair spectrum (NPS) (Zhou et al., 2016) encoding method describes the RNA sequence context of the site by calculating the occurrence frequency of all *k*-spaced nucleotide pairs in the sequence. The *k*-spaced nucleotide pair $n_1\{k\}n_2$ means that there are *k* arbitrary nucleotides between n_1 and n_2, and its occurrence frequency is calculated as follows:

$$d_{n_1\{k\}n_2} = \frac{C\left(n_1\{k\}n_2\right)}{L-k-1} \tag{5}$$

where $C\left(n_1\{k\}n_2\right)$ is the count of $n_1\{k\}n_2$ in this RNA sequence, and L is the sequence length. The parameter k ranges from 1 to d_{max}. The parameter d_{max} is set to 3, so this encoding method transforms an RNA sequence into a vector $\mathbf{D}_{NPS}$ with a dimension of $4 \times 4 \times 3 = 48$.

NPPS Feature

The nucleotide pair position specificity (NPPS) (Xing et al., 2017) encoding method extracts statistical information by calculating the frequency of single nucleotide and *k*-spaced nucleotide pairs at specific locations. Based on the positive training dataset, we can get the frequency matrix

$$F_s^+ = \begin{bmatrix} f^+_{s(A,1)} & \cdots & f^+_{s(A,L)} \\ \vdots & \ddots & \vdots \\ f^+_{s(G,1)} & \cdots & f^+_{s(G,L)} \end{bmatrix} \tag{6}$$

$$F_d^+ = \begin{bmatrix} f^+_{d(AA,1)} & \cdots & f^+_{d(AA,L-k-1)} \\ \vdots & \ddots & \vdots \\ f^+_{d(GG,1)} & \cdots & f^+_{d(GG,L-k-1)} \end{bmatrix} \tag{7}$$

where the element of F_s^+ is the frequency of single nucleotide appearing at each location in the positive training dataset; the element of F_d^+ is the frequency of *k*-spaced nucleotide pair appearing at each location in the positive training dataset; and L is the sequence length. The frequency matrix F_s^- and F_d^- are calculated similarly on the negative training dataset.

Assuming that the *i*-th nucleotide is “A” and the $(i+k)$-th nucleotide is “C”, p_i^+ is calculated through conditional probability formula and frequency matrix:

$$p_i^+ = \frac{f^+_{d(AC,i)}}{f^+_{s(C,i+k)}} \tag{8}$$

NPPS encoding method transforms a sequence into a vector $\mathbf{D}_{NPPS} = \left[p_{k+2}, \ldots, p_L\right]$ with a dimension of $L-k-1$, where $p_i = p_i^+ - p_i^-$.

NCP-ND Feature

Different nucleotides have different chemical properties. According to the difference of ring structure (purine or pyrimidine), hydrogen bond (strong or weak), and functional group (amino or keto), nucleotide A, U, C, and G can be represented by (1, 1, 1), (0, 1, 0), (0, 0, 1), and (1, 0, 0), respectively (Golam Bari et al., 2013).

The nucleotide density (ND) is used to measure the relevance between the frequency and position of the *i*-th nucleotide n_i in the sequence:

$$d_{n_i} = \frac{1}{i}\sum_{j=1}^{L} t\left(n_j\right), \quad t\left(q\right) = \begin{cases} 1, & \mathit{if}\ n_j = q \\ 0, & \mathit{othercase} \end{cases} \tag{9}$$

where L is the sequence length. Combined with the chemical properties of nucleotides, each sequence is transformed into a vector $\mathbf{D}_{NCP-ND}$ with a dimension of $L \times 4$.

EIIP Feature

This encoding method uses the electron-ion interaction pseudopotentials (EIIP) values (Nair and Sreenadhan, 2006) to represent the nucleotide in the sequence. The EIIP values of nucleotides A, T (we replace T with U), C, G are 0.1260, 0.1340, 0.0806, and 0.1335, respectively. Thus the dimension of the vector $\mathbf{D}_{EIIP}$ is equal to the sequence length.

BPB Feature

The Bi-profile Bayes (BPB) encoding method was first proposed by (Shao et al., 2009), and then has been successfully applied in other fields of bioinformatics. This method uses the occurrence frequency $f_{i,n}$ of the *i*-th nucleotide n to estimate the posterior probability $p_{i,n}$, and transforms a sequence into the following vector:

$$\mathbf{D}_{BPB} = \left[f^+_{1,n}, f^-_{1,n}, f^+_{2,n}, f^-_{2,n}, \ldots f^+_{L,n}, f^-_{L,n}\right] \tag{10}$$

where n is the *i*-th nucleotide of the sequence; $f^+_{i,n}$ denotes the frequency of nucleotide n appearing at the *i*-th position of the sequence in the positive training dataset, while $f^-_{i,n}$ denotes the frequency of nucleotide n appearing at the *i*-th position of sequence in the negative training dataset. L is the sequence length. The dimension of the vector $\mathbf{D}_{BPB}$ is $2 \times L$.

Graph Embeddings

Network Construction

To extract the graph embedding feature of each sample, we construct a network based on the whole dataset. Each sample in the dataset is taken as a node, and the relationships between samples are taken as edges. Generally, edges exist two similar sample nodes. The fast linear neighbor similarity approach (FLNSA) (Zhang et al., 2017, 2019) is a method to extract “sample-sample” similarity, which has been successfully applied to many bioinformatics classification tasks. In this study, FLNSA is utilized to calculate the similarity between samples.

First, we extract sequence-derived features and use the feature fusion strategy to transform all the samples in the dataset into *n*-dimensional vector $\{x_1, x_2, \ldots, x_m\}$, where $x_i\ (0 < i \leq m)$ is the vector of the *i*-th sample. Then these vectors are concentrated into a matrix $X \in R^{m \times n}$, each row of which represents a sample vector. FLNSA tries to minimize the objective function:

$$\min_{w} \frac{1}{2}\left\|X-\left(C\bigodot W\right)X\right\|_F^2+\frac{\mu}{2}\sum_{i=1}^{m}\left\|\left(C\bigodot W\right)\mathbf{e}\right\|_F^2 \tag{11}$$

$$s.t.\left(C\bigodot W\right)\mathbf{e} = \mathbf{e}, W \geq 0$$

where $\bigodot$ is the Hadamard product operator; $||\cdot||_F$ represents the Frobenius norm and μ is the regularization coefficient. **e** is an m-dimensional column vector with all elements equal to 1. The element $w_{i,j}$ of matrix $W \in R^{m \times m}$ represents the reconstruction contribution weight of the sample x_j to the sample x_i, and is used to quantify the similarity between two samples. The element of indicator $C \in R^{m \times m}$ is

$$c_{i,j} = \begin{cases} 1 & x_j \in N(x_i) \\ 0 & x_j \notin N(x_i) \end{cases} \tag{12}$$

where $N(x_i)$ denotes the set of all neighbors of x_i. The Euclidean distances between x_i and other samples are calculated and the nearest $c\,(0 < c < m)$ samples are selected to form $N(x_i)$. FLNSA uses the Lagrange method to get matrix W. After mathematical derivation, the Equation (13) is obtained.

$$W_{ij}=\begin{cases} \frac{W_{ij}\left(XX^T+\mu\mathbf{ee}^T\right)_{ij}}{\left(\left(C\bigodot \mathrm{W}\right)XX^T+\mu\left(C\bigodot \mathrm{W}\right)\mathbf{ee}^T\right)_{ij}} & x_j \in N(x_i) \\ 0 \quad x_j \notin N(x_i) \end{cases} \tag{13}$$

Randomly generated matrix W was updated according to Equation (13) until convergence. Taking W as the adjacency matrix, an undirected weighted graph G is obtained. The graph embedding methods require a connected graph as input. Note that if G is not connected, we can increase c (the number of neighborhoods of a sample). Under the condition of ensuring the connectivity of the graph, the edges whose weights are lower than the threshold t are removed and the weights of the remaining edges are set to 1. Finally, an undirected unweighted graph is constructed based on the dataset.

SocDim

The social-dimension-based (SocDim) (Tang and Liu, 2009) method is proposed by Lei Tang and Huan Liu to solve the relational learning between nodes in social networks. This method extracts latent dimensions from networks and uses them as distributed representations, which involves community detection tasks.

SocDim uses Modularity (Newman, 2006) which measures community structure through degree distribution to extract potential dimensions. Modularity considers dividing the network into non-overlapping communities, measures the deviation between the network and uniform random graphs with the same degree distribution, and then obtains the modularity matrix B defined as follows:

$$B=A-\frac{\mathbf{dd}^T}{2m} \tag{14}$$

where A is the interaction matrix of the network; **d** is a column vector composed of the degrees of each node; m is the number of nodes. Subsequently, SocioDim extracts the dimensions from the top eigenvectors of the modularity matrix B.

Node2Vec

Node2Vec (Grover and Leskovec, 2016) attempts to design a graph embedding model that can train efficiently and retain the neighborhood information of nodes to the maximum extent. The embedding vectors of nodes are learned through the skip-gram model. Different from DeepWalk, Node2Vec proposes biased random walk instead of truncated random walk to control the search space. Node2vec considers the homophily (nodes from the same community have similar embeddings) and structural equivalence (nodes that share similar roles have similar embeddings), thus there are two classic search strategies: Breadth-first Sampling (BFS) and Depth-first Sampling (DFS).

GraRep

GraRep (Cao et al., 2015) proposes a graph embedding model that can be learned from weighted graphs and integrate global structure information of the graph. GraRep forms k different vectors by separating k kinds of relationships. For a specific k, GraRep samples a set of k-step paths from the graph. The k-step path which starts with node v_w and ends with node v_c is denoted as (v_w, v_c). For all pairs, it increases the probability of the pairs come from the graph and decreases the probability of the pairs do not come from the graph. Based on the normalized adjacency matrix, GraRep obtains W^k for different values of k, and each column vector of W^k represents an embedding of the node. Finally, this method concatenates all the k-step representations W^1, W^2, ..., W^k.

CatBoost Classifier

CatBoost (Dorogush et al., 2018; Prokhorenkova et al., 2018) is an improved implementation of gradient enhanced decision trees (GDBT) algorithm developed by Yandex. It has demonstrated excellent performance on many classification and regression tasks. Compared with other advanced gradient boosting algorithms such as XGBoost (Chen and Guestrin, 2016) and lightBGM (Ke et al., 2017), CatBoost has the following advantages: (1) It can better process categorical features. (2) To solve the problem of gradient bias and prediction shift, ordered boosting is proposed instead of the classic GDBT gradient estimation algorithm. (3) The requirement of super parameter tuning is reduced.

CatBoost uses oblivious decision trees (Langley and Sage, 1994) as base predictors. As oblivious decision trees are balanced, they can prevent overfitting. Moreover, it optimizes the traditional boosting algorithm which transforms the category features into numerical features, and the algorithm of calculating the leaf value to improve the generalization ability of the model. Since the CatBoost algorithm is running on GPU, the model is trained efficiently and parallelly.

Evaluation Metrics

Our predictor predicts whether the adenosine at the center of an RNA sequence segment is an m^{6}A site. We used the following metrics to evaluate the performance of binary classification predictors: accuracy (ACC), Matthew's correlation coefficient (MCC), sensitivity (SEN), specificity (SPE), and F1. These metrics are calculated as follows:

$$ACC = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \quad (15)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\,(TN+FP)\,(TP+FP)\,(TN+FN)}} \quad (16)$$

$$SEN = \frac{TP}{TP+FN} \times 100\% \quad (17)$$

$$SPE = \frac{TN}{TN+FP} \times 100\% \quad (18)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (19)$$

where TP is the number of true positive samples; TN is the number of true negative samples; FP is the number of false positive samples; FN is the number of false negative samples.

Additionally, the receiver operating characteristic (ROC) curve is also an important measurement to evaluate the performance of classifiers, and the area under receiver operating characteristic curve (AUC) is the quantitative indicator. High values of AUC indicate better performance of predictors.

RESULTS

We redivided the four datasets introduced in section "Datasets" into the training sets and test sets with the ratio of 4:1, respectively. The training datasets were used to train models and the test datasets were utilized to evaluate model performance.

TABLE 1 | The performance of m6AGE against other existing predictors.

Datasets	Predictors	Metrics				
		ACC (%)	MCC	SEN (%)	SPE (%)	AUC
A101	m6AGE	**89.11**	**0.7822**	90.49	**87.68**	**0.9500**
	M6A-HPCS	86.43	0.7286	86.64	86.22	0.9284
	Targetm6A	87.36	0.7471	87.65	87.06	0.9358
	RAM-NPPS	83.86	0.6777	86.44	81.21	0.9077
	M6APred-EL	86.02	0.7205	85.63	86.43	0.9055
	DeepM6ASeq	87.77	0.7595	**93.32**	82.05	0.9419
A25	m6AGE	**87.97**	**0.7708**	74.65	98.85	**0.8867**
	M6A-HPCS	68.35	0.3577	61.97	73.56	0.7238
	Targetm6A	82.91	0.6542	76.06	88.51	0.8370
	RAM-NPPS	82.91	0.6538	**77.46**	87.36	0.8621
	M6APred-EL	87.34	0.7642	71.83	**100.00**	0.8464
	DeepM6ASeq	77.85	0.5515	67.61	86.21	0.8054
H41	m6AGE	**90.93**	**0.8325**	**81.94**	**100.00**	0.9181
	M6A-HPCS	71.46	0.4336	64.76	78.22	0.7765
	Targetm6A	90.49	0.8249	81.06	**100.00**	**0.9205**
	RAM-NPPS	90.49	0.8249	81.06	**100.00**	0.9051
	M6APred-EL	89.82	0.8136	79.74	**100.00**	0.9132
	DeepM6ASeq	86.50	0.7566	73.57	99.56	0.9051

The optimal value of each evaluation metric is marked in bold.

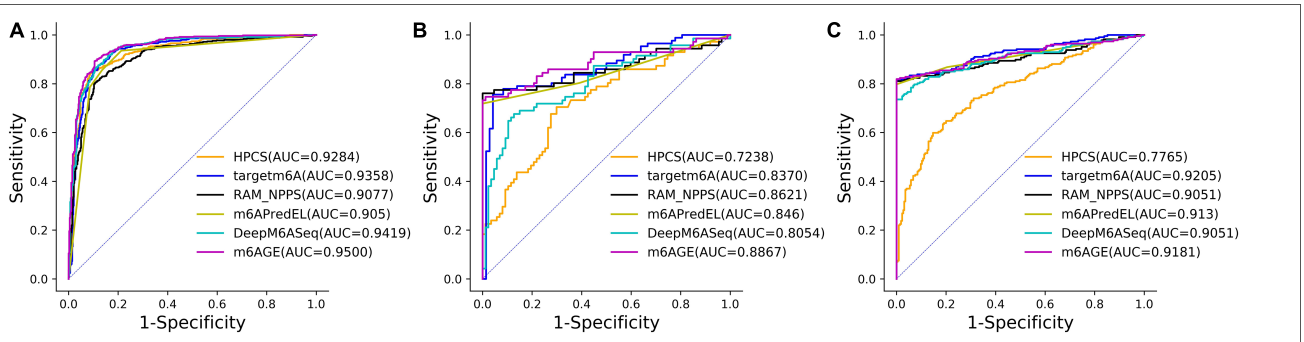


FIGURE 2 | The ROC curves of m6AGE and comparing predictors on three datasets. **(A)** The ROC curves on the *A101* dataset. **(B)** The ROC curves on the *A25* dataset. **(C)** The ROC curves on the *H41* dataset.

Due to the difference between datasets, we selected suitable sequence-derived features for each dataset. For *A101*, the PseKNC, CTD, and NPS features were selected; For *A25*, the EIIP, NPPS, NPS, PseKNC, and NCP-ND were selected; For *S21*, the NPPS and NCP-ND features were selected; For *H41*, the NCP-ND, PseKNC, and NPPS features were selected.

Comparison With Existing Predictors

In this section, we compared the performance of our predictor m6AGE with several other state-of-the-art predictors, including M6A-HPCS (Zhang et al., 2016), Targetm6A(Li et al., 2016), RAM-NPPS (Xing et al., 2017), M6APred-EL (Wei et al., 2018), and DeepM6ASeq (Zhang and Hamada, 2018). M6A-HPCS uses PseDNC and DACC features and a support vector machine (SVM) classifier to identify m^{6}A sites. Targetm6A utilizes position-specific kmer propensities (PSKP) feature and SVM classifier. RAM-NPPS uses the NPPS feature and SVM classifier to identify m^{6}A sites. M6APred-EL creates an ensemble model with PseKNC, PSKP, and NCP-ND features. DeepM6ASeq develops a deep learning framework and uses one-hot encoding for the identification of m^{6}A sites. The predictor M6A-HPCS, M6APred-EL, Targetm6A, and RAM-NPPS were reproduced faithfully, and their parameters were optimized by grid search with five-fold cross-validation. All predictors were trained and evaluated on the same dataset for fairness of comparison.

The evaluation results were summarized in **Table 1**. We employed ACC, MCC, SEN, SPE, and AUC as evaluation metrics, and compared the evaluation metrics of m6AGE with five other predictors on three datasets: *A101*, *A25*, and *H41*. As shown in **Table 1,** our predictor m6AGE achieved all optimal values on three datasets, except for SEN and SPE on the *A25* dataset, and AUC on the *H41* dataset.

TABLE 2 | The performance of different predictors on *S21* dataset.

Predictors	Metrics				
	SEN (%)	SPE (%)	F1	MCC	AUC
m6AGE	68.68	83.02	**0.5723**	**0.4593**	**0.8103**
HPCS	71.70	46.63	0.3622	0.1459	0.6330
Targetm6A	70.57	76.73	0.5260	0.3984	0.7818
RAM-NPPS	66.42	81.49	0.5440	0.4218	0.7778
M6APred-EL	**78.59**	75.20	0.5554	0.4433	0.7899
DeepM6ASeq	63.77	**83.38**	0.5460	0.4253	0.8056

The optimal value of each evaluation metric is marked in bold.

On the *A101* dataset, m6AGE obtained the optimal ACC, MCC, SPE, and AUC with 89.11%, 0.7822, 87.68%, and 0.9500, which is 1.34%, 0.0227, 5.63%, and 0.0081 higher than the suboptimal predictor DeepM6ASeq, respectively.

On the *A25* dataset, m6AGE obtained the optimal ACC, MCC, and AUC with 87.97%, 0.7708, and 0.8867. Its Acc and MCC is 0.63% and 0.0066 higher than the suboptimal value of predictor M6APred-EL. Its AUC is 0.0246 higher than the suboptimal value of predictor RAM-NPPS.

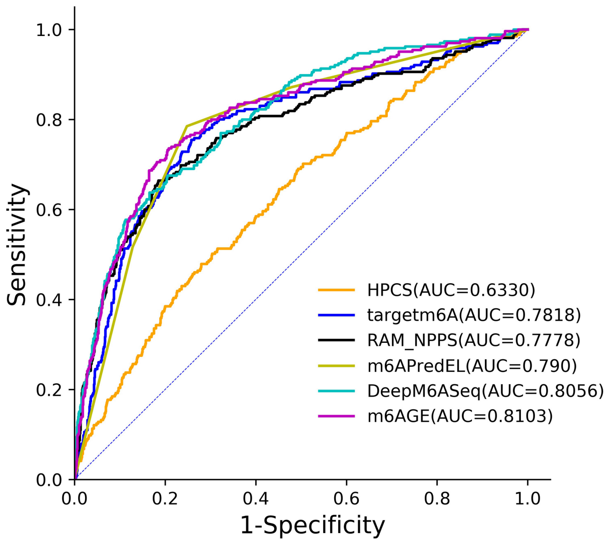


FIGURE 3 | The ROC curves of m6AGE and comparing predictors on the *S21* datasets.

On the *H41* dataset, m6AGE obtained the optimal ACC, MCC, SEN, and SPE with 90.93%, 0.8325, 81.94%, and 100%, which is 0.44%, 0.0076, 0.88%, and 0 higher than the predictor Targetm6A and RAM-NPPS, respectively.

The ROC curves of these predictors on three datasets were plotted in **Figure 2**. As shown in **Figure 2**, our predictor outperformed other predictors on the *A101* and *A25* datasets. Although the AUC of m6AGE on dataset *H41* is lower than other predictors, m6AGE achieved the optimal value of ACC, MCC, SEN, and SPE. These evaluation results demonstrate that our predictor m6AGE is superior to other predictors in terms of these three datasets.

Performance on Imbalanced Dataset

The non-m^{6}a sites on mRNA are much more than m^{6}A sites, so testing the performance of our predictor on imbalanced datasets is of great importance. The imbalance ratio of the *S21* dataset is about 1:4. We redivided the *S21* dataset, and randomly selected 80% samples as the training set, and the remaining 20% samples as the test set.

CatBoost solves the imbalance data issues by setting weights for each class or sample. The weight of each class is generally inversely proportional to the number of its samples. The metrics F1 and MCC are usually used as the evaluation criteria for imbalanced datasets (Zhao et al., 2018; Wang et al., 2019; Dou et al., 2020). We compared m6AGE with five other predictors on the *S21* dataset.

The evaluation results were summarized in **Table 2**. The optimal value of each evaluation metric is marked in bold. As shown in **Table 2,** our predictor m6AGE got the optimal values of F1, MCC, and AUC with 0.5723, 0.4593, and 0.8103.

The ROC curves of these predictors on the *S21* dataset were plotted in **Figure 3**. As shown in **Figure 3**, our predictor outperformed other predictors on the *S21* dataset.

TABLE 3 | The performance of different classifiers.

Datasets	Classifiers	Metrics				
		ACC (%)	MCC	SEN (%)	SPE (%)	AUC
A101	CatBoost	**89.11**	**0.7822**	**90.49**	87.68	**0.9500**
	Random forest	87.67	0.7534	87.04	88.31	0.9377
	Logistic regression	89.00	0.7800	89.07	**88.94**	0.9489
	Decision tree	80.99	0.6197	82.39	79.54	0.8096
A25	CatBoost	**87.97**	**0.7708**	74.65	98.85	**0.8867**
	Random forest	87.34	0.7642	71.83	**100.00**	0.8729
	Logistic regression	79.11	0.5767	74.65	82.76	0.8562
	Decision tree	81.65	0.6349	**84.51**	79.31	0.8191
H41	CatBoost	**90.93**	**0.8325**	81.94	**100.00**	**0.9181**
	random forest	89.38	0.8031	79.74	99.11	0.9098
	Logistic regression	86.95	0.7422	82.38	91.56	0.9125
	Decision tree	86.28	0.7258	**85.46**	87.11	0.8629

The optimal value of each evaluation metric is marked in bold.

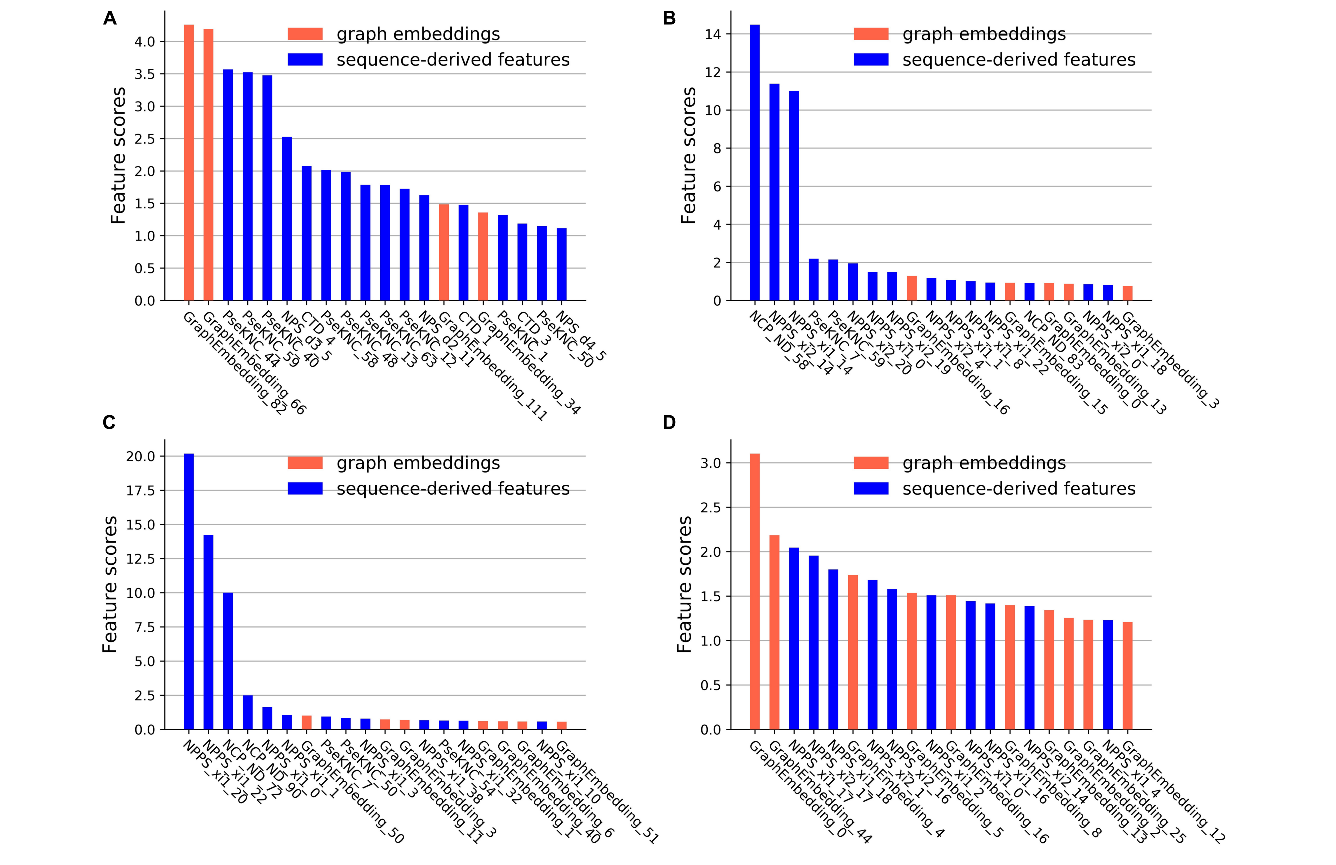


FIGURE 4 | The feature importance scores on the four datasets. **(A)** The feature importance scores on the *A101* datasets. **(B)** The feature importance scores on the *A25* datasets. **(C)** The feature importance scores on the *H41* datasets. **(D)** The feature importance scores on the *S21* datasets.

Comparison With Different Classifiers

To further demonstrate the effectiveness of CatBoost, we compared it with other popular classifiers, including Random Forest, Logistic Regression, and Decision Tree, which are commonly and widely used in bioinformatics classification tasks. All classifiers were trained and assessed under the same conditions for a fair comparison.

The prediction results were summarized in **Table 3**. We compared the prediction results with three other classifiers on the *A101*, *A25*, and *H41* dataset. The evaluation metrics used are

ACC, MCC, SEN, SPE, and AUC. As shown in **Table 3**, CatBoost achieved all optimal metrics on three datasets, except for SPE on the *A101* dataset and SEN on the *A25* and *H41* dataset.

Feature Importance Analysis

CatBoost can output the scores of feature importance, which reflect the contributions of the features in specific feature space for identifying m^6A sites. The first 20 important features and their scores on the four datasets were plotted in **Figure 4**.

On the *A101* dataset, the first three important sequence-derived features are "PseKNC_44", "PseKNC_59", and "PseKNC_40", which correspond to the occurrence frequency of "GUA", "UGU", and PseKNC_40 respectively, On the *A25* dataset, the first three important sequence-derived features are "NCP_ND_58", "NPPS_xi2_14", and "NPPS_xi1_14," which correspond to the position +1 (Assuming that the position of m^6A site is 0), +2 and +4, +2 and +3, respectively; On the *H41* dataset, the first three important sequence-derived features are "NPPS_xi1_20", "NPPS_xi1_22", and "NCP_ND_72", which correspond to the position 0 and +1, +2 and +3, −3, respectively; On the *S21* dataset, the first three important sequence-derived features are "NPPS_xi1_17", "NPPS_xi2_17", and "NPPS_xi1_18," which correspond to the position +6 and +7, +6 and +8, +7 and +9, respectively.

In addition, graph embeddings account for 20%, 25%, 35%, and 50% of the top 20 important features in the four datasets, respectively, which indicates that graph embeddings could supplement the information of the sequence-derived features.

DISCUSSION

The methods for extracting sequence features are indispensable for building a reliable predictor. Contributing sequence features, such as the physical and chemical properties of nucleotides, the frequency of k-nucleotides, and the frequency of specific positions, can fully reflect the information related to the m^6A site recognition. In this study, we integrated and selected suitable sequence-derived features for each dataset. However, most of the feature encoding methods are based on the primary sequence, and only a few of them calculate the frequency of nucleotides in the training dataset, so it is difficult to obtain more helpful information from the whole dataset. This paper innovatively introduces a feature extraction method based on the graph embedding methods as a supplement to sequence-derived features. First of all, a network is constructed based on the whole dataset and sequence-derived features. Samples are abstracted as nodes of the network, and the similarity relationships between samples are abstracted as edges. This network reflects global information of the whole dataset. Then, graph embedding (neighborhood-based node embedding) methods are used to learn the feature representation of each node in an unsupervised manner. The graph embedding features of samples contain the related information with other samples. Finally, we integrate sequence-derived features and graph embeddings based with the feature fusion strategy. Therefore, the final input features can reflect the information of samples more comprehensively.

It is also significant to choose an appropriate classifier. CatBoost is a GBDT algorithm, which shows excellent performance in many classification tasks. Because of its good effect of restraining overfitting and fast running, the CatBoost algorithm is selected to train our predictor m6AGE.

To further prove the effectiveness of our predictor, we compare the evaluation results with that of other existing m^6A site predictors. The results show that our predictor m6AGE outperforms other existing methods. In the future, we will apply m6AGE to more m^6A site datasets and seek more suitable graph embedding methods. It is worth mentioning that the computational framework proposed in this study is possible to extend to other bioinformatics site identification tasks.

The source code of m6AGE is available at https://github.com/bokunoBike/m6AGE. Users can download and run it on the local machines. The data is imported through the file paths of the positive training set, negative training set, and test set. Then m6AGE is trained and generates prediction results. Note that the corresponding python packages need to be installed first (see GitHub page for details). For a new dataset, our predictor will automatically select the appropriate sequence-derived features (or specified by the users in the corresponding configuration file) according to the feature importance scores.

CONCLUSION

The identification of N^6-methyladenosine (m^6A) modification sites on RNA is of biological significance. In this study, a novel computational framework called "m6AGE" is proposed to predict and identify the m^6A sites on mRNA. Our predictor combines sequence-derived features with the features extracted by graph embedding methods. The context information of sites is directly extracted from primary sequences by the sequence-derived features, and the global information is extracted by the graph embeddings. Experiments showed that the proposed m6AGE achieved successful prediction performance on four datasets across three species. It could be expected that m6AGE would be a powerful computational tool for predicting and identifying the m^6A modification sites on mRNA.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. Codes and data are available here: https://github.com/bokunoBike/m6AGE which contains detailed steps to run m6AGE.

AUTHOR CONTRIBUTIONS

YW and RG conceived the algorithm and developed the program. RG, YW, and SY wrote the manuscript and prepared the datasets. YW and SY helped with manuscript editing, design. XMH, LH, and KH helped to revise the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 62072212), the Development Project of Jilin Province of China (Nos. 20200401083GX, 2020C003), and Guangdong Key Project for Applied Fundamental Research (No.2018KZDXM076). This work was also supported by Jilin Province Key Laboratory of Big Data Intelligent Computing (No. 20180622002JC).

REFERENCES

Cao, S., Lu, W., and Xu, Q. (2015). "GraRep: learning graph representations with global structural information," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, (New York, NY:Association for Computing Machinery), 891–900. doi: 10.1145/2806416.2806512

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY:Association for Computing Machinery), 785–794. doi: 10.1145/2939672.2939785

Chen, K., Lu, Z., Wang, X., Fu, Y., Luo, G.-Z., Liu, N., et al. (2015a). High-Resolution N 6 -methyladenosine (m 6 A) map using photo-crosslinking-assisted m 6 a sequencing. *Angew. Chemie*127, 1607–1610.doi: 10.1002/ange.201410647

Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K. C. (2015b). IRNA-Methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*490, 26–33. doi: 10.1016/j.ab.2015.08.021

Chen, W., Feng, P., Ding, H., and Lin, H. (2016). Identifying N 6-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Mol. Genet. Genomics*291, 2225–2229.doi: 10.1007/s00438-016-1243-7

Chen, W., Tang, H., and Lin, H. (2017). MethyRNA : a web server for identification of N–methyladenosine sites. *J. Biomol. Struct. Dyn.*1102, 1–5. doi: 10.1080/07391102.2016.1157761

Chen, W., Tran, H., Liang, Z., Lin, H., and Zhang, L. (2015c). Identification and analysis of the N6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci. Rep.*5:13859. doi: 10.1038/srep13859

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.*47:e41. doi: 10.1093/nar/gkz074

Chen, Z., Zhao, P., Li, F., Wang, Y., Smith, A. I., Webb, G. I., et al. (2020). Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief. Bioinform.*21, 1676–1696. doi: 10.1093/bib/bbz112

Desrosiers, R., Friderici, K., and Rottman, F. (1974). Identification of methylated nucleosides in messenger RNA from novikoff hepatoma cells. *Proc. Natl. Acad. Sci. U.S.A.*71, 3971L–3975. doi: 10.1073/pnas.71.10.3971

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*485, 201–206. doi: 10.1038/nature11112

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv* [Preprint]arXiv:1810.11363,

Dou, L., Li, X., Ding, H., Xu, L., and Xiang, H. (2020). IRNA-m5C_NB: a novel predictor to identify RNA 5-methylcytosine sites based on the naive bayes classifier. *IEEE Access*8, 84906–84917. doi: 10.1109/ACCESS.2020.2991477

Golam Bari, A. T. M., Reaz, M. R., Choi, H. J., and Jeong, B. S. (2013). "DNA encoding for splice site prediction in large DNA sequence," in *Database Systems for Advanced Applications.DASFAA 2013. Lecture Notes in Computer Science*, Vol. 7827, eds B.Hong, X.Meng, L.Chen, W.Winiwarter, and W.Song (Berlin: Springer), doi: 10.1007/978-3-642-40270-8_4

Grover, A., and Leskovec, J. (2016). "Node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY:Association for Computing Machinery), 855–864. doi: 10.1145/2939672.2939754

Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., Lin, H., Chen, W., et al. (2014). INuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*30, 1522–1529. doi: 10.1093/bioinformatics/btu083

Huang, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2018). BERMP: a cross-species classifier for predicting m 6 a sites by integrating a deep learning algorithm and a random forest approach. *Int. J. Biol. Sci.*14, 1669–1677. doi: 10.7150/ijbs.27819

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*30, 3147–3155. doi: 10.1016/j.envres.2020.110363

Langley, P., and Sage, S. (1994). "Oblivious decision trees and abstract cases," in *Proceedings of the Working Notes of the AAAI-94 Workshop on Case-Based Reasoning*, (Seattle, WA:AAAI Press), 113–117.

Li, G. Q., Liu, Z., Shen, H. B., and Yu, D. J. (2016). TargetM6A: identifying N6-Methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. *IEEE Trans. Nanobiosci.*15, 674–682. doi: 10.1109/TNB.2016.2599115

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., and Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods*12, 767–772. doi: 10.1038/nmeth.3453

Liu, K., Cao, L., Du, P., and Chen, W. (2020). im6A-TS-CNN: identifying the N6-methyladenine site in multiple tissues by using the convolutional neural network. *Mol. Ther. Nucleic Acids*21, 1044–1049.doi: 10.1016/j.omtn.2020.07.034

Liu, K., and Chen, W. (2020). IMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. *Bioinformatics*36, 3336–3342. doi: 10.1093/bioinformatics/btaa155

Luo, G. Z., Macqueen, A., Zheng, G., Duan, H., Dore, L. C., Lu, Z., et al. (2014). Unique features of the m6A methylome in *Arabidopsis thaliana*. *Nat. Commun.*5:5630. doi: 10.1038/ncomms6630

Meyer, K. D., and Jaffrey, S. R. (2014). The dynamic epitranscriptome : N 6 -methyladenosine and gene expression control. 1974.*Nat. Rev. Mol. Cell Biol.*15, 313–326. doi: 10.1038/nrm3785

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Nair, A. S., and Sreenadhan, S. P. (2006). A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation*1, 197–202.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*103, 8577–8582. doi: 10.1073/pnas.0601602103

Nilsen, T. W. (2014). Internal mRNA methylation finally finds functions stirring the simmering. *Science*343, 1207–1208. doi: 10.1126/science.1249340

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). "CatBoost: Unbiased Boosting with Categorical Features', in *Proceedings of the 32nd International Conference on Neural Information Processing Systems NIPS'18*. (Red Hook, NY, Unites States: Curran Associates Inc.), 6639–6649. Availble online at: https://nips.cc/Conferences/2018

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., et al. (2013). High-Resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*155, 1409–1421. doi: 10.1016/j.cell.2013.10.047

Shao, J., Xu, D., Tsai, S. N., Wang, Y., and Ngai, S. M. (2009). Computational identification of protein methylation sites through Bi-profile Bayes feature extraction. *PLoS One*4:e4920. doi: 10.1371/journal.pone.0004920

Tang, L., and Liu, H. (2009). "Relational learning via latent social dimensions," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY:Association for Computing Machinery), 817–825. doi: 10.1145/1557019.1557109

Tong, X., and Liu, S. (2019). CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res.*47, e43–e43. doi: 10.1093/nar/gkz087

Wan, Y., Tang, K., Zhang, D., Xie, S., Zhu, X., Wang, Z., et al. (2015). Transcriptome-wide high-throughput deep m6A-seq reveals unique differential m6A methylation patterns between three organs in *Arabidopsis thaliana*. *Genome Biol.*16:272. doi: 10.1186/s13059-015-0839-2

Wang, H., Ma, Y., Dong, C., Li, C., Wang, J., and Liu, D. (2019). CL-PMI: a precursor microRNA identification method based on convolutional and long short-term memory networks. *Front. Genet.*10:967. doi: 10.3389/fgene.2019.00967

Wang, X., and Yan, R. (2018). RFAthM6A: a new tool for predicting m6A sites in *Arabidopsis thaliana*. *Plant Mol. Biol.*96, 327–337. doi: 10.1007/s11103-018-0698-9

Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids*12, 635–644.doi: 10.1016/j.omtn.2018.07.004

Xing, P., Su, R., Guo, F., and Wei, L. (2017). Identifying N6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Sci. Rep.*7:46757. doi: 10.1038/srep46757

Zhang, M., Sun, J. W., Liu, Z., Ren, M. W., Shen, H. B., and Yu, D. J. (2016). Improving N6-methyladenosine site prediction with heuristic selection of nucleotide physical–chemical properties. *Anal. Biochem.*508, 104–113. doi: 10.1016/j.ab.2016.06.001

Zhang, W., Chen, Y., Tu, S., Liu, F., and Qu, Q. (2017). "Drug side effect prediction through linear neighborhoods and multiple data source integration," in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Shenzhen: IEEE), 427–434. doi: 10.1109/BIBM.2016.7822555

Zhang, W., Tang, G., Wang, S., Chen, Y., Zhou, S., and Li, X. (2019). "Sequence-derived linear neighborhood propagation method for predicting lncRNA-miRNA interactions," in *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Vol. 1, (Madrid: IEEE), 50–55. doi: 10.1109/BIBM.2018.8621184

Zhang, Y., and Hamada, M. (2018). DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinformatics*19(Suppl. 19):524. doi: 10.1186/s12859-018-2516-4

Zhao, Z., Peng, H., Lan, C., Zheng, Y., Fang, L., and Li, J. (2018). Imbalance learning for the prediction of N6-Methylation sites in mRNAs. *BMC Genomics*19:574. doi: 10.1186/s12864-018-4928-y

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.*44:e91. doi: 10.1093/nar/gkw104

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N 6 -methyladenosine sites from mRNA. *RNA*25, 205–218. doi: 10.1261/rna.069112.118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 03 June 2021
doi: 10.3389/fgene.2021.670232

MI_DenseNetCAM: A Novel Pan-Cancer Classification and Prediction Method Based on Mutual Information and Deep Learning Model

Jianlin Wang[1], *Xuebing Dai*[1], *Huimin Luo*[1], *Chaokun Yan*[1]*, *Ge Zhang*[1] *and Junwei Luo*[2]*

[1] *School of Computer and Information Engineering, Henan University, Kaifeng, China,* [2] *College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China*

The Pan-Cancer Atlas consists of original sequencing data from various sources, provides the opportunity to perform systematic studies on the commonalities and differences between diverse cancers. The analysis for the pan-cancer dataset could help researchers to identify the key factors that could trigger cancer. In this paper, we present a novel pan-cancer classification method, referred to MI_DenseNetCAM, to identify a set of genes that can differentiate all tumor types accurately. First, the Mutual Information (MI) was utilized to eliminate noise and redundancy from the pan-cancer datasets. Then, the gene data was further converted to 2D images. Next, the DenseNet model was adopted as a classifier and the Guided Grad-CAM algorithm was applied to identify the key genes. Extensive experimental results on the public RNA-seq data sets with 33 different tumor types show that our method outperforms the other state-of-the-art classification methods. Moreover, gene analysis further demonstrated that the genes selected by our method were related to the corresponding tumor types.

Keywords: pan-cancer, cancer classification, DenseNet, guided grad-CAM algorithm, RNA-seq data

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Xiaoqing Peng,
Central South University, China
Juan Wang,
Inner Mongolia University, China

***Correspondence:**
Chaokun Yan
ckyan@henu.edu.cn
Junwei Luo
luojunwei@hpu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *20 February 2021*
Accepted: *10 May 2021*
Published: *03 June 2021*

Citation:
Wang J, Dai X, Luo H, Yan C, Zhang G and Luo J (2021) MI_DenseNetCAM: A Novel Pan-Cancer Classification and Prediction Method Based on Mutual Information and Deep Learning Model. Front. Genet. 12:670232. doi: 10.3389/fgene.2021.670232

1. INTRODUCTION

Cancer, known as the "the king of the diseases," is a serious threat to human health. In 2020, 1,806,590 new cancer cases and 606,520 cancer deaths are projected to occur in the United States (Siegel et al., 2020). Cancer accurate prediction in the early stage is a challenging subject that has drawn worldwide concern due to the high morbidity and mortality of cancer (Kourou et al., 2015). However, the existing medical equipment and clinical symptoms are not sensitive to the changes at the molecular level, and it is difficult to make early diagnosis for potential patients. Some potential patients cancer may be advanced when they are first diagnosed (Sakri et al., 2018), resulting in increased mortality from cancer. If cancer can be detected early and treated appropriately, the survival time of patients will be greatly increased. Therefore, identifying a set of genes that can characterize the type and stage of cancer is the key to effective treatment. These genes may serve as biomarkers to efficiently diagnose diseases and accurately classify cancer types. Furthermore, since The Cancer Genome Atlas (TCGA) project was launched, TCGA project has so far generated a pan-cancer atlas of 33 types of cancer. Therefore, extensive studies about pan-cancer have been researched, among which pan-cancer classification is an important perspective.

In recent years, advances in sequencing technology have led to a significant decrease in the cost of accumulating biological data. A large amount of biological data laid an important foundation for researchers to identify some key cancer biomarkers and enable accurate cancer classification prediction in the early stage. However, the tough challenges also come from the characteristics of these data (i.e., high dimensionality, severely limited samples and containing a large portion of irrelevant genes), which hinders the rapid and accurate cancer classification and prediction (Saeys et al., 2007). In order to solve this problem, feature selection techniques can be applied to analyze the possible cancer-causing genes from massive cancer gene data. The feature selection aims to represent high-dimensional data with fewer features while improves the prediction accuracy of classification models. In general, feature selection can be categorized into two types: filter methods, wrapper methods (Huang et al., 2007). Usually, filter methods have much less computational complexity compared with other methods. Some filter methods, such as MI, IG (Martín-Valdivia et al., 2008), Relief (Urbanowicz et al., 2018), have been applied to data analysis for gene expression data well.

However, most traditional tumor classification studies only focus on the same tumor type, the heterogeneity among different tumor types is usually neglected (Lawrence et al., 2013; Lyu and Haque, 2018). Tumor heterogeneity is reflected in the obvious differences between different tumor cells at the molecular level of genomic, transcriptomic, proteome and so on. Therefore, in order to understand and capture the commonalities and differences between diverse cancers, TCGA later launched the Pan-Cancer analysis project (Weinstein et al., 2013). Pan-cancer analysis is a study that integrates multiple tumor types. In recent years, the research and analysis of pan-cancer have been increasing gradually, and people hope to find the genes related to tumors so as to accurately predict the type of cancer. It has been suggested that specifications of therapies according to tumor types differentiated may maximize the efficacy of the patients (Golub et al., 1999; Alizadeh et al., 2000; Van't Veer et al., 2002). At present, there have been many studies (Kourou et al., 2015; Li et al., 2017) using machine learning (ML) algorithms to analyze pan-cancer data sets and demonstrate its effectiveness in cancer classification and prediction. For example, Li et al. proposed a GA/KNN method to classify 9,096 samples from 31 different tumor types and obtained a set of genes that could correctly classify 90% of the samples. Deep learning has made unprecedented breakthroughs in various classification tasks recently and has been widely applied due to its excellent classification performance. A strength of deep learning is its ability to learn end to end, automatically discovering multiple levels of representation to achieve a prediction task (Wainberg et al., 2018).

In the study, a deep learning approach, MI_DenseNetCAM was proposed to classify 33 different types of tumors based on high-dimensional RNA-Seq gene expression data. Then, the Guided grad-CAM algorithm was used to identify the key genes that played an important role in the classification process. We evaluated the method with performance metrics such as recall, precision and F1 score, and the results demonstrate that the proposed method takes full advantage of the information in the pan-cancer data sets and achieved an overall test accuracy of 96.81%. Compared with the existing methods, our proposed method provides superior performance in the classification accuracy of 33 tumor types. The main contributions of this paper can be summarized as follows:

- For the noise and redundancy of the pan-cancer data sets, the Min-Max normalization and MI was adopted to preprocess the data, which can screen out the highly correlated genes to improve the performance of the classification model. Moreover, we evaluated the impact of different data preprocessing strategies on the classification performance.
- For the pan-cancer data set, the DenseNet model was utilized as a classifier to classify and predict tumor types. Compared with other classifiers, the DenseNet model achieved better performance whilst requiring fewer parameters and computation cost.
- Extensive experiments and analyses have been carried out on the pan-cancer data set in terms of evaluation indicators, and the experimental results demonstrate that our proposed method is very promising. Some of the genes identified by our method have already been verified.

The remainder of this paper is organized as follows: In section 2, we review related works. In section 3, the detailed implementation of the proposed pan-cancer classification method is elaborated. We described the experimental results and analysis in section 4. Finally, we summarize the paper and discuss the future works in section 5.

2. RELATED WORK

The goal of the pan-cancer analysis was to assemble data from the separate disease projects to build a data set spanning multiple tumor types (Weinstein et al., 2013). Through the analysis and interpretation of these data to find the commonalities and differences across various tumor types. At present, many machine learning and deep learning methods have been applied to the analysis of pan-cancer data. Next, we conduct a review of the latest studies in the field of pan-cancer analysis.

Hsu and Si (2018) focused on using machine learning (ML) to build a reliable classification model which can recognize 33 types of cancer patients. They applied five ML algorithms, namely decision tree (DT), k nearest neighbor (kNN), linear support vector machine (linear SVM), polynomial support vector machine (ploy SVM), and artificial neural network (ANN) to analyze the data set of pan-cancer. The results show that linear SVM with a 95.8% accuracy rate is the best classifier among the five classification algorithms.

Kang et al. (2019) proposed a new method for the classification of multiple tumor types by using relaxed Lasso selection feature subsets and an improved support vector machine (GenSVM) as the classifier. GenSVM is a general multiclass support vector machine, which compared with the other three classifiers (KNN, L_1logreg, L_2logreg) on the four multi-class datasets, the experimental results showed that GenSVM has better generality, flexibility and achieve higher classification accuracy with fewer features in multi-classification problems.

Li et al. (2017) undertook the development of a pan-cancer atlas to recognize 9,096 TCGA tumor samples representing 31 tumor types. They applied k-nearest neighbors (KNN) to classify 31 different types of tumor, and embedded genetic algorithm to improve the accuracy of the KNN classifier. This method achieved an accuracy of 90% across 31 tumor types.

In recent years, the deep learning (DL) method was also used to classify and identify cancer types. In paper (Danaee et al., 2017) the author used a stacked auto-encoder first to extract high-level features from the expression values and then input these features into a single layer ANN network to decide whether the sample is a tumor or not. The accuracy of using such a method reached 94%. However, as to the multi-classification problem, because this method has more complicated network structure and parameter setting, in order to save time cost, the author only conducted the experiments on breast cancer.

Khalifa et al. (2020) introduced a novel optimized deep learning approach based on binary particle swarm optimization with decision tree (BPSO-DT) and CNN to classify different types of tumor. The results showed that the proposed method achieved an overall testing accuracy of 96.6%. However, they classified only five different tumor types (KIRC, BRCA, LUSC, LUAD, and UCEC), and did not analyze all the pan-cancer data sets.

Lyu and Haque (2018) designed a new method that embedded the high dimensional RNA-Seq data into 2-D images and used a CNN to make classification of the 33 tumor types. This method achieved 95.59% accuracy for all 33 tumor types. However, the method proposed by Lyu et al. failed to achieve good classification performance on tumor datasets with small samples, which increases the risk of overfitting.

3. MATERIALS AND METHODS

In this section, a novel framework for the classification of pan-cancer, referred to MI_DenseNetCAM has been proposed. First, we preprocess the original data set, and then embed the data into a 2-D image. Then, we train a DenseNet model with the generated images. Next, the trained model and Guided Grad-Cam algorithm are applied to generate the heat map. Furthermore, some important genes can be obtained. The workflow of the proposed method is shown in **Figure 1A**.

3.1. Datasets

We conducted experiments to evaluate the proposed method on the RNA-seq data sets of 33 types of cancers. RNA-seq, also known as transcriptomic sequencing, can accurately analyze gene expression differences and gene structure variations, and reveal specific biological processes and molecular mechanisms in the process of disease occurrence. Therefore, we use the normalized-level3 RNA-seq gene expression data to construct our experiment dataset. The datasets are available for download from http://gdac.broadinstitute.org/. These data sets, which contain 33 different tumor types. The data for each type of tumor is high-dimensional, with 20,531 columns. **Table 3** gives a detailed description of the number of samples and genes in these datasets.

3.2. Data Preprocessing

Firstly, data from 33 different tumor types are collected and integrated, and then the genes in the data set are compared with the annotation files (downloaded from NCBI), so as to screen out the genes that did not exist in the annotation files. About 1,000 genes were not found in the annotation file, therefore, these 1,000 genes and corresponding expression levels need to be removed from the data set. Secondly, genes are ordered based on the chromosome number because adjacent genes are more likely to interact with each other. Thirdly, the data set is normalized by Min-Max normalization, which scales the data to a small interval, thus leads to get the solution quickly. The Min-Max normalization is defined by Equation (1).

$$y = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where X represents a column of data in the pan-cancer data set, X_{min} and X_{max} represent the minimum and maximum values in a column of data.

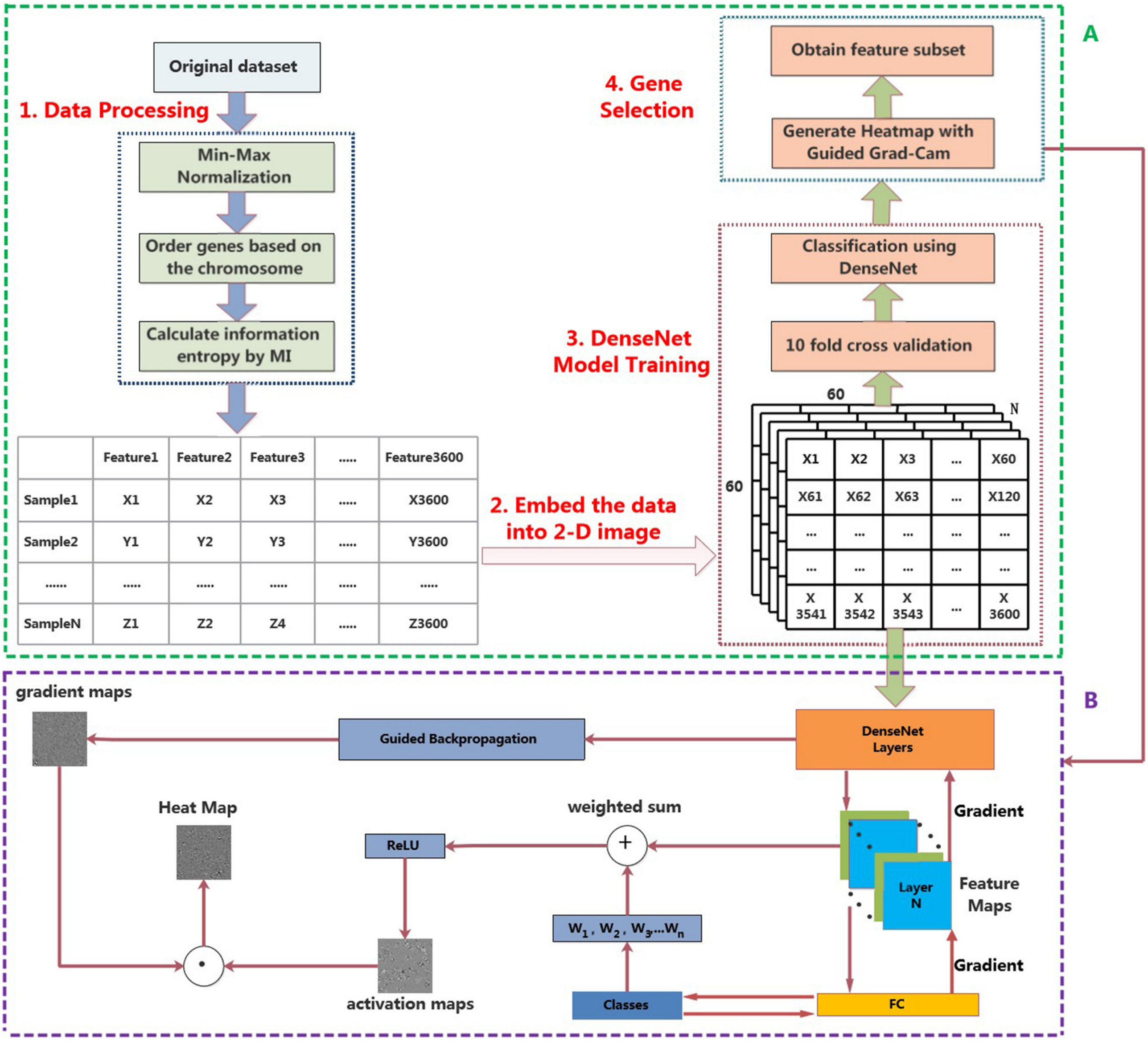
After normalization of gene data, we further adopted Mutual Information (MI) to calculate the correlation between the gene and the label to decide whether to select the gene. MI is a feature ranking approach based on information entropy (Kraskov et al., 2004; Martín-Valdivia et al., 2008). In the domain of feature selection, Sharmin et al. (2019) used MI as a metric to measure the degree of correlation between features and category labels. The more mutual information between the two, the more important this feature is. The mutual information between two random variables *X* and *Y* is as follows:

$$I(X, Y) = \sum_{x,y} P(x, y) log \frac{P(x, y)}{P(x)P(y)} \tag{2}$$

Where, *P(x,y)* represents the joint probabilistic mass function, *P(x)* and *P(y)* represent edge probability density functions. The closer the relationship between *X* and *Y* is, the greater the value of *I(X, Y)* will be. If the two variables are independent, the value of *I(X, Y)* is 0.

When mutual information is applied to feature selection, then random variable *X* represents the feature and random variable *Y* represents the label, the value of *I(X, Y)* represents the correlation between the ith feature and the label. The greater the value, the greater the correlation between the feature and the label, and vice versa. Therefore, we can sort features in terms of the information entropy by MI method and select important features.

After mutual information, the number of genes was further reduced to N. Through the subsequent experiment of different N values, N is set to 3,600. Then, convert the data corresponding to the selected important features into an image format. The data from each sample are successively put into each pixel of the image in order to reconstruct the data from a 1-D array into a 2-D image. In other words, the array with the shape of 3600*1 is turned into a two-dimensional image with the shape of 60*60, and the data needs to be normalized to [0,255]. The result of this step is to generate images that correspond to the samples in the dataset. The resulting 2-D images will be used to train the DenseNet model.



	Feature1	Feature2	Feature3	.....	Feature3600
Sample1	X1	X2	X3	.....	X3600
Sample2	Y1	Y2	Y3	.....	Y3600
......	.....	.....	.....	.....	.....
SampleN	Z1	Z2	Z4	.....	Z3600

X1	X2	X3	...	X60
X61	X62	X63	...	X120
...	...	...	...	...
...	...	...	...	...
X 3541	X 3542	X 3543	...	X 3600

FIGURE 1 | The workflow of MI_DenseNetCAM. **(A)** Cancer classification and prediction through MI and deep learning combined analysis from pan-cancer datasets. **(B)** Principle diagram of the Guided Grad-Cam algorithm.

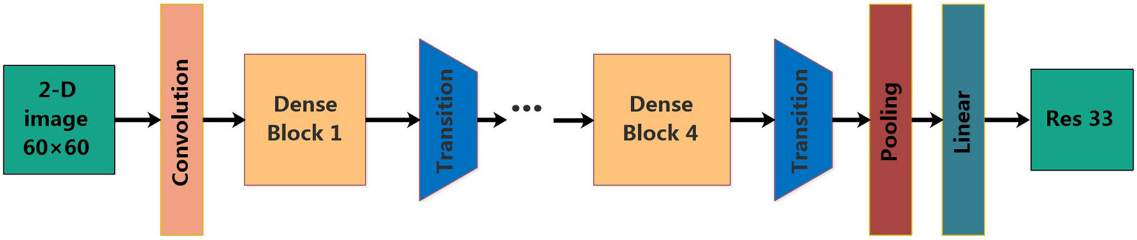


FIGURE 2 | The structure of the DenseNet.

3.3. Model Training

Deep neural models based on Convolutional Neural Network (CNN) have enabled unprecedented breakthroughs in a variety of image classification tasks, some famous architectures such as Resnet (He et al., 2016a) and inception (Szegedy et al., 2015) have excellent performance. In the Imagenet (Deng et al., 2009) challenge, CNN achieved a significant classification accuracy margin over classical machine learning methods. However, with the increase of layers, the traditional neural network will encounter a series of problems, such as gradient vanishing, feature reuse decreasing, parameter number increasing significantly, longer training time and classification accuracy decreasing (He et al., 2016b). In order to solve these problems, Huang et al. proposed a new method, DenseNet (Huang et al., 2017), which is a convolutional neural network with dense connections. Dense Net connects all layers directly to each other to ensure the maximum information flow between each layer in the network, in other words, the input of any next layer in the network is the superposition of the output of all previous layers. In this way, each layer can access the gradient directly from the loss function and the original input signal, yielding models that are easy to train and highly parameter efficient. Further, the dense connections have a regularizing effect, which reduces the risk of overfitting for small sample training tasks (Huang et al., 2017). The structure of the DenseNet is as shown in **Figure 2**.

The DenseNet model consists of four Dense Blocks, and each Dense Block is composed of batch normalization layer (BN) + ReLU + 1 × 1 convolutional layer (Conv 1×1) + BN + ReLU + Conv 3 × 3. The layers between two adjacent blocks are referred to as transition layers, which are composed of BN + ReLU + Conv 1 × 1 + Average Pooling 2 × 2. Pooling denotes the global average pool and Linear denotes the fully connected layer.

The optimizer plays an extremely significant role in deep learning training. It is used to update the weight parameters in the training process, which is related to whether the training can converge quickly and achieve high accuracy. In this paper, we use the Adam optimization algorithm. Compared with the traditional stochastic gradient descent algorithm, the advantage of the Adam algorithm is that it can design independent adaptive learning rates for different parameters, so as to obtain a higher training effect. For the classification task, cross entropy is generally used as the loss function. Moreover, In order to get a better training effect and ensure the robustness of the classification, make full use of the generated 2-D images and obtain reliable and stable models, we use 10-fold cross validation to evaluate the quality of the model during the training of DenseNet.

3.4. Screen Out Important Genes

After the DenseNet model is trained, the important genes can be obtained through two stages. First, the Guided Grad-Cam algorithm can be applied to generate heat maps, it can locate the regions related to categories in the image, indicating why the convolutional neural network is classified in this way. Then, match the high-intensity pixels in the heat map with the gene names in the original data set to obtain the important genes that contribute more to the classification.

The Guided Grad-Cam algorithm provides a technique for visual interpretation of how the convolutional neural network model makes decisions. The detailed procedure for generating heat map through the Guided Grad-Cam algorithm is as follows.

- Step1 Obtain Gradient maps
  First, the Guided backpropagation algorithm is used to calculate the gradient of the convolutional layer's feature value relative to the input layer, so as to obtain the feature gradient maps.
- Step2 Obtain Activation maps
  After feature extraction of the original image through the convolutional layer and the pooling layer, the convolutional neural network output a set of feature maps. A pixel in the feature map corresponds to a region in the original image. If the product of pixel value and weight in the feature map is >0, CNN believes that this region in the original image has features related to categories. The Guided Grad-Cam algorithm calculates the average gradient of each feature map relative to the classification probability to obtain a set of weights. After calculating the weights of all the feature maps, the weighted sum with the feature maps can be used to obtain the activation maps. Finally, the activation maps are processed using the ReLU activation function, retaining only the features of the activation maps that are useful for the category. If you do not add the ReLU activation function, you will bring in pixels belonging to other categories, which will affect the interpretation.
- Step3 Obtain Heat maps
  Superposition the gradient maps and activation maps to obtain the heat map for visualization of the convolutional neural network. The heat map shows the extent to which the pixel at the corresponding position in the original image affects the classification result. The overall structure of the Guided Grad-Cam algorithm is shown in **Figure 1B**.

Based on the DenseNet model and the Guided Grad-Cam algorithm, we can obtain heat maps with high resolution and category discriminability for displaying the importance of genes. Since the Guided Grad-Cam algorithm generates a heat map for each input image, in order to avoid the influence of noise on the experimental results, we averaged all the heat maps. In addition, the intensity of the pixel value in the heat map represents the influence of the pixel at the corresponding position in the original image on the classification result, so the gene corresponding to the position with the largest pixel value in the heat map is an important feature. In other words, the higher the pixel value and the higher the intensity in the heat map, the greater the contribution of these pixels to the final classification, that is their existence affects the classification most. So, important genes can be realized by looking for points with high pixel intensity in the heat map. The specific methods to achieve the following.

When the gene expression data is converted into 2D images, the corresponding expression values of each gene are sequentially mapped to the pixels in the images, as shown in **Figure 1A**. In order to screen out the important genes, firstly, we can convert the pixel value in the heat map into a 1D array based on their original order. Then, find out the index corresponding to the maximum pixel value, the corresponding gene name can be found from the original data according to the index value.

4. EXPERIMENTS AND RESULTS

In order to verify the performance of our method, we compared it with the other four state-of-the-art methods, namely MI_KNN, Relaxed Lasso and generalized multi-class support vector machine (rL-GenSVM) (Kang et al., 2019), Variance_CNN (Var_CNN) (Lyu and Haque, 2018) and ExtraTrees-SVM (ET-SVM) (Hsu and Si, 2018). These methods can realize multiple classification and feature selection of tumors, and have achieved a good classification effect in biomedical data. Our experiment consists of two parts. Firstly, we conducted an experiment on the classification performance of the model to verify that our method could achieve better classification effect in 33 different tumor types. Then, we evaluated the corresponding classification errors of 5 methods when selecting different gene numbers, indicating that our method can obtain a better subset of features. All experiments were executed on a computer server with Windows 7 operating system, Intel Core(TM) i7-10700 CPU (2.9 GHz), 32 GB RAM, 8 GB Nvidia GeForce RTX 2080 SUPER, using Python language.

4.1. Evaluation Metrics

Since pan-cancer classification is a multi-classification problem, we use accuracy to measure the performance. At the same time, to evaluate the performance of the proposed architecture, more performance measures need to be investigated in this research. There are also precision, recall and F1Score (Goutte and Gaussier, 2005) to measure performance in a classification problem. The evaluation indicator is defined as follows.

Accuracy: the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set. The calculation formula is shown below.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

Precision: It represents how many of the samples predicted to be positive are correct. The calculation formula is shown below.

$$P = \frac{TP}{TP + FP} \tag{4}$$

Recall rate: This is how much of the positive sample was predicted correctly. The calculation formula is shown below.

$$R = \frac{TP}{TP + FN} \tag{5}$$

F1-Score: The harmonic mean of the precision rate and recall rate. It is a combination of precision rate and recall rate. The calculation formula is shown below.

$$F1Score = \frac{2PR}{P + R} \tag{6}$$

This research uses the 10-fold cross-validation to calculate accuracy, precision, recall and F1Score.

TABLE 1 | Parameter settings.

Algorithm	Parameter
MI_DenseNetCAM	learning_rate = 0.0001, num_epochs=200, batch_size = 32, growth_rate = 16, compression_factor = 0.5, image_dimension = 60
MI_KNN	n_neighbors = 5
Var_CNN	learning_rate = 0.0001, num_epochs = 200, batch_size = 500
rL-GenSVM	phi = 1/3, *p* = 1, kernel = "rbf," epsilon = 1e-3, lambda = 1e-9, gama = 1e-8, kappa = 2
ET-SVM	C = 0.004, kernel = "linear," decision_function_shape = "ovo," gama = 1

TABLE 2 | The experimental results of five methods.

Method	Accuracy	Precision	Recall	F1-Score
MI_DenseNetCAM	96.81%	96.89%	96.81%	96.85%
MI_KNN	92.61%	92.46%	92.61%	92.40%
Var_CNN	95.59%	95.54%	95.59%	95.43%
rL-GenSVM	87.29%	87.73%	87.29%	86.91%
ET-SVM	90.73%	90.22%	90.73%	89.99%

4.2. Parameters Settings

In this section, the parameter values of all methods are given in **Table 1**. For Var_CNN, rL-GenSVM and ET-SVM, we chose parameter values according to relevant literature (Hsu and Si, 2018; Lyu and Haque, 2018; Kang et al., 2019). For the proposed method, the values of growth_rate and compression_factor are set to 16, 0.5, respectively, which has been analyzed and evaluated in previous studies (Huang et al., 2017). Based on our experimental analysis, the value of image_dimension is set to 60. Similar to Var_CNN, we take the same value for learning_rate and num_epochs.

4.3. Comparison With Other Methods

In this section, we compare the average accuracy, precision, recall and F1-score of MI_KNN, Var_CNN, rL-GenSVM and ET-SVM algorithm. The overall classification results of these methods on 33 tumor types are shown in **Table 2**.

It can be seen from **Table 2**, in terms of accuracy, precision, recall and f1-score, the proposed method MI_DenseNetCAM performs best on the pan-cancer datasets. At the same time, a comparison of accuracy as to each class is shown in **Table 3**. From the previous two experiments, we can see that DenseNet has higher accuracy when using the same preprocessing algorithm. Then, we compared the method with that in the literature (Lyu and Haque, 2018), which makes a similar contribution to our study. Although the overall classification result is only 1.22% higher than Var_CNN algorithm, in terms of the specific accuracy of each class, our method performs better. Especially in ACC, CESC, CHOL, ESCA, MESO and PAAD datasets. Meanwhile, compared with Var_CNN algorithm, our method also has better performance in small sample datasets. The accuracy of our method is 100, 75, and 99% respectively for dataset ACC, CHOL

TABLE 3 | Benchmark datasets.

Tumor type	Cohort	Instances	MI_Dense NetCAM	MI_ KNN	Var_ CNN	rL-GenSVM	ET-SVM
Adrenocortical carcinoma	ACC	79	1	0.95	0.95	0.63	0.92
Bladder urothelial carcinoma	BLCA	408	0.98	0.87	0.97	0.53	0.78
Breast invasive carcinoma	BRCA	1093	0.99	0.99	0.99	0.92	0.99
Cervical and endocervical cancers	CESC	304	0.95	0.88	0.93	0.65	0.86
Cholangiocarcinoma	CHOL	36	0.75	0.58	0.56	0.40	0
Colon adenocarcinoma	COAD	457	0.95	0.99	0.95	0.82	0.98
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	48	1	1	1	1	1
Esophageal carcinoma	ESCA	184	0.85	0.69	0.77	0.50	0.45
Glioblastoma multiforme	GBM	160	0.95	0.92	0.94	0.83	0.81
Head and Neck squamous cell carcinoma	HNSC	520	0.99	0.95	0.98	0.96	0.94
Kidney Chromophobe	KICH	66	0.89	0.75	0.87	0.80	0.64
Kidney renal clear cell carcinoma	KIRC	533	0.94	0.93	0.95	0.89	0.95
Kidney renal papillary cell carcinoma	KIRP	290	0.94	0.86	0.93	0.82	0.83
Acute Myeloid Leukemia	LAML	179	1	1	1	1	1
Brain Lower Grade Glioma	LGG	516	1	0.95	0.98	0.96	0.98
Liver hepatocellular carcinoma	LIHC	371	0.97	0.96	0.97	0.91	0.96
Lung adenocarcinoma	LUAD	515	0.95	0.91	0.95	0.91	0.96
Lung squamous cell carcinoma	LUSC	501	0.93	0.85	0.91	0.84	0.82
Mesothelioma	MESO	87	0.99	0.95	0.94	0.89	0.62
Ovarian serous cystadenocarcinoma	OV	304	1	0.98	0.99	1	1
Pancreatic adenocarcinoma	PAAD	178	1	0.97	0.97	0.95	0.64
Pheochromocytoma and Paraganglioma	PCPG	179	1	0.99	1	0.95	0.96
Prostate adenocarcinoma	PRAD	497	0.99	1	1	0.96	0.99
Rectum adenocarcinoma	READ	166	0	0	0.35	0	0
Sarcoma	SARC	259	0.98	0.95	0.97	0.74	0.98
Skin Cutaneous Melanoma	SKCM	469	0.98	0.97	0.98	1	0.96
Stomach adenocarcinoma	STAD	415	0.96	0.90	0.96	0.93	0.98
Testicular Germ Cell Tumors	TGCT	150	1	0.99	0.99	1	0.83
Thyroid carcinoma	THCA	501	1	1	1	1	0.99
Thymoma	THYM	120	1	0.98	0.99	1	0.91
Uterine Corpus Endometrial Carcinoma	UCEC	545	0.95	0.92	0.96	0.95	0.78
Uterine Carcinosarcoma	UCS	57	0.83	0.72	0.81	0.83	0
Uveal Melanoma	UVM	80	1	1	0.99	1	1

and MESO, and the results are higher than the accuracy obtained by Var_CNN. whose accuracy is 95, 56, and 94% respectively. Since the Guided Grad-CAM algorithm generates heat maps based on the prediction results of each class, the higher the precision in each class, the more likely it is to use heat maps to obtain the optimal subset of features. Moreover, our method requires fewer parameters and uses parameters more efficiently, which can be reflected in the size of the model. our model only uses 13.9 M parameters to achieve an accuracy of 96.81%, while the model of Var_CNN uses 295 M parameters to achieve an accuracy of 95.59%. To achieve a similar level of accuracy, our method only requires around 1/21 of the parameters of Var_CNN. Finally, we compared some of the methods introduced in the related work, and the results show that our method also shows superior performance.

Next, we further evaluated the performance of our proposed method. First, we conducted experiments on the DenseNet model without any preprocessing. In terms of Accuracy, Precision, Recall and F1-Score, the DenseNet model without preprocessing can achieve 93.90, 94.03, 93.90, and 93.89% respectively. Then, we evaluated the effects of different preprocess strategies (Variance, Chi2, *F*-Test, MI) on the classification performance. The experimental results are shown in **Table 4**. It can be seen from **Table 4** that preprocessing based on ML can further improve the accuracy of the classifier. Meanwhile, compared with other methods, MI has better performance on all indicators.

4.4. The Impact of Classifiers on Performance

To further evaluate the impact of different classifiers on the performance of our method, four classifiers, namely KNN, CNN, SVM and DenseNet, are selected to conduct experiments on the pan-cancer data set. The experimental results are shown in **Table 5**. Compared with the other three classifiers, the

TABLE 4 | The performance evaluation results of different preprocess strategies.

Method	Accuracy	Precision	Recall	F1-Score
Var_DenseNet	94.46%	94.62%	94.46%	94.37%
Chi2_DenseNet	95.42%	95.54%	95.42%	95.40%
FTest_DenseNet	95.03%	95.20%	95.03%	95.01%
MI_DenseNetCAM	96.81%	96.89%	96.81%	96.85%

TABLE 5 | The performance evaluation results of four different classifiers.

Classifiers	Accuracy	Precision	Recall	F1-Score
MI_KNN	92.61%	92.46%	92.61%	92.40%
MI_CNN	94.30%	94.37%	94.30%	94.28%
MI_SVM	91.53%	91.67%	91.53%	90.97%
MI_DenseNetCAM	96.81%	96.89%	96.81%	96.85%

TABLE 6 | The performance evaluation results of different image dimensions.

Dimensions	Accuracy	Precision	Recall	F1-Score
30 * 30	93.60%	93.54%	93.60%	93.46%
50 * 50	95.03%	94.82%	95.03%	94.85%
60 * 60	96.81%	96.89%	96.81%	96.85%
70 * 70	95.22%	95.41%	95.22%	95.23%
90 * 90	94.17%	94.18%	94.17%	94.07%
110 * 110	92.93%	93.18%	92.93%	92.81%
130 * 130	93.41%	93.88%	93.41%	93.34%

DenseNet model shows better performance in terms of different evaluation indicators.

4.5. The Impact of Image Dimensions on Performance

To further evaluate the impact of image dimension on the performance of the proposed method, in this section, various image dimensions are adopted to conduct experiments. The experimental results are shown in **Table 6**. As can be seen from **Table 6**, MI_DenseNetCAM achieves the best performance when the image dimension is set to 60.

4.6. Evaluate Important Genes

However, discovering some key genes quickly will reduce the workload of following biological experiments, and help the rapid disease diagnosis. As a result, it is meaningful to obtain small gene sets with high classification accuracy. For the issue, we further evaluated the classification performance for all methods based on small scale genes ranges from 20 to 200. The experimental results are shown in **Figure 3**. The results show that the proposed MI_DenseNetCAM is superior to other methods. It can achieve 83.24% accuracy only using 20 genes.

As can be seen from **Figure 3**, in terms of classification accuracy, MI_DenseNetCAM has the best effect, which is obviously superior to the other four methods, while the rL-GenSVM method has the worst effect, with the accuracy can only be up to 86%. For the other three methods of MI_KNN, ET-SVM and Var_CNN, although their performance is unsatisfactory in the case of a small number of genes, their accuracy is improved with the increase of the number of genes. Compared with the other four methods, MI_DenseNetCAM usually requires fewer genes under the condition of the same precision. For example, with the highest accuracy of 86% of rL-GenSVM as the baseline, we compared the number of genes needed to achieve this accuracy with other methods. MI_DenseNetCAM only requires 25 genes to achieve this accuracy, while ET-SVM requires 70 genes, Var_CNN requires 85 genes, and MI_KNN requires the most. It requires 130 genes. In addition, from **Figure 3**, it is obvious that MI_DenseNetCAM can obtain higher prediction accuracy than the other four methods when dealing with the same number of genes. Therefore, both in terms of the number of genes and accuracy, our method can achieve better performance.

4.7. Gene Analysis

In this section, we conduct further analysis and verify the selected genes by the proposed method. These genes selected by our proposed method are lists in **Table 7**.

We selected 40 genes for further analysis, because it can be seen from **Figure 3** that the accuracy of 40 genes was already very high, and the accuracy did not improve significantly with the increase of the number of genes. Next, the KEGG pathway analysis results for 40 genes are obtained using the David website (https://david.ncifcrf.gov/), trying to find out if significantly enriched pathways are related to the tumor. Pathway analyses showed those genes were significantly enriched in 31 KEGG pathways [Log10(P) <-2 or $P<0.01$], which mainly involved in complement activation, cell projection, cellular response, cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis (**Table 8**). Some of these pathways are already involved in cancer development. For example, hsa04610 might contribute to the progression of bladder cancer (Liu et al., 2020). The hsa05133 pathway is related to the hsa04610 pathway, so it also promotes bladder cancer formation. In the hsa04611 pathway, cancer cells migrate to the vasculature and interact with platelets, causing inflammation and promoting mesothelioma growth (Jurasz et al., 2004; Sekido, 2013). The hsa04512 pathway interaction is involved in six critical cancer hallmarks (Pickup et al., 2014). So, the related genes in these pathways can then be viewed as tumor specific biomarkers.

For other genes that are not significantly enriched in the pathway, we can retrieve these genes from the GeneCard database (www.genecards.org/). GeneCard is a searchable, comprehensive and public database containing genetic analysis data that provides concise information on all known and predicted human genes in the genome, proteome, transcription, genetics and function. GeneCard is a comprehensive database of human genes. So the easiest way to see a summary of a gene is to use GeneCard.

As to COAD(Colon adenocarcinoma), LGALS4 is associated with the colon. LGALS4 is a Protein Coding gene, the expression of this gene is restricted to the small intestine, colon, and rectum, and it is under-expressed in colorectal cancer. In the paper (Kim et al., 2013), the authors have demonstrated that LGALS4 is predominantly expressed in the luminal epithelia of

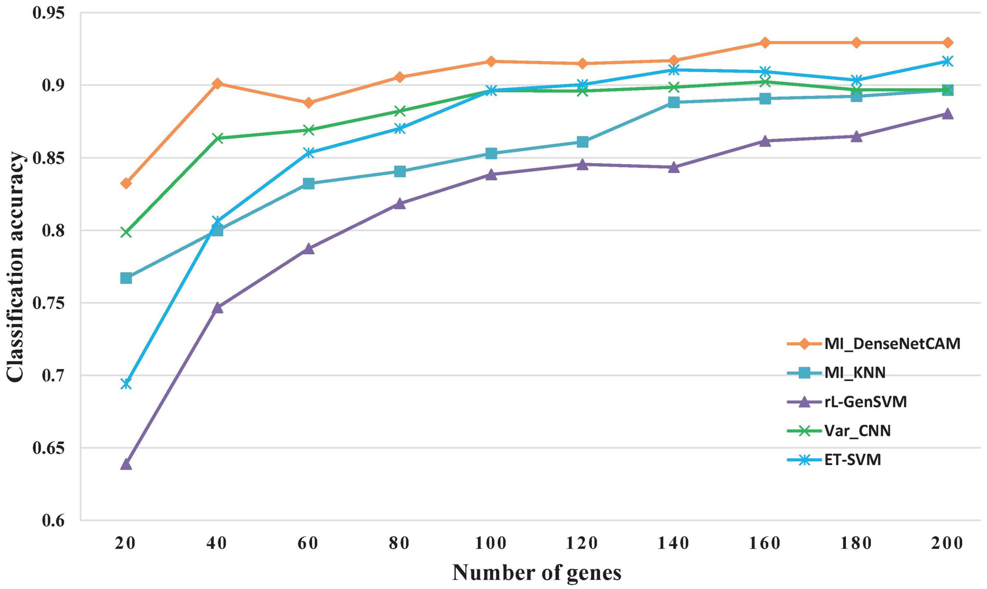


FIGURE 3 | The Classification accuracy of different gene numbers.

the gastrointestinal tract, and its loss of expression plays a key role in colorectal tumorigenesis.

As to GBM(Glioblastoma multiforme), It is a primary brain tumor that develops from astroglial cells. The gene GFAP selected by our method is a protein-coding gene. This gene encodes one of the major intermediate filament proteins of mature astrocytes. It is used as a marker to distinguish astrocytes from other glial cells during development. In the paper (Heiland et al., 2019), the authors demonstrated that tumor associated glial cells are widespread in GBM. In the paper (Tichy et al., 2016), the authors demonstrated that the GFAP gene was over-expression in GBM and that GFAP could be considered as a biomarker of astrocytic pathology in neurological diseases.

As to LUSC(Lung squamous cell carcinoma), The gene SFTPA2 selected by our method is a protein-coding gene. This gene is one of several genes encoding pulmonary-surfactant associated proteins (SFTPA) located on chromosome 10. Mutations in this gene and a highly similar gene located nearby, which affect the highly conserved carbohydrate recognition domain, are associated with idiopathic pulmonary fibrosis. In the paper (Peng et al., 2015), the authors demonstrated that SFTPA2 encodes surfactant protein A that plays a vital role in maintaining normal lung function and has been implicated in various lung diseases, which can accurately distinguished lung cancer from other cancer samples.

As to OV(Ovarian serous cystadenocarcinoma), A product of the MUC1 gene of the genes selected by our method has been used as a marker for different cancers. MUC1 is a Protein Coding gene. In the paper (Hu et al., 2006), the authors demonstrated that MUC1 overexpresses in the majority of ovarian carcinomas and contributes to the metastasis process, promotes tumor formation and metastasis. It plays a role in contributing to ovarian tumor growth.

TABLE 7 | Selected genes.

Number of genes	The name of the gene
40	GSTA1, C4A, COL3A1, PABPC1, COL1A1, KRT13, S100A6, SERPINA1, FGA, MUC2, COL1A2, APOE, KRT5, MALAT1, GFAP, TUBA1A, KRT14, KLK1, ATP1A1, RGS5, SPP1, CLU, S100A9, TF, APOC1, MUC1, ADAM6, SFTPA2, BCAM, TTR, CHGA, SCG2, FASN, PDLIM5, LGALS4, CA2, MYH11, SILV, PGC, TG

As to STAD(Stomach adenocarcinoma), The gene PGC selected by our method is a protein-coding gene. The protein encoded by this gene is a digestive enzyme produced in the stomach, Polymorphisms in this gene are associated with susceptibility to gastric cancers. In the paper (Shen et al., 2017), the authors demonstrated that PGC is a comparatively ideal negative marker of gastric cancer.

As to TCHA(Thyroid carcinoma), The S100A6 gene selected by our method is a protein-coding gene. In the paper (Sofiadis et al., 2010), the authors demonstrated that the expression patterns of S100A6 in thyroid carcinoma are unique compared with those of other carcinomas, and over-expression in thyroid carcinoma. S100A6 gene can be used as a biomarker of Thyroid carcinoma.

In order to more visually show the expression of genes in different tumor samples, we can use heat maps to understand the distribution of data or the differential expression of genes. In the heat map, the gradient color is used to represent the change of values. The data value size can be visually represented by the defined color depth. In addition, each column represents

TABLE 8 | The KEGG pathway analysis.

KEGG Pathways	Description	*P*-Value	Genes
hsa04610	Complement and coagulation cascades	9.50E-09	C3,CLU,C4A,FGA,SERPINA1
hsa05133	Pertussis	6.40E-07	C3,CALML3,SFTPA2,C4A
hsa04974	Protein digestion and absorption	1.22E-06	COL3A1,COL1A2,ATP1A1,COL1A1
hsa05146	Amoebiasis	1.50E-06	COL3A1,MUC2,COL1A2,COL1A1
hsa04611	Platelet activation	4.19E-06	COL3A1,COL1A2,FGA,COL1A1
hsa04918	Thyroid hormone synthesis	3.80E-05	TG,ATP1A1,TTR
hsa04971	Gastric acid secretion	3.95E-05	CALML3,CA2,ATP1A1
hsa04933	AGE-RAGE signaling pathway in diabetic complications	9.05E-05	COL3A1,COL1A2,COL1A1
hsa04926	Relaxin signaling pathway	1.93E-04	COL3A1,COL1A2,COL1A1
hsa04964	Proximal tubule bicarbonate reclamation	2.02E-04	CA2,ATP1A1
hsa04915	Estrogen signaling pathway	2.29E-04	CALML3,KRT14,KRT13
hsa04145	Phagosome	3.02E-04	C3,TUBA1A,SFTPA2
hsa04979	Cholesterol metabolism	8.78E-04	APOE,APOC1
hsa04961	Endocrine and other factor-regulated calcium reabsorption	8.78E-04	KLK1,ATP1A1
hsa04978	Mineral absorption	9.82E-04	TF,ATP1A1
hsa05150	Staphylococcus aureus infection	1.58E-03	C3,C4A
hsa04976	Bile secretion	1.77E-03	CA2,ATP1A1
hsa04512	ECM-receptor interaction	2.49E-03	COL1A2,COL1A1
hsa04970	Salivary secretion	2.71E-03	CALML3,ATP1A1
hsa04972	Pancreatic secretion	3.20E-03	CA2,ATP1A1
hsa04925	Aldosterone synthesis and secretion	3.20E-03	CALML3,ATP1A1
hsa04916	Melanogenesis	3.39E-03	CALML3,TYRP1
hsa04270	Vascular smooth muscle contraction	5.65E-03	CALML3,MYH11
hsa05322	Systemic lupus erythematosus	5.73E-03	C3,C4A
hsa04910	Insulin signaling pathway	6.07E-03	CALML3,FASN
hsa05418	Fluid shear stress and atherosclerosis	6.24E-03	CALML3,GSTA1
hsa01100	Metabolic pathways	6.77E-03	TYRP1,BCAM,FASN,GSTA1,CA2
hsa04261	Adrenergic signaling in cardiomyocytes	7.12E-03	CALML3,ATP1A1
hsa04022	cGMP-PKG signaling pathway	8.84E-03	CALML3,ATP1A1
hsa04530	Tight junction	9.14E-03	MYH11,TUBA1A
hsa05010	Alzheimer disease	9.24E-03	CALML3,APOE

the expression of each gene in different samples, and each row represents the expression of all genes in each sample. A heat map representation of the relative expression levels of the top 40 genes across all tumor samples is shown in **Figure 4**.

From **Figure 4**, we were able to look at the level of expression of each gene in all tumor types. The use of heat maps is more indicative of the relationship between genes and samples. For example, the gene of GFAP was highly expressed in LGG and GBM and low in all other tumors. The gene of LGALS4 was moderately expressed in COAD and READ and low in all other tumors. The heat map visually shows that these genes are differential expression in different tumor samples, which also demonstrates the effectiveness of our proposed method. It is feasible to identify biomarkers with our proposed method.

These results indicate that the genes selected by our method are closely related to the corresponding tumor types, and therefore we can use these selected genes as biomarkers to distinguish different tumors. For the rest of the genes (PABPC1, KRT5, MALAT1, RGS5, SPP1, S100A9, ADAM6, CHGA, SCG2, PDLIM5, SILV), they were neither significantly enriched in the pathway nor found to be tumor-related in GeneCard. The role of these genes in tumor development is unclear, so, pending further study by biological researchers.

5. CONCLUSIONS AND FUTURE WORKS

In recent years, with the rapid development of the new generation of gene sequencing technology, the generated bioinformatics data such as gene, protein and metabolism are generally high-dimensional and complex. There are a lot of important data closely related to life and health in these data. However, due to the high data dimension, it is impossible to analyze all the data. Feature selection technology can effectively screen high-dimensional data, reduce the workload of data analysis by reducing dimensions, find disease-related markers to achieve early and accurate diagnosis.

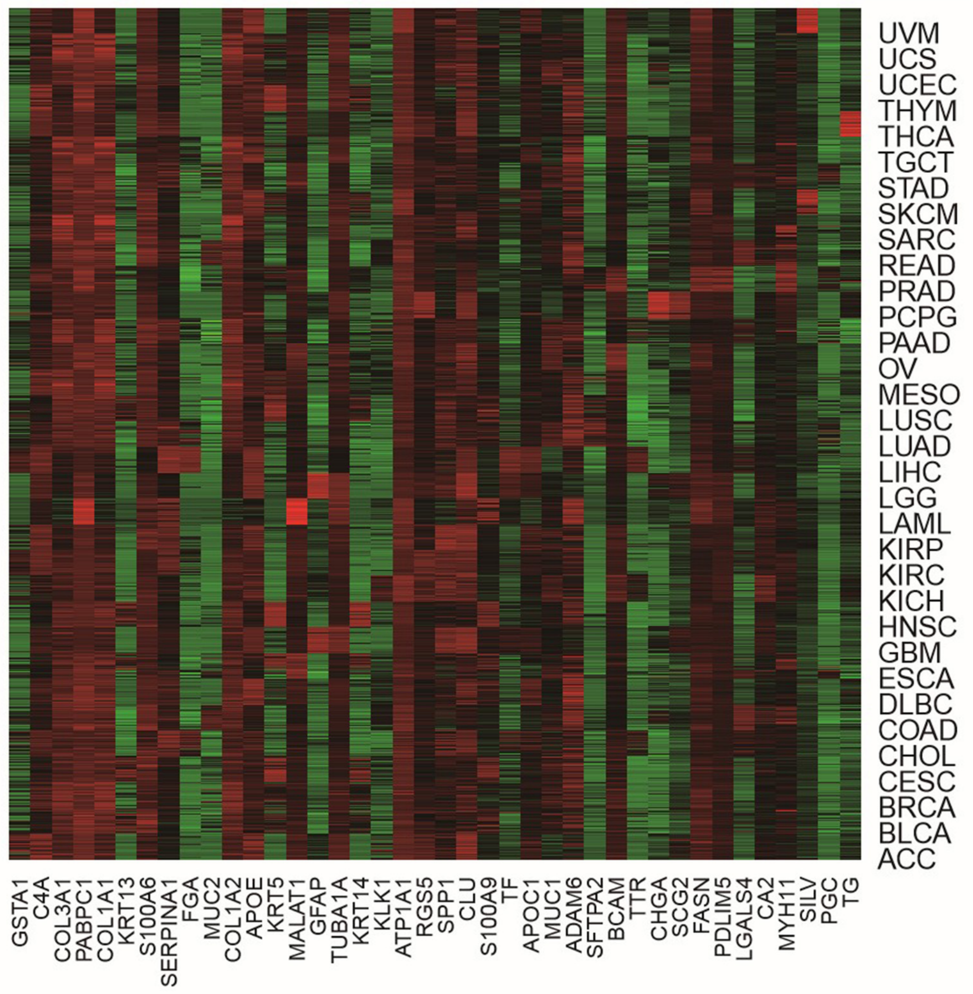


FIGURE 4 | The heat map of the top 40 genes across all tumor samples.

In this paper, we have designed a novel approach to classify different types of cancer, whilst it can be used to find biomarkers associated with tumors. We identified biomarkers that were significantly associated with the pan-cancer studies by innovatively combining the traditional machine learning model and deep learning. The presented results and the performance metrics performed in this research showed that the proposed approach achieved an overall testing accuracy of 96.81%. Moreover, the results of our experiment also demonstrated that the genes selected by our method were related to the corresponding tumor types by means of KEGG pathway analysis and gene query, some of these genes have been used as clinical markers. These biomarkers can be used to quickly identify the type of tumor, so as to detect and treat the tumor in advance and improve the cure rate of the tumor.

The methods presented in this paper are not limited to RNA-Seq data, but also applicable to other types of data. However, the method in this paper still needs improvement. For example, the preprocessing strategy of our method includes not only the filter approach, but also the wrapper approach. So, one of the potential future works is applying a new preprocessing strategy to verify and extend this approach. In conclusion, a novel approach for the classification of pan-cancer has been proposed in this paper, which can accurately predict the type of tumor and find tumor-related biomarkers from high-dimensional biological datasets, have broad application prospects and great scientific research prospects, and is of great significance to human development.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

JW and CY conceived and designed the approach. XD performed the experiments. HL and JL analyzed the data. CY and XD wrote the manuscript. JW and GZ supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 61802113, 61802114, and 61972134). Science and Technology Development Plan Project of Henan Province, (Nos. 202102210173 and 212102210091). China Postdoctoral Science Foundation (No. 2020M672212). Henan Province Postdoctoral Research Project Founding.

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511. doi: 10.1038/35000501

Danaee, P., Ghaeini, R., and Hendrix, D. A. (2017). "A deep learning approach for cancer detection and relevant gene identification," in *Pacific Symposium on Biocomputing 2017* (Hawaii: World Scientific), 219–229.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (Florida: IEEE), 248–255.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science* 286, 531–537. doi: 10.1126/science.286.5439.531

Goutte, C., and Gaussier, E. (2005). "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval* (Santiago de Chile: Springer), 345–359.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas), 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). "Identity mappings in deep residual networks," in European Conference on Computer Vision (Amsterdam: Springer), 630–645.

Heiland, D. H., Ravi, V. M., Behringer, S. P., Frenking, J. H., Wurm, J., Joseph, K., et al. (2019). Tumor-associated reactive astrocytes aid the evolution of immunosuppressive environment in glioblastoma. *Nat. Commun.* 10, 1–12. doi: 10.1038/s41467-019-10493-6

Hsu, Y.-H. and Si, D. (2018). "Cancer type prediction and classification based on rna-sequencing data," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, HI: IEEE), 5374–5377.

Hu, X. F., Yang, E., Li, J., and Xing, P. X. (2006). Muc1 cytoplasmic tail: a potential therapeutic target for ovarian carcinoma. *Exp. Rev. Anticancer Therapy* 6, 1261–1271. doi: 10.1586/14737140.6.8.1261

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Hawaii: IEEE), 4700–4708.

Huang, J., Cai, Y., and Xu, X. (2007). A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recogn. Lett.* 28, 1825–1844. doi: 10.1016/j.patrec.2007.05.011

Jurasz, P., Alonso-Escolano, D., and Radomski, M. W. (2004). Platelet–cancer interactions: mechanisms and pharmacology of tumour cell-induced platelet aggregation. *Br. J. Pharmacol.* 143:819. doi: 10.1038/sj.bjp.0706013

Kang, C., Huo, Y., Xin, L., Tian, B., and Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed lasso and generalized multi-class support vector machine. *J. Theoret. Biol.* 463, 77–91. doi: 10.1016/j.jtbi.2018.12.010

Khalifa, N. E. M., Taha, M. H. N., Ali, D. E., Slowik, A., and Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor rna-seq data: a novel optimized deep learning approach. *IEEE Access* 8, 22874–22883. doi: 10.1109/ACCESS.2020.2970210

Kim, S. W., Park, K. C., Jeon, S. M., Ohn, T. B., Kim, T. I., Kim, W. H., et al. (2013). Abrogation of galectin-4 expression promotes tumorigenesis in colorectal cancer. *Cell. Oncol.* 36, 169–178. doi: 10.1007/s13402-013-0124-x

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E* 69:066138. doi: 10.1103/PhysRevE.69.066138

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., et al. (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC Genomics* 18:508. doi: 10.1186/s12864-017-3906-0

Liu, Y., Xiong, S., Liu, S., Chen, J., Yang, H., Liu, G., et al. (2020). Analysis of gene expression in bladder cancer: possible involvement of mitosis and complement and coagulation cascades signaling pathway. *J. Comput. Biol.* 27, 987–998. doi: 10.1089/cmb.2019.0237

Lyu, B., and Haque, A. (2018). "Deep learning based tumor type classification using gene expression data," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Washington, DC), 89–96.

Martín-Valdivia, M. T., Díaz-Galiano, M. C., Montejo-Raez, A., and Urena-Lopez, L. (2008). Using information gain to improve multi-modal information retrieval systems. *Inform. Proc. Manag.* 44, 1146–1158. doi: 10.1016/j.ipm.2007.09.014

Peng, L., Bian, X. W., Xu, C., Wang, G. M., Xia, Q. Y., Xiong, Q., et al. (2015). Large-scale rna-seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 tcga cancer types. *Sci. Rep.* 5:13413. doi: 10.1038/srep13413

Pickup, M. W., Mouw, J. K., and Weaver, V. M. (2014). The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep.* 15, 1243–1253. doi: 10.15252/embr.201439246

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi: 10.1093/bioinformatics/btm344

Sakri, S. B., Rashid, N. B. A., and Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access* 6, 29637–29647. doi: 10.1109/ACCESS.2018.2843443

Sekido, Y. (2013). Molecular pathogenesis of malignant mesothelioma. *Carcinogenesis*, 34(7):1413–1419. doi: 10.1093/carcin/bgt166

Sharmin, S., Shoyaib, M., Ali, A. A., Khan, M. A. H., and Chae, O. (2019). Simultaneous feature selection and discretization based on mutual information. *Pattern Recogn.* 91, 162–174. doi: 10.1016/j.patcog.2019.02.016

Shen, S., Jiang, J., and Yuan, Y. (2017). Pepsinogen c expression, regulation and its relationship with cancer. *Cancer Cell Int.* 17:57. doi: 10.1186/s12935-017-0426-6

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Sofiadis, A., Dinets, A., Orre, L. M., Branca, R. M., Juhlin, C. C., Foukakis, T., et al. (2010). Proteomic study of thyroid tumors reveals frequent up-regulation of the ca2+-binding protein s100a6 in papillary thyroid carcinoma. *Thyroid* 20, 1067–1076. doi: 10.1089/thy.2009.0400

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1–9.

Tichy, J., Spechtmeyer, S., Mittelbronn, M., Hattingen, E., Rieger, J., Senft, C., et al. (2016). Prospective evaluation of serum glial fibrillary acidic protein (gfap) as a diagnostic marker for glioblastoma. *J. Neurooncol.* 126, 361–369. doi: 10.1007/s11060-015-1978-8

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* 85, 189–203. doi: 10.1016/j.jbi.2018.07.014

Van’t Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536. doi: 10.1038/415530a

Wainberg, M., Merico, D., Delong, A., and Frey, B. J. (2018). Deep learning in biomedicine. *Nat. Biotechnol.* 36, 829–838. doi: 10.1038/nbt.4233

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45:1113. doi: 10.1038/ng.2764

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 28 June 2021
doi: 10.3389/fgene.2021.699510

A Cluster-Based Approach for the Discovery of Copy Number Variations From Next-Generation Sequencing Data

*Guojun Liu and Junying Zhang**

School of Computer Science and Technology, Xidian University, Xi'an, China

The next-generation sequencing technology offers a wealth of data resources for the detection of copy number variations (CNVs) at a high resolution. However, it is still challenging to correctly detect CNVs of different lengths. It is necessary to develop new CNV detection tools to meet this demand. In this work, we propose a new CNV detection method, called CBCNV, for the detection of CNVs of different lengths from whole genome sequencing data. CBCNV uses a clustering algorithm to divide the read depth segment profile, and assigns an abnormal score to each read depth segment. Based on the abnormal score profile, Tukey's fences method is adopted in CBCNV to forecast CNVs. The performance of the proposed method is evaluated on simulated data sets, and is compared with those of several existing methods. The experimental results prove that the performance of CBCNV is better than those of several existing methods. The proposed method is further tested and verified on real data sets, and the experimental results are found to be consistent with the simulation results. Therefore, the proposed method can be expected to become a routine tool in the analysis of CNVs from tumor-normal matched samples.

Keywords: next-generation sequencing, copy number variation, clustering algorithm, abnormal score, Tukey's fences

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Ruifeng Hu,
Harvard Medical School,
United States
Cuncong Zhong,
University of Kansas, United States

****Correspondence:***
Junying Zhang
jyzhang@mail.xidian.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *23 April 2021*
Accepted: *07 June 2021*
Published: *28 June 2021*

Citation:
Liu G and Zhang J (2021) A Cluster-Based Approach for the Discovery of Copy Number Variations From Next-Generation Sequencing Data. Front. Genet. 12:699510. doi: 10.3389/fgene.2021.699510

INTRODUCTION

The copy number variation (CNV) of DNA fragments has been widely recognized as a major type of structural variations, and can cause the amplification or deletion of DNA fragments, the lengths of which are greater than 1 kbp in the human genome (Freeman et al., 2006). Some CNVs, called germline CNVs, are also present in normal tissues of the human body; these generally originate from family inheritance, and can cause cancers and diseases (Kuiper et al., 2010; Krepischi et al., 2012). The CNVs in tumor tissue are generally called somatic CNVs, which are acquired CNVs, and cause tumor formation by oncogene and tumor suppressor gene mutations (Stratton et al., 2009; Beroukhim et al., 2010; Pei et al., 2020). Many experimental studies have proven that CNVs can change the doses of genes and lead to the reorganization of chromosome structure (Sharp et al., 2005; Magi et al., 2017; Pei et al., 2021b), and makes an important contribution to the occurrence and formation of tumors and various disorders (Pei et al., 2021a). For example, it can cause schizophrenia and autism disorders in humans

(Sebat et al., 2007; Cook and Scherer, 2008; Stone et al., 2008). Some studies have shown that CNVs are related to cancer, such as breast and ovarian cancer (Tchatchou and Burwinkel, 2008; Adam and David, 2009; Malek et al., 2011). In practical applications, there is a strong requirement to capture CNVs of various range lengths, which requires the developed tools to have higher resolution and better robustness than previously developed tools to reduce the false positive rate of test results. Therefore, it is still a difficult task to effectively detect CNVs of different lengths.

Compared with traditional (array-based) detection methods (Carter, 2007; Buysse et al., 2009), the detection cost has been greatly reduced and resolution has reached the base-pair level with the emergence of next-generation sequencing technology. In recent years, most related tools for CNV detection using next-generation sequencing data have been developed based on paired-end mapping (PEM) (Korbel et al., 2007) and depth of coverage (DOC) (Yoon et al., 2009) strategies. The basic concept of PEM-based methods is that the insertion size of aligned paired-end reads is significantly different from the insertion size preset by the laboratory (Medvedev et al., 2009). While PEM-based methods can detect amplification, deletion, insertion, translocation, etc., they can only identify those insertion variants whose lengths are less than the preset insertion length. The basic concept of DOC-based methods is that the number of reads aligned to each position of the reference genome is proportional to the number of copies corresponding to that position (Yoon et al., 2009). In principle, DOC-based methods can detect CNVs of various lengths. However, in practical applications, they are more suitable for the detection of long CNVs, and cannot accurately detect the boundaries of the CNVs.

Generally, DOC-based methods require the input of tumor-normal matched samples to detect the tumor genome and effectively capture CNVs. The workflow of this type of method is: (1) input tumor-normal matched samples; (2) obtain read count profiles with SAMtools (Li et al., 2009); (3) bin read count profiles (Chiang et al., 2009) and generate read depth profiles; (4) use the joint read depth information of the tumor-normal matched samples to build a statistical model; (5) choose a suitable threshold to predict CNVs. It is generally believed that the deviation caused by sequencing is consistent in the same areas of the two samples. Therefore, DOC-based methods use the read depth ratio information to eliminate these deviations (GC content and mappability biases) (Bentley et al., 2008; Chiang et al., 2009). Some well-known methods have been developed to detect CNVs from tumor-normal matched samples, including BIC-seq2 (Xi et al., 2016), SeqCNV (Chen et al., 2017), and CNVkit (Talevich et al., 2016). BIC-seq2 preprocesses the sequenced reads, including by calibrating the GC content bias, removing mappability bias, and normalizing reads at the nucleic acid level. Based on the preprocessed data, the segmentation procedure is executed using the bayesian information criterion, by which CNVs are forecasted. It is not sensitive to the detection of short CNVs. SeqCNV extracts the read depth information of the tumor-normal matched samples to build a maximum penalized likelihood estimation model to predict CNVs. It detects a small number of CNVs, most of which are the gain areas and true positives, and its detection is more conservative than that of BIC-seq2. It has a long running time and is not suitable for testing samples with long CNVs. CNVkit is a software toolkit that extracts the information of on- and off-target sequenced reads. It adopts a rolling median method to normalize the GC content bias, mappability bias, and target density bias, and to reduce the impact on the true copy number status. CNVkit detects the CNVs, many of which are deletion regions. However, it is not sensitive to the detection of short CNVs.

In consideration of the limitations of the existing methods, in this study, a new tumor-normal matched sample-based CNV detection method, called CBCNV (cluster-based approach for CNV detection), is proposed for the prediction of CNVs using whole-genome sequencing data. CBCNV extracts the read count profiles of tumor-normal matched samples with SAMtools (Li et al., 2009). The preprocessing program is executed on the read count profiles, which can yield the read depth segment profiles, the dimensions of which are transformed into two-dimensional space. CBCNV adopts the k-means algorithm to cluster the preprocessed read depth segment profiles (Hartigan and Wong, 1979), which can yield clusters of different sizes. The clusters are sorted from largest to smallest according to the number of elements in each cluster. Then, by setting a boundary threshold, these clusters are divided into large and small clusters. Based on the above definition, CBCNV assigns a cluster-based abnormal score for each read depth segment. Using the cluster-based abnormal score profiles, Tukey's fences method is employed to announce candidate CNVs (Zijlstra et al., 2007). The performance of the proposed method is verified using simulated and real data sets, and is compared with several existing CNV detection methods. The experimental results show that the performance of CBCNV is better than several other comparison methods, especially for low-purity samples. In addition, CBCNV is also found to detect some biologically meaningful CNVs, which can provide some valuable reference information for assistance with clinical diagnosis and targeted drug research.

The remainder of this article is organized as follows. Section "Materials and Methods" includes the workflow of CBCNV, data preprocessing, the calculation of cluster-based abnormal score, and the prediction of CNVs. In section "Results," simulation and real experiments are designed, and the experimental results are analyzed and discussed. Section "Discussion and Conclusion" summarizes this research and puts forward ideas for future work.

MATERIALS AND METHODS

Overview of CBCNV

CBCNV is a DOC-based approach that is suitable for the detection of tumor-normal matched samples, and can identify somatic CNVs and germline CNVs from whole-genome sequencing data. The pipeline of CBCNV is described in detail in **Figure 1**. The sequenced tumor-normal matched samples that are composed of a large number of sequenced reads are compared to the reference genome using the BWA tool (Li and Durbin, 2010). Then, the read count profiles of the tumor-normal matched samples are generated with SAMtools (Li et al., 2009).

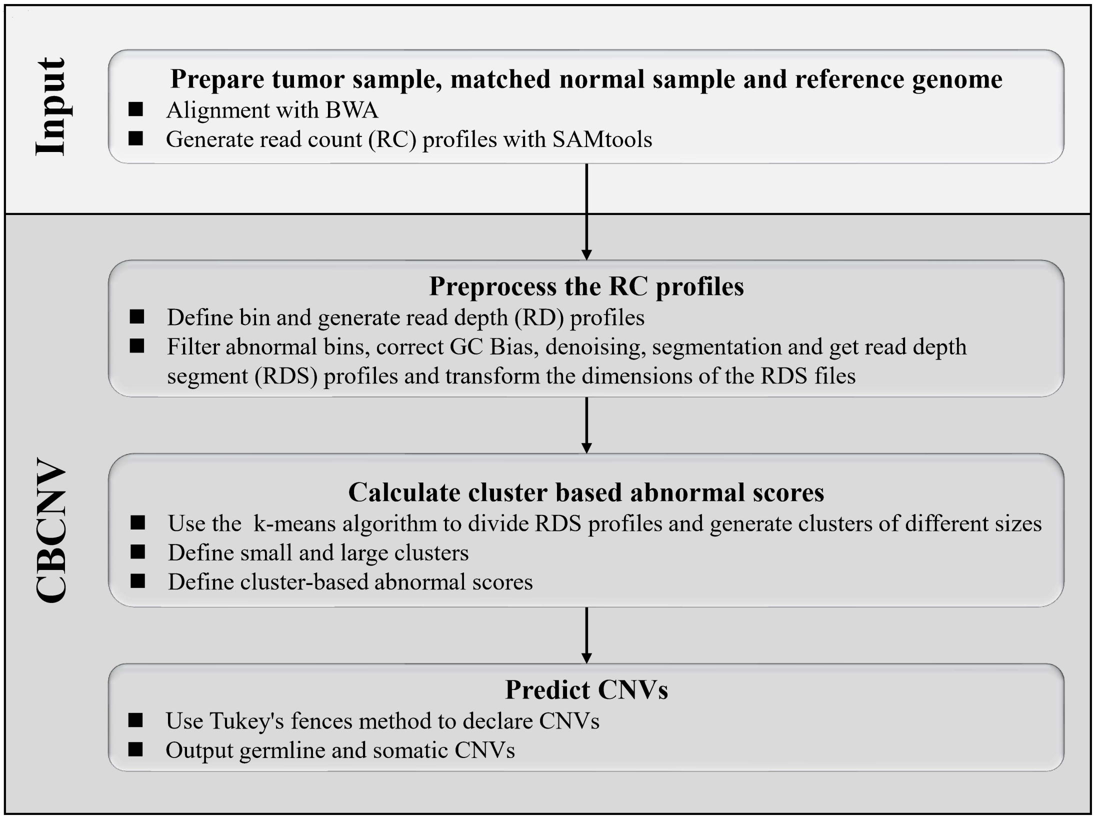

Input
Prepare tumor sample, matched normal sample and reference genome
■ Alignment with BWA
■ Generate read count (RC) profiles with SAMtools
CBCNV
Preprocess the RC profiles
■ Define bin and generate read depth (RD) profiles
■ Filter abnormal bins, correct GC Bias, denoising, segmentation and get read depth segment (RDS) profiles and transform the dimensions of the RDS files
Calculate cluster based abnormal scores
■ Use the k-means algorithm to divide RDS profiles and generate clusters of different sizes
■ Define small and large clusters
■ Define cluster-based abnormal scores
Predict CNVs
■ Use Tukey's fences method to declare CNVs
■ Output germline and somatic CNVs

FIGURE 1 | Overview of the workflow of CBCNV. It is mainly composed of four steps, which includes preparing input files, preprocessing read count profiles, calculating cluster-based abnormal scores, and recording CNVs.

Based on the read count profiles, the following four steps are conducted for CBCNV to complete CNV detection. The first step involves defining the bin, dividing the reference genome into continuous and non-overlapping regions according to the bin size, and generating the read depth profiles. In the second step, the abnormal bins are removed, and the GC content bias is corrected. The read depth profiles are denoised and segmented to generate the read depth segment profiles. The dimensions of the read depth segment profiles are converted from one-dimensional to two-dimensional space. In the third step, the preprocessed read depth segment profiles are clustered via the k-means method to form clusters of different sizes. A boundary value is set to divide large and small clusters. The cluster-based abnormal score is defined based on the following two situations (He et al., 2003): (1) if a read depth segment belongs to a small cluster, the cluster-based abnormal score is defined as the distance between the read depth segment and the center of the large cluster that is the closest to it; (2) if a read depth segment belongs to a large cluster, the cluster-based abnormal score is defined as the distance between the read depth segment and the center of the large cluster. Finally, in the fourth step, Tukey's fences method is employed to predict CNVs (Zijlstra et al., 2007). The CBCNV software is developed based on the R and Python languages (Zhao et al., 2019). Its source code is public, and can be downloaded from https://github.com/gj-123/CBCNV/releases, where users can easily install and use the software according to the instructions.

Data Preprocessing

The sequenced reads are aligned to the reference genome with BWA (Li and Durbin, 2010), and the read count profiles are generated by SAMtools (Li et al., 2009). The reference genome is composed of five types of positions ("A", "T", "G", "C", and "N"). Here, "N" indicates the base positions that cannot be determined during the sequencing process. The sequenced reads cannot be matched to the "N" positions, which are often mistaken for CNV deletion regions. To obtain reasonable read count profiles, a binning strategy is adopted to deal with the "N" positions (Yuan et al., 2018). The read count profiles are divided into continuous and non-overlapping areas according to the bin size. The bins that contain the "N" positions are treated as abnormal bins and filtered out. The mean read count value of each bin is calculated to obtain the read depth profiles. Based on the above processing, Eq. (1) is used to deal with GC content bias (Yoon et al., 2009):

$$RD_i^{'} = RD_i \cdot \frac{RD_m}{RD_{gc}}, \quad (1)$$

where RD_i and $RD_i^{'}$ represent the original and revised read depth values of the i-th bin, respectively, RD_m represents the mean value of the read depth of all bins, and RD_{gc} represents the mean read depth value of the bins, the GC content of which is equal to that of the i-th bin. Sequencing errors and various deviations will lead to a substantial amount of noise in the read depth data, and ultimately false test results. Thus, noise reduction is a necessary step in CNV detection. The fused lasso regression method is adopted to smooth the read depth profile (Tibshirani and Wang, 2008). This method effectively considers the copy number relationship between adjacent read depth signals, which allows a reasonable read depth segment profile to be obtained. Based on the denoised read depth segment profile, Eqs. (2–5) (Li Y. et al., 2019; Liu et al., 2020) are used to transform its dimensions.

$$CN = CN_{norm} \cdot \frac{RDS_i}{RDS_m} \quad 1 \leq i \leq |RDS| \tag{2}$$

$$RDSR = \frac{RDS_i}{RDS_m} \quad 1 \leq i \leq |RDS| \tag{3}$$

$$RDSD = \begin{cases} \frac{\sum_{j=i+1}^{i+L}|RDSR_i - RDSR_j|}{L} & i = 1, 5 \leq L \leq 20 \\ \frac{\sum_{j=1}^{i+L}|RDSR_i - RDSR_j|}{i-1+L} & 1 < i \leq L, 5 \leq L \leq 20 \\ \frac{\sum_{j=i-L}^{i+L}|RDSR_i - RDSR_j|}{2L} & L < i \leq |RDS| - L, 5 \leq L \leq 20 \\ \frac{\sum_{j=i-L}^{|RDS|-1}|RDSR_i - RDSR_j|}{L+|RDS|-i-1} & |RDS| - L < i \leq |RDS| - 1, 5 \leq L \leq 20 \\ \frac{\sum_{j=i-L}^{i-1}|RDSR_i - RDSR_j|}{L} & i = |RDS|, 5 \leq L \leq 20 \end{cases} \tag{4}$$

$$RDS^{'} = \{CN, RDSD\} \tag{5}$$

In Eq. (2), CN_{norm} represents the normal copy number, and its value is equal to 2. Additionally, RDS_i represents the value of the i-th read depth segment, RDS_m represents the mean across all the read depth segments, and CN represents the set of copy number, which is composed of the copy number of all read depth segments. $|RDS|$ represents the number of elements in the read depth segment set. In Eq. (3), $RDSR$ represents a set that is composed of the ratio between RDS_i and RDS_m. In Eq. (4), L represents the number of left and right neighbors of the i-th element of $RDSR$, and is set to 10 by default. $|RDSR_i - RDSR_j|$ represents the absolute value of the difference between $RDSR_i$ and $RDSR_j$, and $RDSD$ represents the set of differences of each element in $RDSR$. In Eq. (5), $RDS^{'}$ represents a two-dimensional data set, which is composed of CN and $RDSD$. This processing step provides two perspectives to observe read depth segments. The first dimension can approximately reflect the copy number status for each read depth segment, which provides a longitudinal and global perspective to observe the trend of copy number changes. The second dimension indirectly reflects the difference between a read depth segment and its surrounding read depth segments, which provides a horizontal and partial perspective to illustrate the relevance of the copy number status of each read depth segment. Moreover, this processing step provides a valid data set for the calculation of cluster-based abnormal scores, which is elaborated in the next subsection.

Calculation of Cluster-Based Abnormal Scores

Based on the $RDS^{'}$ profile, a cluster-based abnormal score is calculated for each read depth segment. Here, each element of $RDS^{'}$ is regarded as an object O. The cluster-based abnormal score is designed based on the concept of CBLOF (He et al., 2003), and is different from the traditional tumor-normal matched samples based CNV detection methods, which utilize read depth information to fit a statistical model and set a threshold to predict CNVs. The cluster-based abnormal score reflects the isolation degree of the local small cluster relative to the large cluster around it, as well as the deviation degree of each object in the large cluster relative to its cluster center, which indirectly reflects the abnormal degree of the copy number of each object. If the cluster-based abnormal score of an object is higher than those of most objects, it is likely a CNV. To further calculate the cluster-based abnormal scores, the definition is subsequently introduced in detail. First, the k-means algorithm is executed on the data set $RDS^{'}$, and can divide the data set to form clusters of different sizes. Equation (6) is used to describe the clustering result:

$$RDSC = \{RDSC_1, RDSC_2, \cdots, RDSC_{k-1,}RDSC_k\}, \tag{6}$$

$$RDSC_i \cap RDSC_j = \emptyset, 1 \leq i \leq k, 1 \leq j \leq k, i \neq j,$$

where RDSC represents a set of k clusters. Second, based on the first step, RDSC is divided into large and small clusters (He et al., 2003), as given by Eqs. (7–11).

$$RDSC^{'} = \{RDSC_1^{'}, RDSC_2^{'}, \cdots, RDSC_{k-1}^{'}, RDSC_k^{'}\} \tag{7}$$

$$|RDSC_1^{'}| + |RDSC_2^{'}| + \cdots + |RDSC_\theta^{'}| \ \geq \ |RDSC^{'}| \cdot x \tag{8}$$

$$\frac{|RDSC_\theta^{'}|}{|RDSC_{\theta+1}^{'}|} \ \geq \ \mathrm{y} \tag{9}$$

$$LRDSC^{'} = \{RDSC_i^{'} | 1 \leq i \leq \theta\} \tag{10}$$

$$SRDSC^{'} = \{RDSC_j^{'} | \theta < j \leq k\} \tag{11}$$

$$RDSC_1^{'}| \geq \ |RDSC_2^{'}| \ \geq \cdots \geq \ |RDSC_{k-1}^{'}| \geq |RDSC_k^{'}|,$$

$$1 \leq i \leq k, 1 \leq j \leq k, i \neq j$$

In Eq. (7), $RDSC^{'}$ represents the sorted cluster set RDSC, which is sorted in descending order. In Eq. (8), | *| represents the number of elements in a cluster, θ represents the boundary threshold of large and small clusters, and x represents a ratio between the total number of objects in the large cluster and the total number of objects in all clusters. The definition of the Eq. (8) is based on the consideration that most objects in $RDSC^{'}$ are not CNVs. Thus, the clusters that contain most of the objects are considered large clusters. Eq. (9) signifies that the size of a large cluster is at least y times the size of a small cluster, and describes the difference in size between the smallest large cluster and the largest small cluster. In Eq. (10), $LRDSC^{'}$

represents the set of large clusters. In Eq. (11), $SRDSC^{'}$ represents the set of small clusters. Finally, based on the preceding definitions, Eq. (12) is constructed to describe the cluster-based abnormal score.

$$CBAS(O) = \begin{cases} min(dist(O, RDSC_i^{'}))O \in RDSC_j^{'}, RDSC_i^{'} \in LRDSC^{'}, RDSC_j^{'} \in \\ SRDSC^{'}, 1 \le i \le \theta, \theta < j \le k \\ dist(O, RDSC_i^{'}) \quad O \in RDSC_i^{'}, RDSC_i^{'} \in LRDSC^{'}, 1 \le i \le \theta \end{cases} \tag{12}$$

In Eq. (12), CBAS (*O*) represents the cluster-based abnormal score of object *O*, which is defined in two cases: (1) if the object *O* originates from a small cluster, the distance between *O* and the center of the closest large cluster is considered as the cluster-based abnormal score of *O*; (2) if the object *O* originates from a large cluster, the distance between *O* and the center of the large cluster is considered as the cluster-based abnormal score of *O*.

Predicting CNVs

Based on the cluster-based abnormal score profiles, the abnormal objects must be identified. For this step, the traditional methods analyze the abnormality of each object, and the users directly select an appropriate threshold to cut off the abnormal objects according to the application scenario. In the proposed method, Tukey's fences method is adopted to determine the abnormal objects. The prediction of abnormal objects consists of the following five steps. (1) the cluster-based abnormal scores of all objects are sorted from smallest to largest. (2) Eq. (13) is defined to evaluate an extreme outer limit:

$$T = CBAS_{Q_3} + w \cdot (CBAS_{Q_3} - CBAS_{Q_1}), \tag{13}$$

where T represents the upper limit of fences, w represents an abnormal weight, $CBAS_{Q_1}$ represents the cluster-based abnormal score of the lower quartile, and $CBAS_{Q_3}$ represents the cluster-based abnormal score of the upper quartile. (3) the basic notion of judging abnormal objects is that the higher the cluster-based abnormal score of an object, the more likely it is to be a CNV. Here, T is used as the baseline to identify abnormal objects. If the cluster-based abnormal score of an object is greater than T, it is considered to be a CNV. If the cluster-based abnormal score of an object is less than or equal to T, it is considered to be a normal area. (4) after the candidate CNVs are determined, their mutation modes (gain or loss) are determined. If the read depth value of a CNV area is greater than or equal to the mean read depth value of all normal areas, it is considered to be a gain area. If the read depth value of a CNV area is less than the mean read depth value of all normal areas, it is considered to be a loss area. (5) finally, somatic CNVs and germline CNVs are further identified. A germline CNV is a genetic variation that may originate from an individual's parents or family. If a CNV exists in both the tumor-normal matched samples, it is regarded as a germline CNV.

Parameter Setting of CBCNV

To effectively use CBCNV, it is necessary to further explain the settings of related parameters, which include the bin size, the number of neighbors (L), the number of clusters (k), and the ratios of large clusters (x), multiples (y), and abnormal weight (w). In this study, the bin size and L are set to 2,000 bp and 10 by default, respectively. Additionally, the values of k, x, and y are set to 5, 0.9, and 5, respectively, which are adopted by referencing published article (He et al., 2003). In Tukey's fences method, w is generally set to 1.5 (Zijlstra et al., 2007). In the proposed method, w is set to 1.5 as the default value. The settings of these default parameter values in the proposed method were determined according to experience and related methods. Users can also adjust these parameters according to their actual needs and application scenarios.

RESULTS

It is necessary to design a reasonable experimental plan to verify the effectiveness and reliability of the proposed method. Aiming at this point, simulation and real experiments were conducted. A simulation experiment is an effective and objective evaluation strategy, which can provide a comparison criterion to quantify the performance of the proposed method. In the simulation experiment, three popular published algorithms (BIC-seq2 (Xi et al., 2016), SeqCNV (Chen et al., 2017), and CNVkit (Talevich et al., 2016)) that can be used to effectively detect tumor-normal matched samples were selected for comparison with CBCNV. The performances of these methods are evaluated from three perspectives. First, the sensitivity and false discovery rate (FDR) of the four methods are evaluated at six CNV length levels. Then, the sensitivity and FDR of each method in the CNV gain and loss regions are analyzed and discussed. Finally, three indicators (recall, precision, and F1-score) are used to comprehensively evaluate the performance of each method. In real data applications, the proposed algorithm was used to detect two pairs of matched breast cancer whole-genome sequencing samples. Because the ground truths of the real data sets are unknown, the number of overlapping events and number of predicted events are adopted to evaluate the performance of each method. To further verify the performance of the proposed method, we use overlapping density score method to quantify performance of each method. The experimental results demonstrate that CBCNV is powerful CNV detection tools.

Application of Simulation Data

Many CNV simulation softwares have been developed and applied to generate next-generation sequencing data. In this study, IntSIM software was selected to generate simulation data sets (Yuan et al., 2017). Before its use, some settings were conducted: (1) the reference genome was prepared; (2) the tumor purity (TP) and sequencing coverage (SC) were set; (3) the number of repetitions of the sample under the configuration of each group was selected. Chromosome 21 of hg19 was entered into the software as a reference genome. The tumor purity was set to 0.2 and 0.3, and sequencing coverage was set to 10× to generate simulated data sets of different configurations, in which 50 samples were generated. Each sample was embedded with 22 regions of variation, which were composed of 12 gains and 10 losses (four heterogeneous losses and six homogeneous losses). The length of the CNV regions ranged from 2 to 100 kb. To fairly evaluate the performance of each method, the default parameters were used for all methods to detect each set of data.

Figure 2 describes the sensitivity of the four methods for the respective detection of CNVs with lengths of 2, 6, 10, 30, 50, and 100 kb under two different configurations, respectively. Two performance indicators (sensitivity and FDR) are adopted to evaluate the resolution of each method. Sensitivity is defined as the value of the number of CNVs correctly detected by a tool divided by the total number of CNVs recorded by the ground truth file. FDR is defined as the value of the number of false positives detected by a tool divided by the total number of CNVs detected by the tool. If a detected event overlaps with the ground truth file by more than 50%, it is considered as a candidate CNV (Hormozdiari et al., 2009). From the figure, it is evident that the sensitivity of each method increased with the increase in tumor purity from 0.2 to 0.3. This demonstrates that tumor purity is one of the key factors that affect CNV detection. In contrast, long CNVs were more easily detected by each method than short CNVs. CBCNV achieved the best sensitivity for all CNV length levels, and BIC-seq2 achieved better sensitivity than the other two methods (SeqCNV and CNVkit) at most CNV length levels. SeqCNV achieved the lowest sensitivity in the cases of CNVs with lengths of 50 and 100 kb, which indicates that it is not sensitive enough to detect long CNVs. CNVkit achieved the lowest sensitivity in the cases of CNVs with lengths of 2 and 6 kb, which indicates that it is not sensitive enough to detect short CNVs. **Figure 3** presents the FDR of each method at the six CNV length levels under two different configurations. In the case of tumor purity = 0.2, CNVkit performed the best in terms of FDR, followed by CBCNV, BIC-seq2 and SeqCNV. Although CNVkit achieved the best FDR, it had the lowest sensitivity. In the case of tumor purity = 0.3, CBCNV performed excellently in terms of FDR, followed by BIC-seq2, CNVkit, and SeqCNV. Considering the two indicators together, CBCNV achieved the best tradeoff between sensitivity and FDR, followed by BIC-seq2, SeqCNV, and CNVkit. Via the preceding analysis and discussion, it can be concluded that CBCNV can detect more CNVs with fewer false positives than the other three methods.

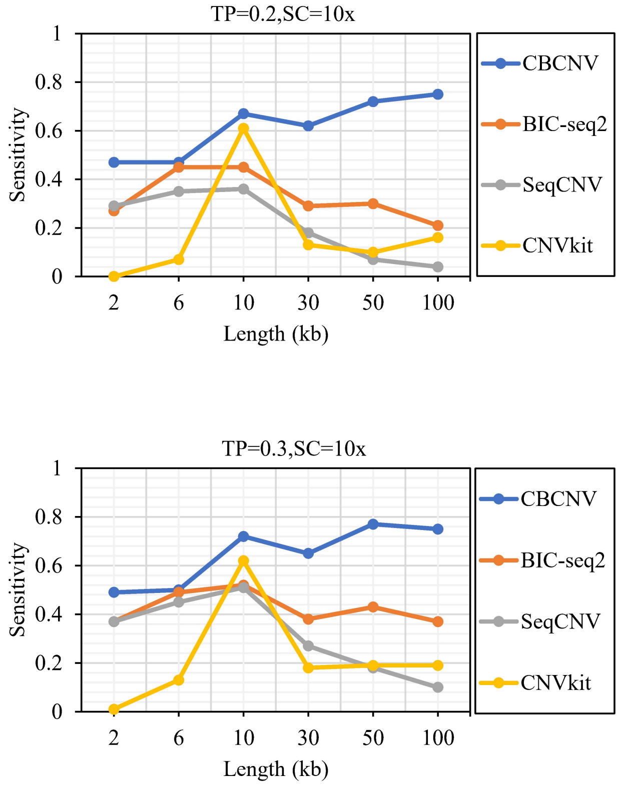


FIGURE 2 | The sensitivity of four methods at the six CNV length levels under two sets of simulation configurations.

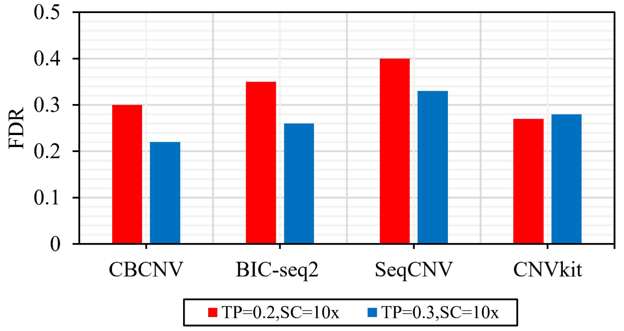


FIGURE 3 | The FDR of each method at the six CNV length levels under two sets of simulation samples.

Based on the simulated data sets, sensitivity and FDR were considered to analyze and evaluate the performances of the compared methods (CBCNV, BIC-seq2, SeqCNV, and CNVkit) in the gain and loss areas, and the averages of the two indicators were calculated across the 50 samples under different setting conditions. In general, the sensitivity of each method was found to increase with the increase in tumor purity, which demonstrates that the performance of each method was very sensitive to tumor purity. Most methods detected the CNV gain areas more sensitively than the CNV loss areas. **Figure 4** describes the sensitivity of each method to the detection of the gain and loss areas under two different sets of conditions. In each set of conditions, CBCNV achieved the highest sensitivity in the gain and loss areas. BIC-seq2 achieved better sensitivity in the gain areas than the other two methods (SeqCNV and CNVkit), and its sensitivity in the loss areas ranked third. The sensitivity of SeqCNV to the detection of the gain areas was between those of BIC-seq2 and CNVkit, and it was insensitive to the detection of the loss areas as compared to the other three methods. CNVkit achieved the lowest sensitivity in the gain areas, but its sensitivity ranked second in the loss areas, which indicates that it is suitable for detecting loss areas. **Figure 5** describes the FDR of each method in the detection of gain and loss areas under two different sets of conditions. When tumor purity was

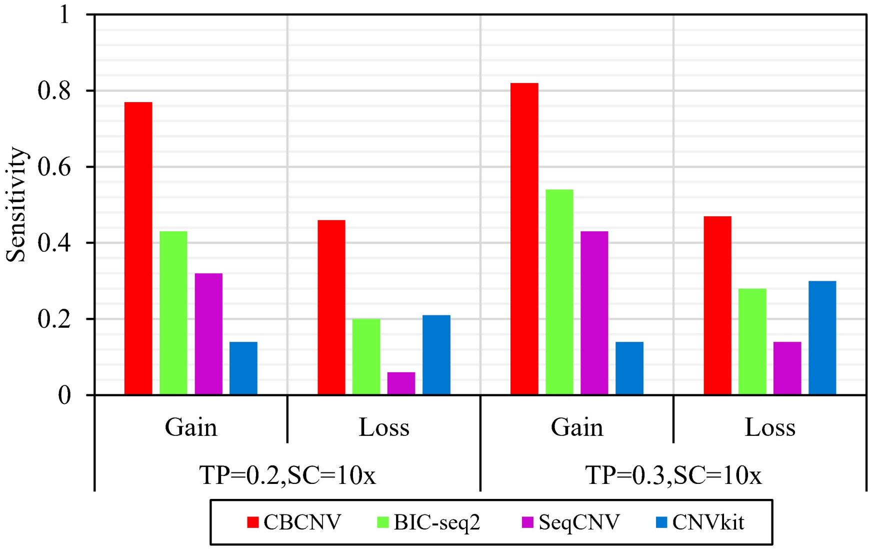


FIGURE 4 | Comparison of the sensitivity of the four methods for detecting gain and loss areas under two sets of simulation settings.

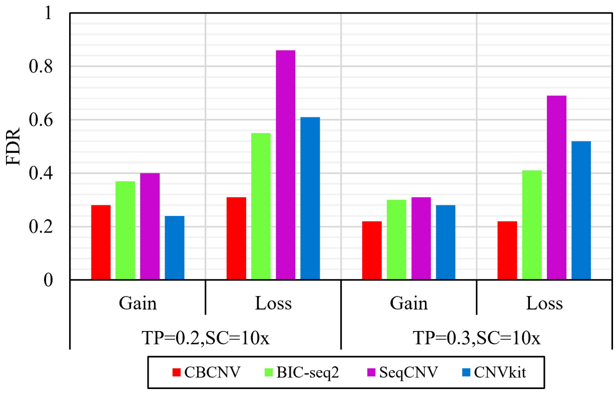


FIGURE 5 | Comparison of the FDR of each method for detecting gain and loss areas under two sets of simulation samples.

equal to 0.2 and 0.3, the FDR of CBCNV ranked second and first in the gain areas, and ranked first in the loss areas. The FDR of BIC-seq2 ranked third and second in the gain and loss areas, respectively. SeqCNV exhibited the largest FDRs in the gain and loss areas, which demonstrates that there were many false positives of the detected CNVs. CNVkit had a medium FDR between those of BIC-seq2 and SeqCNV in the loss areas. In the gain areas, the FDR of CNVkit ranked first when tumor purity = 0.2, and second when tumor purity = 0.3. CNVkit detected the gain areas more accurately than the loss areas. This demonstrates that the DOC-based method detected the gain areas more sensitively than the loss areas. In summary, CBCNV achieved the best tradeoff between sensitivity and FDR in the detection of gain and loss areas.

Three indicators (recall, precision, and F1-score) were adopted to comprehensively evaluate the performance of each method. Recall is defined as the number of correctly detected CNVs divided by the total number of simulated CNVs (Magi et al., 2013). Precision is defined as the number of correctly detected CNVs divided by the total number of detected CNVs (Magi et al., 2013). The F1-score represents the harmonic mean of precision and recall. The three performance indicators are reported as the averages of 50 samples under each set of conditions. **Figure 6** detail the F1-score level of each method, from which it is evident that CBCNV achieved the highest recall, followed by BIC-seq2, SeqCNV, and CNVkit. When tumor purity = 0.3, CBCNV got the best precision rate among all methods. When tumor purity = 0.2, CNVkit performed the best in terms of precision, followed by CBCNV, BIC-seq2, and SeqCNV. Moreover, CBCNV achieved the best tradeoff between precision and recall, followed by BIC-seq2, SeqCNV, and CNVkit, which is consistent with the above experimental results.

Detection of Copy Number Variants From Breast Cancer Samples

To analyze and verify the performance of the proposed method, it was applied to detect four paired whole-genome breast cancer samples (PD4088a, PD4088b, PD4192a, and PD4192b), the details of which were sourced from https://www.ebi.ac.uk/ega/studies under accession EGAS00001000170 (Li Y. Y. et al., 2019). CBCNV was used to detect 22 autosomes in each set of samples, and two well-known methods (BIC-seq2 and CNVkit) were selected for comparison. For a fair comparison, the default parameters were used for these methods to detect the samples. The number of overlapping events and predicted events were used for performance measurement to effectively analyze the advantages and disadvantages of each method. The ground truth file could not be provided in the real data experiment. The number of overlapping events represent the average number of overlapping events for one method and other methods, and

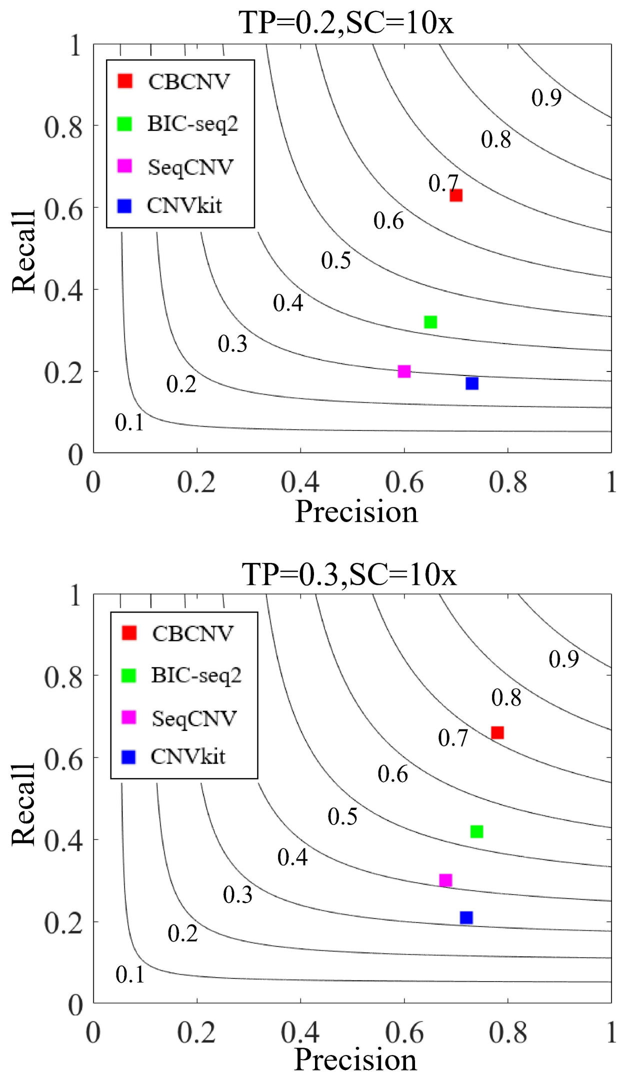


FIGURE 6 | The performance comparison of the four methods in terms of recall, precision and F1-score under two sets of simulation settings. Black curves indicate F1-score levels, which ranges from 0.1 to 0.9. The equations on the left and right sides of the comma represent the tumor purity (TP) and sequencing coverage (SC), respectively.

number of predicted events represents the total number of events predicted by a method. **Table 1** presents the number of overlapping events and predicted events of each method in the 22 autosomes of samples PD4088a and PD4192a, respectively. In the sample PD4088a, it is evident that CBCNV achieved the greatest number of overlapping events and predicted events. BIC-seq2 detects the least number of overlapping events and predicted events, which shows that it is more conservative than the other two methods. CNVkit achieved number of overlapping events and predicted events between CBCNV and BIC-seq2. In the sample PD4192a, CNVkit called a large number of CNV events, but obtained number of overlapping events as many as CBCNV, which means it has detected a large number of non-overlapping events. It indirectly shows that the CNVs detected by CNVkit are likely to contain a large number of false positive events. A small number of overlapping events and predicted events were found by BIC-seq2, performance of which is consistent with the above sample. The number of events detected by CBCNV is between the other two methods, but it obtains the most overlapping events, which fully shows that most of the CNV events detected by CBCNV are true positive events.

In order to further verify the performance of each method, we adopt the evaluation method of overlapping density score (ODS) (Yuan et al., 2018), which is defined by the following equation.

$$ODS = O_m \cdot O_r, \tag{14}$$

Where O_m represents the mean number of overlapping events between one method and other comparison methods, O_r represents O_m divided by the total number of CNV events detected by the method. Here, we use Eq. (14) to calculate ODS for each method, and the comparison results are recorded in **Table 2**. From the experimental results, CBCNV achieve the best ODS in the all samples. ODS of BIC-seq2 are the lowest among all methods. Compared with the above two methods, CNVkit obtain the medium ODS in each group of samples. Overall, CBCNV achieved the best balance between O_m and O_r as compared to the other two methods.

On the basis of the above experiments, we used the catalog of somatic mutations in cancer (COSMIC) database to analyze the biological significance of the detected CNVs. From two pairs of matched breast cancer samples, we found that some of the detected CNVs contained some genes that were related to breast cancer, such as PDZK1 (Kim et al., 2013), XRCC4 (Allen-Brady et al., 2006), Fbxl17 (Mason et al., 2020), ITGBL1 (Li et al., 2015), RORA (Taheri et al., 2017), BAGE (Fujie et al., 1997), AMOTL1 (Couderc et al., 2016), RAP80 (Osorio et al., 2009), PIWIL4 (Wang et al., 2016), CSE1L (Behrens et al., 2001), and USP18 (Tan et al., 2018).

Evaluation of Running Time

Running time is a critical evaluation indicator to evaluate the performance of the methods. For this, CBCNV and the other three methods (BIC-seq2, SeqCNV, and CNVkit) are tested on 50 simulation samples, which are run on a personal computer with

TABLE 1 | Comparison of number of overlapping events (NOE) and predicted events (NPE) for each method on two sets of real samples.

Sample	Indicator	CBCNV	BIC-seq2	CNVkit
PD4088a	NOE	80	19	49
	NPE	510	85	194
PD4192a	NOE	126	20	126
	NPE	482	83	2,156

TABLE 2 | Comparison of ODS for each method on two sets of real samples.

Sample	CBCNV	BIC-seq2	CNVkit
PD4088a	19	6	18
PD4192a	43	7	32

TABLE 3 | Comparison of running time for each method.

Indicator	CBCNV	BIC-seq2	SeqCNV	CNVkit
Running time (s)	39	8	500	182

Intel(R) Core (TM) i7-4710MQ CPU @ 2.50 GHz and 16.0 GB memory. The running time of each method is counted as the averages of 50 simulation samples. As shown in **Table 3**, BIC-seq2 performed the best in terms of running time, followed by CBCNV, CNVkit, and SeqCNV, which shows that CBCNV is a relatively efficient CNV detection tool.

DISCUSSION AND CONCLUSION

In this work, the proposed CBCNV method was developed based on DOC profiles to detect CNVs using next-generation sequencing data, and is suitable for the detection of tumor-normal matched samples. CBCNV uses a local perspective to capture abnormal read depth signals, which are considered to be only a small portion of the overall signals. Its detection concept is different from those of traditional CNV detection methods, which generally construct a statistical model by fitting the read depth signals, then select a reasonable baseline to identify CNVs. Instead, in CBCNV, a clustering algorithm is performed on the read depth segment profile to form clusters of different scales. According to the scales of the clusters, large and small clusters are defined. If a read depth segment originates from a large cluster, its abnormal score is defined as the distance between the read depth segment and the cluster center. If a read depth segment belongs to a small cluster, its abnormal score is defined as the distance between the read depth segment and the center of the closest large cluster. In this way, an abnormal score is assigned to each read depth segment. Based on the abnormal score profile, Tukey's fences method is adopted to predict CNVs (Zijlstra et al., 2007).

Via the analysis of the concepts of the proposed method, the following characteristics are summarized. (1) CBCNV extracts two features of read depth signals, which fully considers the copy number of each read depth segment and the difference in the ratios of adjacent read depth segments. (2) The traditional outlier detection algorithm was effectively converted to detect CNVs. CBCNV uses a local perspective to identify CNVs, which can objectively reflect the actual state of abnormal read depth signals. It does not require the fitting of the distribution of read depth signals, and cluster-based abnormal scores are constructed for each read depth segment signal to effectively identify the copy number status of adjacent read depth signals. (3) Based on the abnormal score of each read depth segment, Tukey's fences method is applied to identify CNVs, which does not require the evaluation of the distribution of abnormal scores.

Simulated data sets were used to evaluate the performance of CBCNV, and three popular algorithms were selected for comparison. First, the sensitivity and FDR of each method for the detection of CNVs of different lengths and in gain and loss regions were analyzed and discussed. Via the analysis of the experimental results, it was found that CBCNV achieved the best tradeoff between sensitivity and FDR. Second, three performance indicators (recall, precision, and F1-score) were adopted to comprehensively evaluate the performance of each method. The experimental results proved that CBCNV achieved the best performance in terms of all three indicators. In real data applications, two sets of whole-genome data were used to evaluate the effectiveness of the proposed method. The experimental results demonstrated that CBCNV achieved the best number of overlapping events and overlapping density scores compared to the other two methods in each group of samples. In summary, CBCNV is an effective and reliable CNV detection tool for using on tumor-normal matched samples.

Some shortcomings of the proposed method were also discovered. For example, the selection of the number of clusters (k) is a very important step that may affect the accuracy of the results. In most application scenarios, the performance of the proposed method was superior under this set of parameter settings, which meets the needs of users in most cases. However, in some unique cases, the performance of this set of parameters may not be suitable. In future research, the data size and characteristics will be fully considered to automatically set the parameter k. In addition, in the present study, only two features of read depth were extracted as the input. In future research, multiple factors of read depth signals will be considered to improve the accuracy of the proposed method. Ultimately, CBCNV will be further expanded (Mao et al., 2021) and proved to effectively detect other types of structural variation in multiple application scenarios.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/ega/studies, EGAS00001000170.

AUTHOR CONTRIBUTIONS

GL participated in the design of the algorithms and the experiments. JZ participated in the design of the entire framework of CNV detection and directed the whole work, and helped to revise the manuscript. Both authors read the final manuscript and agreed on its contents for submission.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grants 91853123 and 11674352.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.699510/full#supplementary-material

REFERENCES

Adam, S., and David, M. (2009). Copy number variations and cancer. *Genome Med.* 1, 62. doi: 10.1186/gm62

Allen-Brady, K., Cannon-Albright, L., Neuhausen, S., and Camp, N. (2006). A role for XRCC4 in age at diagnosis and breast cancer risk. *Cancer Epidemiol. Biomarkers Prevent.* 15, 1306–1310. doi: 10.1158/1055-9965.EPI-05-0959

Behrens, P., Brinkmann, U., Fogt, F., Wernert, N., and Wellmann, A. (2001). Implication of the proliferation and apoptosis associated CSE1L/CAS gene for breast cancer development. *Anticancer Res.* 21, 2413–2417.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899–905. doi: 10.1038/nature08822

Buysse, K., Chiaie, B. D., Coster, R. V., Loeys, B., Paepe, A. D., Mortier, G., et al. (2009). Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *Eur. J. Med. Genet.* 52, 398–403. doi: 10.1016/j.ejmg.2009.09.002

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* 39, S16–S21. doi: 10.1038/ng2028

Chen, Y., Zhao, L., Wang, Y., Cao, M., Gelowani, V., Xu, M., et al. (2017). SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data. *BMC Bioinform.* 18:147. doi: 10.1186/s12859-017-1566-3

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., et al. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103. doi: 10.1038/nmeth.1276

Cook, E., and Scherer, S. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature* 455, 919–923. doi: 10.1038/nature07458

Couderc, C., Boin, A., Fuhrmann, L., Vincent-Salomon, A., Mandati, V., Kieffer, Y., et al. (2016). AMOTL1 promotes breast cancer progression and is antagonized by merlin. *Neoplasia* 18, 10–24. doi: 10.1016/j.neo.2015.11.010

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., Mccarroll, S. A., Altshuler, D. M., et al. (2006). Copy number variation: new insights in genome diversity. *Genome Res.* 16, 949–961. doi: 10.1101/gr.3677206

Fujie, T., Mori, M., Ueo, H., Sugimachi, K., and Akiyoshi, T. (1997). Expression of MAGE and BAGE genes in Japanese breast cancers. *Ann. Oncol.* 8, 369–372. doi: 10.1023/A:1008255630202

Hartigan, J. A., and Wong, M. A. (1979). A K-Means clustering algorithm. *J. R. Stat. Soc.* 28, 100–108. doi: 10.2307/2346830

He, Z. Y., Xu, X. F., and Deng, S. C. (2003). Discovering cluster-based local outliers. *Pattern Recognition Lett.* 24, 1641–1650. doi: 10.1016/S0167-8655(03)00003-5

Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* 19, 1270–1278. doi: 10.1101/gr.088633.108

Kim, H., Abd Elmageed, Z., Ju, J., Naura, A., Abdel-Mageed, A., Varughese, S., et al. (2013). PDZK1 is a novel factor in breast cancer that is indirectly regulated by Estrogen through IGF-1R and promotes estrogen-mediated growth. *Mol. Med.* 19, 253–262. doi: 10.2119/molmed.2011.00001

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318, 420–426. doi: 10.1126/science.1149504

Krepischi, A., Pearson, P. L., and Rosenberg, C. (2012). Germline copy number variations and cancer predisposition. *Future Oncol.* 8, 441–450. doi: 10.2217/fon.12.34

Kuiper, R. P., Ligtenberg, M. J., Hoogerbrugge, N., and Kessel, A. G. V. (2010). Germline copy number variation and cancer risk. *Curr. Opin. Genet. Dev.* 20, 282–289. doi: 10.1016/j.gde.2010.03.005

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, X.-Q., Du, X., Li, D.-M., Kong, P.-Z., Sun, Y., Liu, P.-F., et al. (2015). ITGBL1 Is a Runx2 transcriptional target and promotes breast cancer bone metastasis by activating the TGFβ signaling pathway. *Cancer Res.* 75, 3302–3313. doi: 10.1158/0008-5472.CAN-15-0240

Li, Y., Yuan, X., Zhang, J., Yang, L., Bai, J., and Jiang, S. (2019). SM-RCNV: a statistical method to detect recurrent copy number variations in sequenced samples. *Genes Genomics* 41, 529–536. doi: 10.1007/s13258-019-00788-9

Li, Y. Y., Zhang, J. Y., and Yuan, X. G. (2019). BagGMM: calling copy number variation by bagging multiple Gaussian mixture models from tumor and matched normal next-generation sequencing data. *Digital Signal Processing* 88, 90–100. doi: 10.1016/j.dsp.2019.01.025

Liu, G. J., Zhang, J. Y., Yuan, X. G., and Wei, C. (2020). RKDOSCNV: a local kernel density-based approach to the detection of copy number variations by using next-generation sequencing data. *Front. Genet.* 11:569227. doi: 10.3389/fgene.2020.569227

Magi, A., Pippucci, T., and Sidore, C. (2017). XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genomics* 18:747. doi: 10.1186/s12864-017-4137-0

Magi, A., Tattini, L., Cifola, I., D'Aurizio, R., Benelli, M., Mangano, E., et al. (2013). EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* 14:R120. doi: 10.1186/gb-2013-14-10-r120

Malek, J. A., Mery, E., Mahmoud, Y. A., Al-Azwani, E. K., Roger, L., Huang, R., et al. (2011). Copy number variation analysis of matched ovarian primary tumors and peritoneal metastasis. *PLoS One* 6:e28561. doi: 10.1371/journal.pone.0028561

Mao, Y. F., Yuan, X. G., and Cun, Y. P. (2021). A novel machine learning approach (svmSomatic) to distinguish somatic and germline mutations using next-generation sequencing data. *Zool. Res.* 42, 246–249. doi: 10.24272/j.issn.2095-8137.2021.014

Mason, B., Flach, S., Teixeira, F., Garcia, R., Rueda, O., Abraham, J., et al. (2020). Fbxl17 is rearranged in breast cancer and loss of its activity leads to increased globalO-GlcNAcylation. *Cell. Mol. Life Sci.* 77, 2605–2620. doi: 10.1007/s00018-019-03306-y

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* 6, S13–S20. doi: 10.1038/nmeth.1374

Osorio, A., Barroso, A., Garcia, M., Martinez-Delgado, B., Urioste, M., and Benitez, J. (2009). Evaluation of the BRCA1 interacting genes RAP80 and CCDC98 in familial breast cancer susceptibility. *Breast Cancer Res. Treatment* 113, 371–376. doi: 10.1007/s10549-008-9933-4

Pei, G., Hu, R., Dai, Y., Manuel, A., Zhao, Z., and Jia, P. (2021a). Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations. *Nucleic Acids Res.* 49, 53–66. doi: 10.1093/nar/gkaa1137

Pei, G., Hu, R., Jia, P., and Zhao, Z. (2021b). DeepFun: a deep learning sequence-based model to decipher non-coding variant effect in a tissue- and cell type-specific manner. *Nucleic Acids Res. [online ahead of print]* gkab429. doi: 10.1093/nar/gkab429

Pei, G., Hu, R., Dai, Y., Zhao, Z., and Jia, P. (2020). Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network. *Oncogene* 39, 5031–5041. doi: 10.1038/s41388-020-1343-z

Sebat, J., Lakshmi, B., Malhotra, D., and Troge, J. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445–449. doi: 10.1126/science.1138659

Sharp, A. J., Locke, D. P., McGrath, S. D., and Cheng, Z. (2005). Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* 77, 78–88. doi: 10.1086/431652

Stone, J. L., O'Donovan, M. C., Gurling, H., and Kirov, G. K. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455, 237–241. doi: 10.1038/nature07239

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* 458, 719–724. doi: 10.1038/nature07943

Taheri, M., Omrani, M. D., Noroozi, R., Ghafouri-Fard, S., and Sayad, A. (2017). Retinoic acid-related orphan receptor alpha (RORA) variants and risk of breast cancer. *Breast Dis.* 37, 21–25. doi: 10.3233/BD-160248

Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* 12:e1004873. doi: 10.1371/journal.pcbi.1004873

Tan, Y., Zhou, G., Wang, X., Chen, W., and Gao, H. (2018). USP18 promotes breast cancer growth by upregulating EGFR and activating the AKT/Skp2 pathway. *Int. J. Oncol.* 53, 371–383. doi: 10.3892/ijo.2018.4387

Tchatchou, S., and Burwinkel, B. (2008). Chromosome copy number variation and breast cancer risk. *Cytogenetic Genome Res.* 123, 183–187. doi: 10.1159/000184707

Tibshirani, R., and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* 9, 18–29. doi: 10.1093/biostatistics/kxm013

Wang, Z., Liu, N., Shi, S., Liu, S., and Lin, H. (2016). The role of PIWIL4, an argonaute family protein, in breast cancer. *J. Biol. Chem.* 291, 10646–10658. doi: 10.1074/jbc.M116.723239

Xi, R. B., Lee, S., Xia, Y. C., Kim, T. M., and Park, P. J. (2016). Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* 44, 6274–6286. doi: 10.1093/nar/gkw491

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Yuan, X. G., Bai, J., Zhang, J. Y., Yang, L., Duan, J., Li, Y., et al. (2018). CONDEL: detecting copy number variation and genotyping deletion zygosity from single tumor samples using sequence data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1141–1153. doi: 10.1109/TCBB.2018.2883333

Yuan, X. G., Zhang, J. Y., and Yang, L. Y. (2017). IntSIM: an integrated simulator of next-generation sequencing data. *IEEE Trans. Biomed. Eng.* 64, 441–451. doi: 10.1109/TBME.2016.2560939

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). PyOD: a Python toolbox for scalable outlier detection. *J. Machine Learn. Res.* 20:96.

Zijlstra, W. P., van der Ark, L. A., and Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behav. Res.* 42, 531–555. doi: 10.1080/00273170701384340

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fgene.2021.697279

Investigating Different DNA Methylation Patterns at the Resolution of Methylation Haplotypes

Xiaoqing Peng[1]*, *Yiming Li*[1], *Xiangyan Kong*[2], *Xiaoshu Zhu*[3] *and Xiaojun Ding*[3]*

[1] *Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, China,* [2] *School of Computer Science and Engineering, Central South University, Changsha, China,* [3] *School of Computer Science and Engineering, Yulin Normal University, Yulin, China*

Different DNA methylation patterns presented on different tissues or cell types are considered as one of the main reasons accounting for the tissue-specific gene expressions. In recent years, many methods have been proposed to identify differentially methylated regions (DMRs) based on the mixture of methylation signals from homologous chromosomes. To investigate the possible influence of homologous chromosomes on methylation analysis, this paper proposed a method (MHap) to construct methylation haplotypes for homologous chromosomes in CpG dense regions. Through comparing the methylation consistency between homologous chromosomes in different cell types, it can be found that majority of paired methylation haplotypes derived from homologous chromosomes are consistent, while a lower methylation consistency was observed in the breast cancer sample. It also can be observed that the hypomethylation consistency of differentiated cells is higher than that of the corresponding undifferentiated stem cells. Furthermore, based on the methylation haplotypes constructed on homologous chromosomes, a method (MHap_DMR) is developed to identify DMRs between differentiated cells and the corresponding undifferentiated stem cells, or between the breast cancer sample and the normal breast sample. Through comparing the methylation haplotype modes of DMRs in two cell types, the DNA methylation changing directions of homologous chromosomes in cell differentiation and cancerization can be revealed. The code is available at: https://github.com/xqpeng/MHap_DMR.

Keywords: methylation haplotype, differentially methylated region, cell differentiation, homologous chromosomes, methylation consistency, hypomethylation consistency

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Hao Lin,
University of Electronic Science and Technology of China, China
Leyi Wei,
Shandong University, China

Correspondence:
Xiaoqing Peng
xqpeng@csu.edu.cn
Xiaojun Ding
ding.xiaojun@foxmail.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *19 April 2021*
Accepted: *01 June 2021*
Published: *28 June 2021*

Citation:
Peng X, Li Y, Kong X, Zhu X and Ding X (2021) Investigating Different DNA Methylation Patterns at the Resolution of Methylation Haplotypes. Front. Genet. 12:697279. doi: 10.3389/fgene.2021.697279

1. INTRODUCTION

In recent years, the revealing of the mechanisms behind the diseases has been performed from different angles, such as mutated genes, altered DNA methylation (Eden et al., 2003; Baylin, 2005), non-coding RNAs (Yan et al., 2017, 2018; Chen et al., 2019; Lan et al., 2020), microbes (Yan et al., 2019, 2021), etc. Differentially methylated regions (DMRs) are the main explanation accounting for the diversity of gene expression in different cell types in a body. Differentiation-associated differential methylation profiles were observed on cell types under different stages of development and differentiation (Laurent et al., 2010). Recent studies show that altered DNA methylation has a very close relationship with diseases. In cancer genomes, the promoter regions of tumor suppressor genes are altered to be hypermethylated (Baylin, 2005), while the promoter regions of tumor genes are altered to be hypomethylated (Eden et al., 2003). Identifying DMRs can promote revealing the mechanisms in tissue-specific/diseases-specific gene expression (Scott et al., 2020) and tissue-specific DMRs can be used as feature markers in identifying the tissues-of-origin of cfDNAs in noninvasive diagnosis (Hu et al., 2019; Xiaoqing et al., 2020).

Infinium HumanMethylation450 BeadChip and Infinium MethylationEPIC BeadChip provide a convenient way to measure the methylation levels of CpG sites in CpG islands and gene regions. In BreadChips, the methylation level of a certain CpG site is estimated by using the ratio of intensities between methylated and unmethylated alleles. In recent years, due to the development of sequencing technology, bisulfite sequencing (BS-Seq) makes to reveal the methylation status of each cytosine site on a read become possible. The numbers of methylated cytosines and unmethylated ones of each cytosine site can be measured, respectively. Recently, by using deep-learning, DNA methylation status of each cytosine site can be deduced from Nanopore sequencing reads (Ni et al., 2019). In both BeadChip and BS-Seq, molecules derived from two homologous chromosomes are not discriminated and are always processed together.

Based on the methylation profiles extracted from BeadChips or BS-Seq data, many methods have been proposed to identify DMRs in different tissues or cell types. These methods can be roughly divided into two categories: differentially methylated cytosine site (DMC)-based methods and candidate region-based methods. In DMC based methods, methylation levels of CpG sites can be calculated based the raw methylation information of CpG sites (Catoni et al., 2018; Condon et al., 2018; Xu et al., 2020), estimated by beta-binomial distribution considering the biological variances and sample variances (Feng et al., 2014; Park et al., 2014; Lea et al., 2015; Wu et al., 2015; Park and Wu, 2016; Wen et al., 2016) or estimated by considering the spatial correlation (Hansen et al., 2012; Hebestreit et al., 2013; Wu et al., 2015; Sun and Yu, 2016). Then, DMCs are identified and DMRs are formed by merging the neighboring DMCs satisfying some defined criteria, such as DMCs with p-values less than a certain threshold, the distance between the DMCs less than a cutoff value, and the minimum number of DMCs required in a region.

In candidate region-based methods, there are two types of candidate regions, including sample-dependent candidate regions and sample-independent ones. The sample-independent candidate regions are predefined on the genome with a fixed-size or sliding window (Stockwell et al., 2014; Wang et al., 2015; Catoni et al., 2018). The sample-dependent candidate regions are generated according to the coverage, the depth of CpG sites, the methylation levels of CpG sites in samples, and the methylation changes of CpG sites among multi-samples. Then DMRs are identified by comparing the methylation of regions among different samples (Su et al., 2012; Lokk et al., 2014; Liu et al., 2015; Jühling et al., 2016; Gomez et al., 2019).

As we known, the allele-specific methylation is a special phenomenon of DNA methylation, which is that the methylation of an allele on two homologous chromosomes is not consistent. With the development of high-throughput sequencing technology, the region capture based sequencing and the genome-wide sequencing have been widely used for detecting allele-specific methylation sites. Some strategies and algorithms also contribute to improve the identification of allele-specific methylation (Cheung et al., 2017; Abante et al., 2020). However, the research on identifying allele-specific methylation is limited to the alleles, and the influence of homologous chromosomes on methylation analysis should be investigated genome wide.

In the methods of identifying DMRs, the reads from homologous chromosomes are processed together, and the methylation levels of CpG sites are calculated based on the mixture of reads from homologous chromosomes. The influence of homologous chromosomes on methylation analysis was not considered and investigated. To investigate the possible influence of homologous chromosomes on methylation analysis, we construct methylation haplotypes for homologous chromosomes on the sample-independent candidate regions. Then the methylation consistency of paired methylation haplotypes from homologous chromosomes in different cell types is compared. Further, DMRs are identified at the resolution of methylation haplotypes. The proposed method in this paper not only can be applied to methylation analysis, but also can provide a clear explanation for the methylation difference at the resolution of methylation haplotypes.

2. MATERIALS AND METHODS

In this paper, two methods, MHap and MHap_DMR, are proposed to construct methylation haplotypes and identify DMRs based on methylation haplotypes, respectively. MHap is a method for constructing methylation haplotypes, which consists of four steps. Firstly, it generates sample-independent candidate regions based on genomic information, such as CpG islands and CpG dense regions. Then, for the BS-seq data of each sample, it classifies CpG sites into homozygous and heterozygous ones, and then constructs methylation haplotypes for each candidate region. Finally, the paired methylation haplotypes of homologous chromosomes are represented by methylation haplotype modes (MHMs). MHap_DMR is the method designed to identify DMRs based on methylation haplotypes. The framework of MHap and MHap_DMR is shown in **Figure 1** and the detail of each step in the proposed methods will be described in the following subsections.

2.1. Materials

To investigate the influence of homologous chromosomes on methylation analysis, 12 WGBS datasets of 10 different tissues/cell types are involved in this study, including two lower leg skin samples and two tibial nerver samples downloaded from the ENCODE project (The ENCODE Project Consortium, 2012) (access sample id: ENCSR930WUY, ENCSR128RMY, ENCSR752OCM, and ENCSR658MZU), breast cancer sample and normal breast sample in the GEO database under the accession number GSE29069 (Hon et al., 2012), adipose-derived stem (ADS) cells and mature fat cells (adipocytes derived from the ADS cells) in the NCBI SRA database under the accession number SRA023829.2 (Lister et al., 2011), embryonic stem cells (hESCs) and foreskin fibroblasts (hESC-Fibro cells) in the GEO database under the accession number GSE19418 (Laurent et al.,

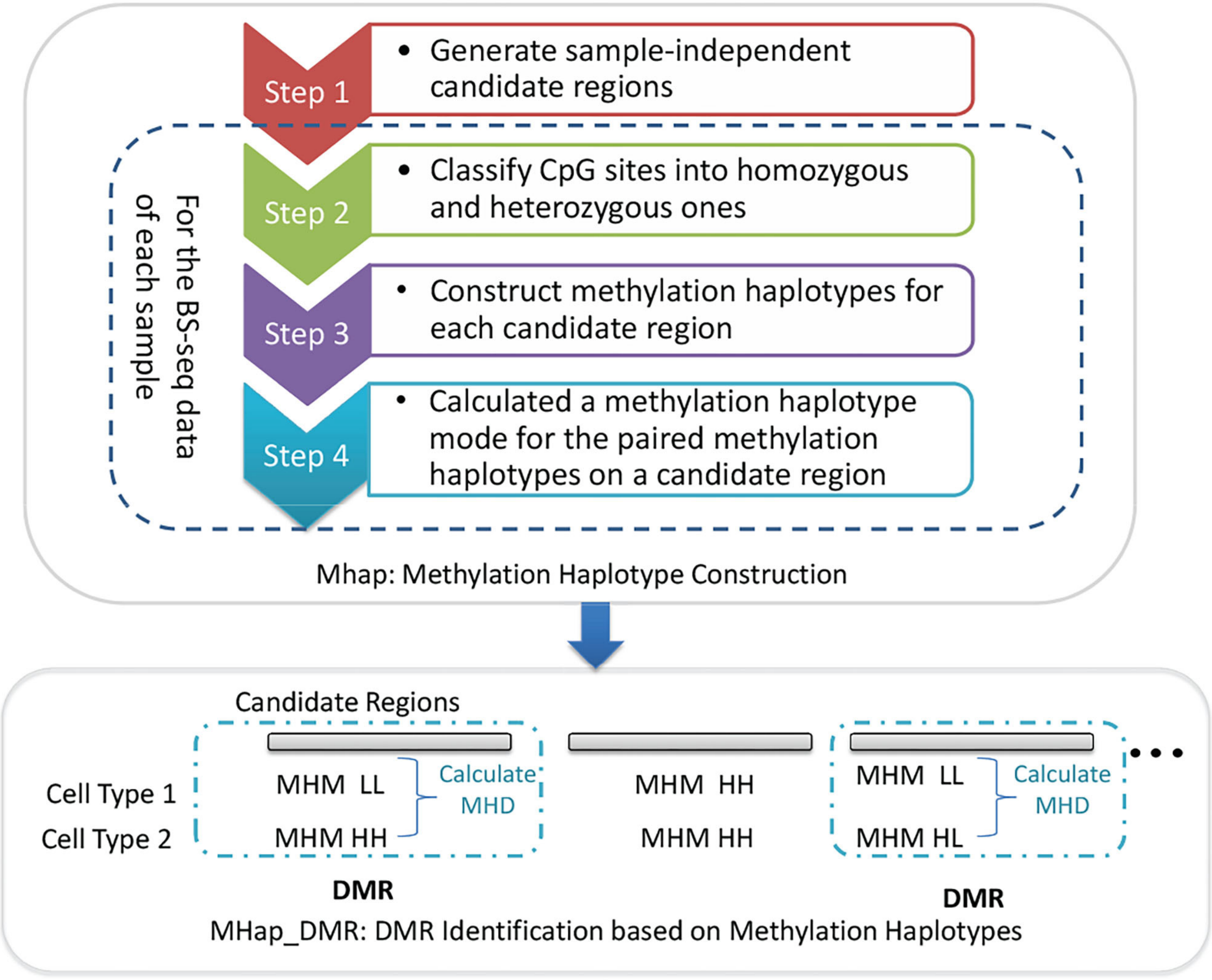

Step 1
• Generate sample-independent candidate regions
Step 2
• Classify CpG sites into homozygous and heterozygous ones
Step 3
• Construct methylation haplotypes for each candidate region
Step 4
• Calculated a methylation haplotype mode for the paired methylation haplotypes on a candidate region
For the BS-seq data of each sample
Mhap: Methylation Haplotype Construction
Candidate Regions
Cell Type 1
Cell Type 2
MHM LL
MHM HH
Calculate MHD
MHM HH
MHM HH
MHM LL
MHM HL
Calculate MHD
DMR
DMR
MHap_DMR: DMR Identification based on Methylation Haplotypes

FIGURE 1 | The framework of MHap and MHap_DMR.

2010), mature B cells and hematopoietic stem cells in the GEO database under the accession number GSE31971 (Hodges et al., 2011). The WGBS datasets were aligned to the human reference genome (hg38) and the methylation statuses of cytosines on reads were called by using Bismark (Krueger and Andrews, 2011).

2.2. MHap: Methylation Haplotype Construction

Due to the limited read lengths and the uneven distribution of CpG sites, it is challenging to construct two complete methylation haplotypes for two homologous chromosomes. Thus, sample-independent candidate regions are predefined on CpG dense regions, and methylation haplotypes are constructed for homologous chromosomes in these regions. MHap is proposed to construct methylation haplotypes for homologous chromosomes based on the overlapping methylation statuses of heterozygous methylated CpG sites on reads. The details of MHap is described as following.

2.2.1. Generate Sample-Independent Candidate Regions

MHap generates sample-independent candidate regions based on the CpG island information and the distance between neighboring CpG sites. In order not to hide local methylation signals, CpG islands are usually divided into a number of candidate regions, each of which contains at least 7 CpG sites. For other regions with densely located CpG sites, a distance-based clustering algorithm is applied to generating candidate regions, which contains at least 7 CpG sites also and the distances between neighboring CpG sites are not >20 bp. As shown in **Table 1**, for each chromosome, the number of candidate regions and the corresponding averages of CpG numbers and region lengths are listed. Then, MHap will construct methylation haplotypes for

TABLE 1 | The number, the average number of CpGs, and the length of candidate regions in each chromosome.

Chromosome	Num. of candidate regions	Ave. Num. of CpGs	Ave. length of candidate regions
chr1	26,643	10.48	92.11
chr2	20,446	10.46	91.71
chr3	14,013	10.46	93.08
chr4	13,316	10.49	95.06
chr5	14,411	10.46	93.70
chr6	14,378	10.49	93.86
chr7	16,809	10.42	92.33
chr8	12,429	10.43	92.64
chr9	13,798	10.44	91.80
chr10	13,482	10.46	91.90
chr11	14,194	10.43	91.38
chr12	13,107	10.42	93.56
chr13	7,503	10.41	93.78
chr14	9,217	10.44	91.07
chr15	9,173	10.49	89.84
chr16	14,837	10.36	91.21
chr17	17,285	10.45	92.24
chr18	6,419	10.55	92.72
chr19	20,663	10.51	95.05
chr20	8,844	10.40	90.39
chr21	5,597	10.70	92.66
chr22	8,153	10.38	87.39
chrX	10,687	10.42	97.36
chrY	1,982	10.31	103.24

homologous chromosomes on these candidate regions. of the candidate regions.

2.2.2. Classify CpG Sites Into Homozygous and Heterozygous Ones

The flow char of classifying CpG sites into homozygous and heterozygous ones is illustrated as in **Figure 2**. For each sample, the reads falling in candidate regions are collected. In these candidate regions, firstly, CpG sites with depth less than a threshold Th_{dp} are filtered out. Then the remaining CpG sites are classified into homozygous sites and candidate heterozygous sites(CHSs) based on the types of methylation statuses and the corresponding depths. If a CpG site has only one methylation status with depth not less than Th_{dp}, it is considered as a homozygous site. If it has two methylation statuses and the depth of each status is not less than half of Th_{dp}, it is considered as a CHS.

Due to the sequencing errors and the bisulfite conversion rates, the identified CHSs inevitably contain false-positives. The joint methylation statuses of neighboring CHSs on the same reads can help to distinguish true-positives from false-positives. Thus, the joint methylation statuses of two neighboring CHSs on the covering reads are extracted and can be represented as 00/11/01/10 patterns. In MHap, the frequency of each pattern on two neighboring CHSs is calculated, and patterns with frequency <2 are filtered. Then, one or two true-positive patterns are identified according to the ratios of the corresponding frequencies to the total frequency of all patterns or to the maximum frequency. If there is a pattern with the maximum frequency among other patterns and the ratio of its frequency to the total frequency of all patterns is above a threshold (recommended as 0.6), it is considered as the only one true-positive pattern on the two neighboring CHSs. Otherwise, if there are two patterns with higher frequencies than other patterns and the ratio of the second maximum frequency to the first maximum frequency is not less than a threshold (recommended as 0.4), it is considered that there are two true-positive patterns on the two neighboring CHSs. Then two neighboring CHSs are reclassified into homozygous or heterozygous ones based on the true-positive patterns.

Pairs of neighboring CHSs are processed sequentially. Assume there are three successive CHSs (u, v, w). During the processing of two successive pairs (u, v) and (v, w), the unbalance join depths may result in a conflict on the classification of the overlapped CHS v. To handle with this conflict, a confidence score is calculated for each pair of neighboring CHSs, computed as the ratio of the total frequency of true-positive patterns on two sites to the maximum depth among three CpG sites, as defined in Equation (1). If $conf(u, v) >= conf(v, w)$, the class of v will be not changed, and the class of w will be determined based on the joint methylation statuses of (v, w) with the given class of v. If $conf(u, v) < conf(v, w)$, the class of v will be revised based on the true-positive patterns of (v, w).

$$conf(u, v) = \frac{\sum_{p}^{p \in TP} f(p)}{\max(d(u), d(v), d(w))} \tag{1}$$

where TP denotes the set of true-positive patterns of (u, v), $f(p)$ denotes the frequency of pattern p, and $d(u)$, $d(v)$, and $d(w)$ denote the depths of u, v, and w, respectively.

2.2.3. Construct Methylation Haplotypes for Each Candidate Region

After classifying CpG sites into homozygous and heterozygous ones, the skeletons of two methylation haplotypes are constructed by linking the patterns of neighboring heterozygous sites sequentially. Then, a pair of methylation haplotypes are constructed by padding the homozygous CpG sites into the skeletons.

2.2.4. Definition of Methylation Haplotype Mode

Each methylation haplotype can be represented by a 0–1 string. To simplify the comparison between methylation haplotypes, each methylation haplotype is converted into a label based on its 0–1 string, defined in Equation (2). Then, two labels of the paired methylation haplotypes on a candidate region, denoted as LL, HL, LN, LM, NN, MM, MN, HN, HM or HH, are termed as a methylation haplotype mode (MHM).

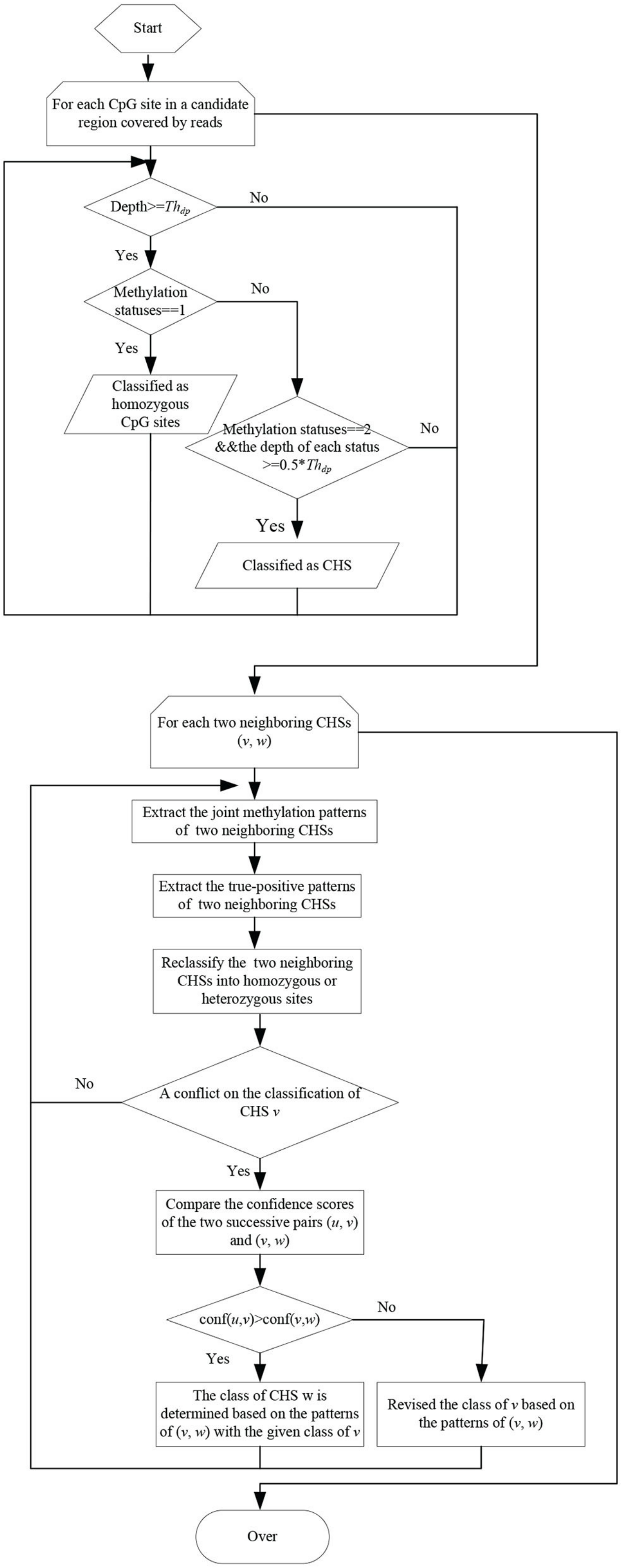

Start
For each CpG site in a candidate region covered by reads
Depth>=Th_{dp}
No
Yes
Methylation statuses==1
No
Yes
Classified as homozygous CpG sites
Methylation statuses==2 &&the depth of each status >=0.5*Th_{dp}
No
Yes
Classified as CHS
For each two neighboring CHSs (v, w)
Extract the joint methylation patterns of two neighboring CHSs
Extract the true-positive patterns of two neighboring CHSs
Reclassify the two neighboring CHSs into homozygous or heterozygous sites
No
A conflict on the classification of CHS v
Yes
Compare the confidence scores of the two successive pairs (u, v) and (v, w)
conf(u,v)>conf(v,w)
No
Yes
The class of CHS w is determined based on the patterns of (v, w) with the given class of v
Revised the class of v based on the patterns of (v, w)
Over

FIGURE 2 | The flowchart of classifying CpG sites into homozygous and heterozygous ones.

TABLE 2 | Statistics of candidate regions with methylation haplotypes in different samples.

Sample	Num. of candidate regions with VMHs	Ave. Num. of CpG sites in candidate regions	Ave. Num. of covered CpG sites in VMHs
Mature fat cells	249,253	10.51	5.91
Adipose-derived stem cells	256,671	10.51	6.08
Breast cancer sample	233,973	10.49	6.69
Normal breast sample	223,692	10.49	6.22
Hematopoietic stem cells	172,536	10.44	6.38
Mature B cells	138,053	10.44	5.20
Embryonic stem cells	220,970	10.63	6.05
Foreskin fibroblasts	213,317	10.66	6.10
Lower_leg_skin_1	228,263	10.30	8.67
Lower_leg_skin_2	244,369	10.36	8.96
Tibial_nerve_1	239,034	10.36	8.81
Tibial_nerve_2	225,728	10.31	8.67

$$Label(s) = \begin{cases} L, \text{if} MH(s) \leq 0.25 \\ N, \text{elseif} MH(s) \leq 0.5 \\ M, \text{elseif} MH(s) \leq 0.75 \\ H, \text{else} \end{cases} \tag{2}$$

where $MH(s) = \frac{\sum_{i=1}^{len(s)} (s_i - 0)}{len(s)}$, s represents the 0–1 string of a methylation haplotype, $len(s)$ represents the length of s, and s_i is the i-th character in s.

2.3. Map_DMR: DMR Identification Based on Methylation Haplotypes

Based on the MHMs of each candidate region among different samples, MHap_DMR identifies DMRs by comparing the MHMs directly. If the MHMs are identical, the candidate region is considered as a non-DMR. Otherwise, a methylation haplotype difference (MHD) between a pair of samples or groups is calculated, defined as in Equation (3). Then, the methylation difference among multi groups on the region can be defined as the maximum MHD among pairs of groups.

$$\begin{aligned} MHD(g_i, g_j) = {} & \max(abs(MH(g_{i1}) - MH(g_{j1})), abs(MH(g_{i2}) \\ & -MH(g_{j2}))) \end{aligned} \tag{3}$$

where g_i and g_j denote group i and j, g_{i1} and g_{j1} denote the 0–1 strings of methylation haplotypes with higher MH values in g_i and g_j, respectively, and g_{i2} and g_{j2} denote the 0–1

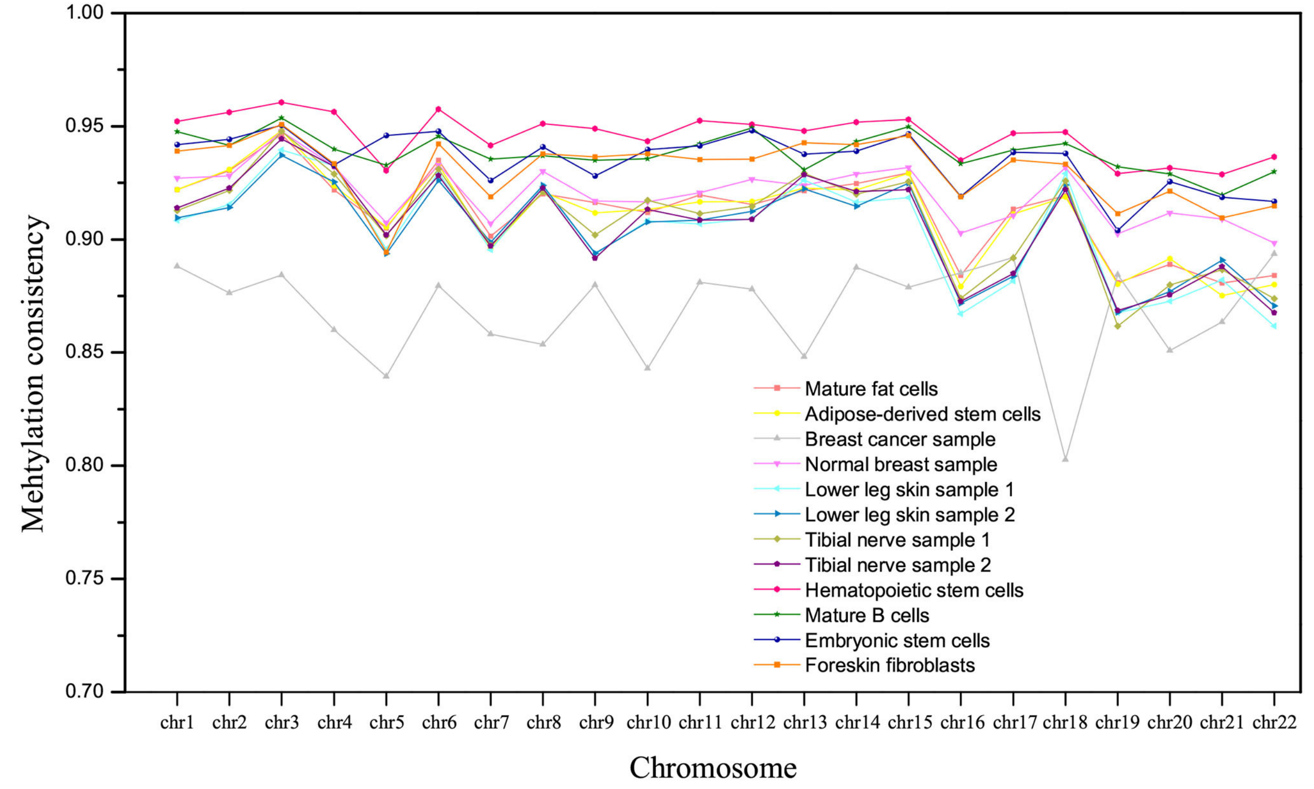


FIGURE 3 | The comparison of methylation consistency of homologous chromosomes in different tissues/cell types.

strings of methylation haplotypes with lower *MH* values in g_i and g_j, respectively.

To investigate the influence of homologous chromosomes on methylation analysis, we applied MHap to construct methylation haplotypes for 12 WGBS datasets of 10 different tissues/cell types. MHap constructs methylation haplotypes for each sample based on the alignment file and candidate regions. Methylation haplotypes covering more than 3 CpG sites are defined as valid methylation haplotypes (VMHs). **Table 2** lists the number of candidate regions with VMHs contained by each sample. It can be observed that the average number of CpG sites in these candidate regions is >10, and the average number of covered CpG sites in VMHs is ranging from 5.9 to 8.9.

3. RESULT

3.1. Majority of Methylation Haplotypes Are Consistent Between Homologous Chromosomes

MHMs *HH* and *LL* denote that the paired methylation haplotypes of two homologous chromosomes are simultaneously hypermethylated (*HH*) or hypomethylated (*LL*). Both the *HH* and *LL* are considered as consistent MHMs. Then, the methylation consistency between two homologous chromosomes in a sample can be defined as the ratio of the number of CpGs in VMHs with consistent MHMs to that in all VMHs.

The methylation consistency of homologous autosomes in different tissues/cell types is compared, as shown in **Figure 3**. For normal tissues or cell types, the methylation consistency is above 90% on average, especially in hematopoietic stem cells. A lower methylation consistency can be observed in the breast cancer sample, which is about 86% on all the homologous chromosomes.

The methylation consistency of chromosome X indicates the gender of a sample. In **Figure 4**, it can be observed that three samples with methylation consistency above 94% are derived from male, while samples with methylation consistency ranging from 54 to 72% are derived from female which is much lower than that of other homologous autosomes. It coincides with the previous studies that the methylation between two homologous chromosome X in female are different, one of which is inactive and highly methylated (Mohandas et al., 1981; Goto and Monk, 1998).

Further, we compared the hypomethylation consistency in different samples. The hypomethylation consistency between two homologous chromosomes in a sample can be defined as the ratio of the number of CpGs in VMHs with consistent MHM *LL* to that in all VMHs. From **Figure 5**, we can observe that the hypomethylation consistency of derived cells is higher than that of the corresponding undifferentiated stem cells, which is consistent with the former studies that methylation decrease with

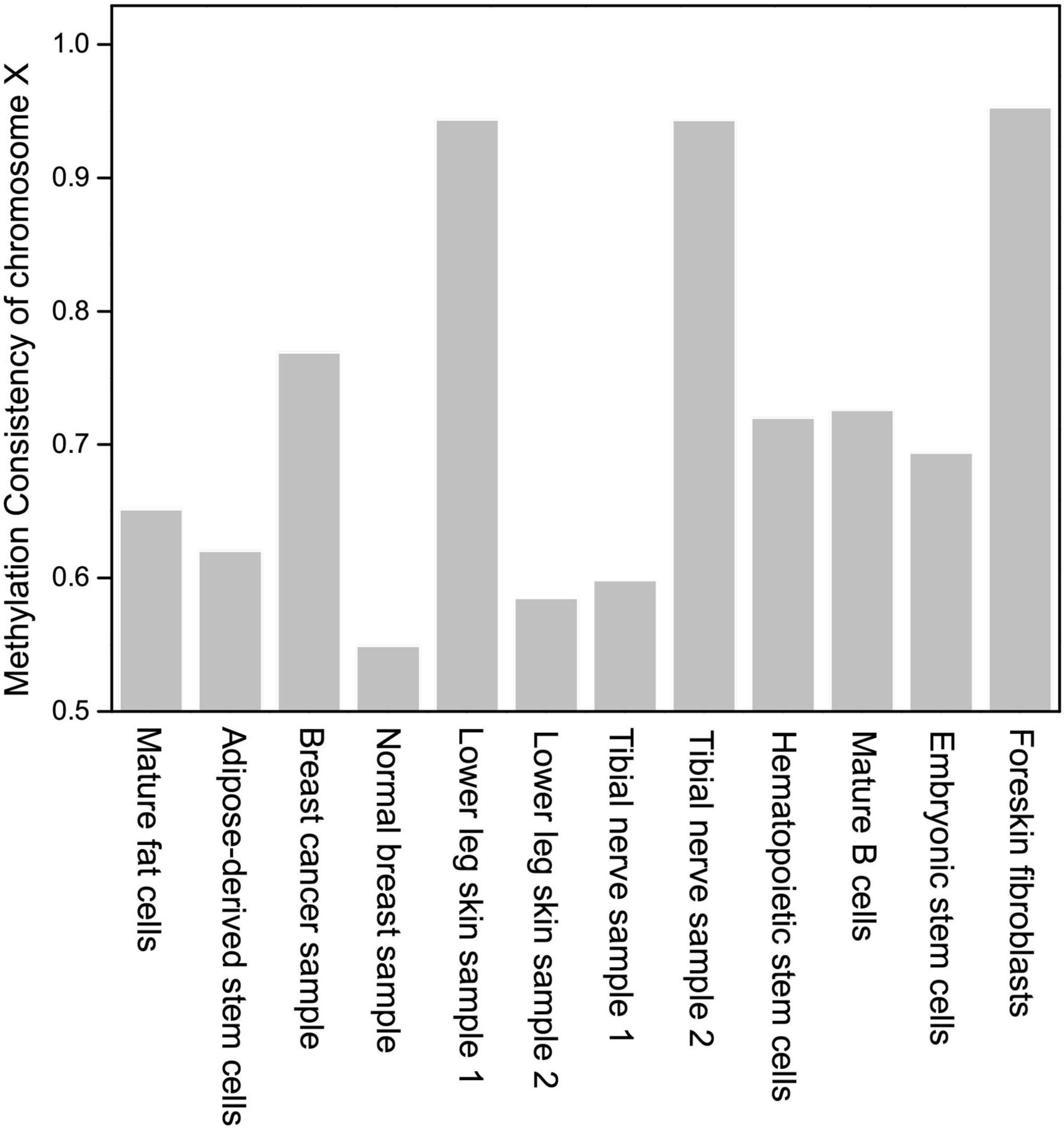


FIGURE 4 | The comparison of methylation consistency of chromosome X in different samples.

the degree of differentiation increased (Laurent et al., 2010). In **Figure 5**, we can find that the mature fat cells are more hypomethylated than adipose-derived stem cells, mature B cells are more hypomethylated than hematopoietic stem cells, and foreskin fibroblasts are more hypomethylated than embryonic stem cells. It is also noted that the hypomethylation consistency of breast cancer sample is much lower than that of normal breast sample on homologous chromosome.

In addition, it is interesting to observe that the tissues/cell types can be roughly clustered into three groups according to the hypomethylation consistency, as shown in **Figure 5**. Lower leg skin and tibial nerve have similar hypomethylation consistency and they belong to the ectoderm. The hypomethylation consistency of mature fat cells, adipose-derived stem cells, mature B cells, hematopoietic stem cells and the normal breast sample are similar, and these tissues/cell types belong to the mesoderm. The hESCs and hESC-Fibro cell types have high hypomethylation consistency in chromosomes, which are higher than that of other tissues/cell types.

3.2. Identifying DMRs Between Two Samples

MHap_DMR was applied to identify DMRs in four pairs of samples, including breast cancer vs. normal breast, mature fat

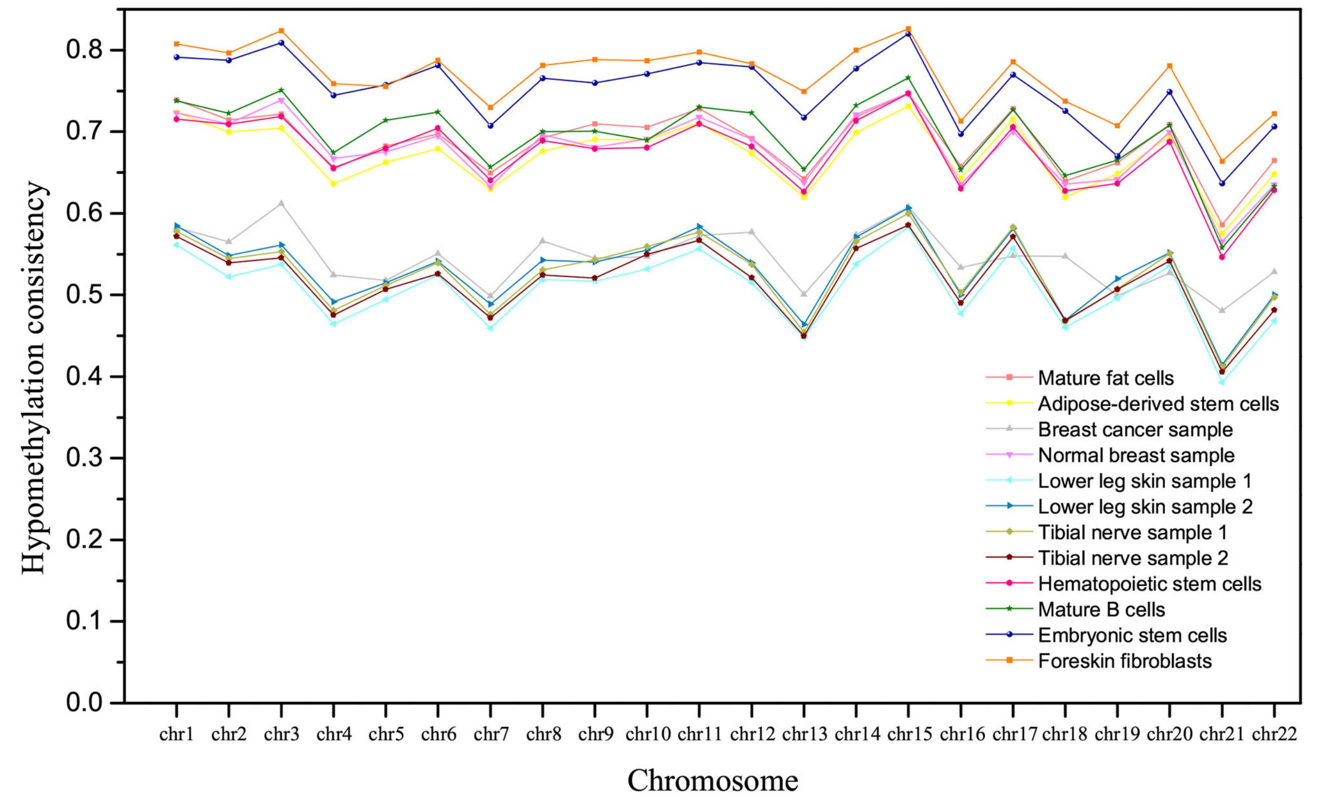


FIGURE 5 | The comparison of hypomethylation consistency of homologous chromosomes in different tissues/cell types.

TABLE 3 | Four types of DMRs identified by MHap_DMR for each pair of samples.

Pairs of samples	**Type 1 DMR (hypo- vs. non-hypo)**	**Type 2 DMR (consistent hypo- vs. semi-hypo)**	**Type 3 DMR (consistent hyper- vs. semi-hyper)**	**Type 4 DMR with other modes**
Mature fat cells vs. Adipose-derived stem cells	1,156 (LL vs. HH: 1) (HL vs. HH: 459)	1,032 (LL vs. HL: 574)	583	309
Breast cancer vs. Normal breast	20,138 (LL vs. HH: 15,175) (HL vs. HH: 650)	1,351 (LL vs. HL: 1,351)	0	0
Hematopoietic stem cells vs. Mature B cells	1,468 (LL vs. HH: 391) (HL vs. HH: 223)	625 (LL vs. HL: 286)	182	257
Embryonic stem cells vs. Foreskin fibroblasts	6,856 (LL vs. HH: 2,698) (HL vs. HH: 1,465)	2,812 (LL vs. HL: 1,490)	914	340

cells vs. adipose-derived stem cells, embryonic stem cells (hESCs) vs. foreskin fibroblasts (hESC-Fibro cells), and mature B cells vs. hematopoietic stem cells. In this study, MHap_DMR reports the DMRs with $p < 0.05$ and $MHD > 0.5$.

Based on the MHMs of samples on DMRs, the identified DMRs can be further classified into four groups: 1. hypomethylation mode (a MHM containing *L*) vs. non-hypomethylation mode (a MHM not containing *L*); 2. hypomethylation consistent mode *LL* vs. semi-hypomethylation mode (an unconsistent MHM containing *L*); 3. hypermethylation consistent mode *HH* vs. semi-hypermethylation mode (an unconsistent MHM containing

TABLE 4 | The number of DMRs identified by different methods.

Pairs of samples	MHap_DMR	CpG_MPs	DMRCaller	SMART	Metilene
Mature fat cells vs. Adipose-derived stem cells	3,080	932	4,081	2,152	44,359
Breast cancer vs. Normal breast	21,489	233,298	861,108	353,565	357,980
Hematopoietic stem cells vs. Mature B cells	2,532	26,526	172,475	50,453	75,180
Embryonic stem cells vs. Foreskin fibroblasts	10,922	130,376	385,877	282,617	338,631

H); 4. DMR with other modes. The number of these types of DMRs between each pair of samples is listed in **Table 3**.

To investigate the methylation changing directions at the methylation haplotype level, the number of some subtypes of DMRs in Type 1 and Type 2 DMRs is specified. For example, in Type 1 DMRs, the number of DMRs with hypomethylation consistent mode *LL* vs. hypermethylation consistent mode *HH* and the number of DMRs with hypomethylation unconsistent mode *HL* vs. hypermethylation consistent mode *HH* are listed.

In Type 1 DMRs, it can be observed that there is only 1 DMR with hypomethylation consistent mode *LL* vs. hypermethylation mode consistent *HH* in mature fat cells and adipose-derived stem cells. It may indicate that the methylation statuses of two homologous chromosomes are seldom changed simultaneously during the differentiation from adipose-derived stem cells to mature fat cells.

In Type 1 DMRs between breast cancer and normal breast, it can be observed that there are 13,173 DMRs with hypermethylation consistent mode *HH* in breast cancer and hypomethylation consistent mode *LL* in normal breast, while there are only 2,002 DMRs with hypomethylation consistent mode *LL* in breast cancer and hypermethylation consistent mode *HH* in normal breast. It suggests that many regions with hypomethylation consistent mode *LL* in normal breast become hypermethylated in breast cancer, while a small quantity of regions with hypermethylation consistent mode *HH* in normal breast become hypomethylated in breast cancer. Further, comparing the number of four types of DMRs between breast cancer and normal breast, it may indicate that, in breast cancer, the methylation statuses of homologous chromosomes changes in the same direction (hypomethylated or hypermethylated) simultaneously in many cases. The MHMs of DMRs among different samples can indicate the methylation changing directions of homologous chromosomes in cell differentiation and cancerization.

3.3. Compared With Comparative Methods

To further demonstrate the performance of MHap_DMR, four comparative tools were also applied to these four pairs of samples, including CpG_MPs (Su et al., 2012), DMRCaller (Catoni et al., 2018), SMART (Liu et al., 2015), and Metilene (Jühling et al., 2016). The default parameter settings were adopted when running these methods.

The numbers of DMRs identified by different methods are compared, as shown in **Table 4**. Metilene always predicts a larger number of DMRs with low methylation level differences than other methods. MHap_DMR predicts a smaller number of DMRs than other methods, because it works on candidate regions predefined on the CpG dense regions. All the methods report a largest number of DMRs between breast cancer sample and normal breast sample, and a second largest number of DMRs between embryonic stem cells and foreskin fibroblasts. This consistency indicates that DNA methylation is altered a lot in cancerization, and the methylation difference between embryonic stem cells and foreskin fibroblasts is larger than that between other types of stem cells and the cells derived from these stem cells.

4. CONCLUSION

In this paper, MHap is developed to construct methylation haplotypes for homologous chromosomes in CpG dense regions. Through the analysis based on methylation haplotypes of homologous chromosomes, we found that majority of methylation haplotypes are consistent between homologous autosomes, while a lower methylation consistency was observed in the breast cancer sample. Further, the hypomethylation consistency of derived cells is higher than that of the corresponding undifferentiated stem cells. The hypomethylation consistency can be used as a feature for cell clustering. DMRs identified by MHap_DMR based on methylation haplotypes can help to investigate the methylation changing directions of homologous chromosomes in cell differentiation and cancerization.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: the ENCODE project (https://www.encodeproject.org/) through the access sample ids ENCSR930WUY, ENCSR128RMY, ENCSR752OCM, and ENCSR658MZU, the GEO database (https://www.ncbi.nlm.nih.gov/geo/) through GEO accession numbers GSE29069, GSE19418, and GSE31971, and the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) under the accession number SRA0238292.

AUTHOR CONTRIBUTIONS

XP and XD conceived and designed the approach. XP and YL performed the experiments. YL and XK analyzed the data. XP wrote the manuscript. XP and XZ supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 61702555, 61662028, and U1909208), the National Key R&D Program of China (No. 2018YFC0910504), 111 Project (No. B18059), and Hunan Provincial Science and Technology Program (2018WK4001).

REFERENCES

Abante, J., Fang, Y., Feinberg, A., and Goutsias, J. (2020). Detection of haplotype-dependent allele-specific DNA methylation in WGBS data. *Nat. Commun.* 11, 1–13. doi: 10.1038/s41467-020-19077-1

Baylin, S. B. (2005). DNA methylation and gene silencing in cancer. *Nat. Rev. Clin. Oncol.* 2:S4. doi: 10.1038/ncponc0354

Catoni, M., Tsang, J. M., Greco, A. P., and Zabet, N. R. (2018). DMRcaller: a versatile R/bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucl. Acids Res.* 46:e114. doi: 10.1093/nar/gky602

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y.-P. P., et al. (2019). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. 18, 1106–1112 doi: 10.1109/TCBB.2019.2936476

Cheung, W. A., Shao, X., Morin, A., Siroux, V., Kwan, T., Ge, B., et al. (2017). Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biol.* 18, 1–21. doi: 10.1186/s13059-017-1173-7

Condon, D. E., Tran, P. V., Lien, Y.-C., Schug, J., Georgieff, M. K., Simmons, R. A., et al. (2018). Defiant:(dmrs: easy, fast, identification and annotation) identifies differentially methylated regions from iron-deficient rat hippocampus. *BMC Bioinformatics* 19:31. doi: 10.1186/s12859-018-2037-1

Eden, A., Gaudet, F., Waghmare, A., and Jaenisch, R. (2003). Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* 300:455. doi: 10.1126/science.1083557

Feng, H., Conneely, K. N., and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucl. Acids Res.* 42:e69. doi: 10.1093/nar/gku154

Gomez, L., Odom, G. J., Young, J. I., Martin, E. R., Liu, L., Chen, X., et al. (2019). coMethDMR: accurate identification of co-methylated and differentially methylated regions in epigenome-wide association studies with continuous phenotypes. *Nucl. Acids Res.* 47:e98. doi: 10.1093/nar/gkz590

Goto, T., and Monk, M. (1998). Regulation of x-chromosome inactivation in development in mice and humans. *Microbiol. Mol. Biol. Rev.* 62, 362–378. doi: 10.1128/MMBR.62.2.362-378.1998

Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* 13:R83. doi: 10.1186/gb-2012-13-10-r83

Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* 29, 1647–1653. doi: 10.1093/bioinformatics/btt263

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., et al. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell* 44, 17–28. doi: 10.1016/j.molcel.2011.08.026

Hon, G. C., Hawkins, R. D., Caballero, O. L., Lo, C., Lister, R., Pelizzola, M., et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* 22, 246–258. doi: 10.1101/gr.125872.111

Hu, X., Li, M., Wang, L., Li, X., Wu, F.-X., and Wang, J. (2019). "Classification of schizophrenia by iterative random forest feature selection based on DNA methylation array data," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (San Diego, CA), 807–811. doi: 10.1109/BIBM47256.2019.8983308

Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., and Hoffmann, S. (2016). metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* 26, 256–262. doi: 10.1101/gr.196394.115

Krueger, F., and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 27, 1571–1572. doi: 10.1093/bioinformatics/btr167

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). LDICDL: LncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. doi: 10.1109/TCBB.2020.3034910

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., et al. (2010). Dynamic changes in the human methylome during differentiation. *Genome Res.* 20, 320–331. doi: 10.1101/gr.101907.109

Lea, A. J., Tung, J., and Zhou, X. (2015). A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genet.* 11:e1005650. doi: 10.1371/journal.pgen.1005650

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., et al. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471:68. doi: 10.1038/nature09798

Liu, H., Liu, X., Zhang, S., Lv, J., Li, S., Shang, S., et al. (2015). Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucl. Acids Res.* 44, 75–94. doi: 10.1093/nar/gkv1332

Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., et al. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.* 15:3248. doi: 10.1186/gb-2014-15-4-r54

Mohandas, T., Sparkes, R., and Shapiro, L. (1981). Reactivation of an inactive human x chromosome: evidence for x inactivation by DNA methylation. *Science* 211, 393–396. doi: 10.1126/science.6164095

Ni, P., Huang, N., Zhang, Z., Wang, D.-P., Liang, F., Miao, Y., et al. (2019). DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics* 35, 4586–4595. doi: 10.1093/bioinformatics/btz276

Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). Methylsig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* 30, 2414–2422. doi: 10.1093/bioinformatics/btu339

Park, Y., and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* 32, 1446–1453. doi: 10.1093/bioinformatics/btw026

Scott, C. A., Duryea, J. D., MacKay, H., Baker, M. S., Laritsky, E., Gunasekara, C. J., et al. (2020). Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol.* 21, 1–23. doi: 10.1186/s13059-020-02065-5

Stockwell, P. A., Chatterjee, A., Rodger, E. J., and Morison, I. M. (2014). DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics* 30, 1814–1822. doi: 10.1093/bioinformatics/btu126

Su, J., Yan, H., Wei, Y., Liu, H., Liu, H., Wang, F., et al. (2012). CpG_MPs: identification of CPG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucl. Acids Res.* 41:e4. doi: 10.1093/nar/gks829

Sun, S., and Yu, X. (2016). Hmm-fisher: identifying differential methylation using a hidden Markov model and fisher's exact test. *Stat. Appl. Genet. Mol. Biol.* 15, 55–67. doi: 10.1515/sagmb-2015-0076

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57. doi: 10.1038/nature11247

Wang, Z., Li, X., Jiang, Y., Shao, Q., Liu, Q., Chen, B., et al. (2015). swDMR: a sliding window approach to identify differentially methylated regions based on whole genome bisulfite sequencing. *PLoS ONE* 10:e0132866. doi: 10.1371/journal.pone.0132866

Wen, Y., Chen, F., Zhang, Q., Zhuang, Y., and Li, Z. (2016). Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. *Bioinformatics* 32, 3396–3404. doi: 10.1093/bioinformatics/btw497

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., et al. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucl. Acids res.* 43:e141. doi: 10.1093/nar/gkv715

Xiaoqing, P., Hong-Dong, L., Fang-Xiang, W., and Jianxin, W. (2020). Identifying the tissues-of-origin of circulating cell-free DNAs is a promising way in noninvasive diagnostics. *Brief. Bioinformatics* 22:bbaa060. doi: 10.1093/bib/bbaa060

Xu, Z., Xie, C., Taylor, J. A., and Niu, L. (2020). ipDMR: identification of differentially methylated regions with interval p-values. *Bioinformatics* 37, 711–713. doi: 10.1093/bioinformatics/btaa732

Yan, C., Duan, G., Wu, F.-X., Pan, Y., and Wang, J. (2019). BRWMDA: predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 17, 1595–1604. doi: 10.1109/TCBB.2019.2907626

Yan, C., Duan, G., Wu, F.-X., Pan, Y., and Wang, J. (2021). MCHMDA: Predicting microbe-disease associations based on similarities and low-rank matrix completion. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 18, 611–620. doi: 10.1109/TCBB.2019.2926716

Yan, C., Wang, J., Ni, P., Lan, W., Wu, F.-X., and Pan, Y. (2017). DNRLMF-MDA: predicting microRNA-disease associations based on similarities of microRNAs and diseases. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 233–243. doi: 10.1109/TCBB.2017.2776101

Yan, C., Wang, J., and Wu, F.-X. (2018). DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC Bioinformatics* 19(Suppl. 19):520. doi: 10.1186/s12859-018-2522-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 02 July 2021
doi: 10.3389/fgene.2021.695597

Transcriptomic Analysis of Gene Networks Regulated by U11 Small Nuclear RNA in Bladder Cancer

Zhenxing Wang[1,2,3]†, Xi Wang[1,2]†, Yaobang Wang[1,2]†, Shaomei Tang[4], Chao Feng[1,2], Lixin Pan[1,2], Qinchen Lu[1,2], Yuting Tao[1,2], Yuanliang Xie[1,2,5], Qiuyan Wang[1,2]* and Zhong Tang[1,2,6]**

[1] Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, China, [2] Guangxi Key Laboratory for Genomic and Personalized Medicine, Guangxi Collaborative Innovation Center for Genomic and Personalized Medicine, Nanning, China, [3] Department of Clinical Laboratory, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [4] Department of Gastroenterology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [5] Department of Urology, The Affiliated Cancer Hospital of Guangxi Medical University, Nanning, China, [6] School of Information and Management, Guangxi Medical University, Nanning, China

OPEN ACCESS

Edited by:
Chen Qingfeng,
Guangxi University, China

Reviewed by:
Ximiao He,
Huazhong University of Science and Technology, China
Jinping Liu,
University of Pennsylvania, United States
Huiyu Zhou,
University of Leicester, United Kingdom

***Correspondence:**
Yuanliang Xie
136641161@qq.com
Qiuyan Wang
qiuyanwang510@yahoo.com
Zhong Tang
18978868502@qq.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 April 2021*
Accepted: *11 June 2021*
Published: *02 July 2021*

Citation:
Wang Z, Wang X, Wang Y, Tang S, Feng C, Pan L, Lu Q, Tao Y, Xie Y, Wang Q and Tang Z (2021) Transcriptomic Analysis of Gene Networks Regulated by U11 Small Nuclear RNA in Bladder Cancer. Front. Genet. 12:695597. doi: 10.3389/fgene.2021.695597

Small nuclear RNA is a class of non-coding RNA that widely exist in the nucleus of eukaryotes. Accumulated evidences have shown that small nuclear RNAs are associated with the regulation of gene expression in various tumor types. To explore the gene expression changes and its potential effects mediated by U11 snRNA in bladder cancer cells, U11 snRNA knockout and overexpressed cell lines were constructed and further used to analyze the gene expression changes by RNA sequencing. The differentially expressed genes were found to be mainly enriched in tumor-related pathways both in the U11 knockout and overexpression cell lines, such as NF-kappa B signaling pathway, bladder cancer and PI3K-Akt signaling pathway. Furthermore, alternative splicing events were proposed to participate in the potential regulatory mechanism induced by the U11 knockout or overexpression. In conclusion, U11 may be involved in the regulation of gene expression in bladder cancer cells, which may provide a potentially new biomarker for clinical diagnosis and treatment of bladder cancer.

Keywords: U11 small nuclear RNA, bladder cancer, T24, transcriptomic analysis, gene networks

INTRODUCTION

Bladder cancer is one of the most common urological cancers, ranking ninth among all malignant tumors worldwide and sixth among men (Ploeg et al., 2009; Bray et al., 2018). In the United States, bladder cancer ranks fourth among all malignant tumors, with 74,000 new cases of bladder cancer in 2015, including 56,320 males and 17,680 females, and the estimated fatal cases were 16,000, including 11,510 males and 4,490 females. In South Asia and Western Asia, the incidence and mortality of bladder cancer also rank top among all malignant tumors (Salim et al., 2010). Although the incidence and mortality of bladder cancer in China are lower than the world average level, there is a trend of increasing incidence in some cities (Feng et al., 2019), which seriously threatens the survival health and life quality of patients. Therefore, it is of great significance to study the mechanism of the occurrence and development of bladder cancer, and further to improve the diagnosis and treatment rate of bladder cancer.

Cajal bodies, also known as coiled bodies, were first discovered in the nucleus of nerve cells by Ramony Cajal in 1903 (Hebert, 2010). Cajal bodies are widely found in the nucleus of higher eukaryotes, and their number and size can vary with species. Cajal bodies are more abundant in somatic cells of differentiated tissues and some cells with higher metabolic activity, such as muscle, neurons, and tumor cells (Hearst et al., 2009; Machyna et al., 2013). At the same time, the numbers and sizes of Cajal bodies are related to the cell cycle, and their numbers often reach the maximum in the G1/S phase. In general, Cajal bodies depolymerize in the M phase, and the reassembly process depends on the level of gene transcription and the rate of proliferation of the cell. Relevant studies have shown that there are a large number of factors involved in the splicing of mRNA precursors, rRNA precursor processing, histone pre-mRNA 3′ end processing and telomere maintenance in the components of Cajal bodies, suggesting that Cajal bodies may be a site for the assembly and function of ribonucleoprotein (RNP) (Hebert, 2013). Cajal bodies can also serve as a platform for effective modification responses in transcriptionally active cells requiring high levels of RNP, such as neuronal cells and tumor cells. Studies have found that zebrafish embryos are unable to complete embryonic development due to the lack of coilin and Cajal bodies. The depletion of coilin and Cajal bodies was mainly characterized by deficits in snRNP biogenesis and expression of spliced mRNA, while mature snRNPs can partially rescue embryonic lethal phenotypes (Strzelecka et al., 2010).

At present, more than dozens of proteins have been found to localize or interact with Cajal bodies (Machyna et al., 2013). Coilin is recognized as a marker protein and a major component of Cajal bodies (Andrade et al., 1991). One of the most significant structural features of Cajal bodies is the accumulation of a large number of non-coding small RNAs, which include U1, U2, U3, U4, U5, U6, U7, U8, U11, U12, U13, U14, U64, U6atac, and U87 scaRNA, etc. These small non-coding RNAs were once thought to be mainly involved in post-transcriptional modification of RNA. However, with the development of high-depth sequencing technology and bioinformatics technology, the functions of these small non-coding RNAs have been further recognized and understood. Studies have found that they may be involved in gene expression, genome structure organization, and other functions (Lui and Lowe, 2013). In recent years, non-coding RNAs have attracted much attention as a special way of gene expression regulation.

Our previous studies innovatively proposed the notion that Cajal bodies can be simultaneously linked with multiple small molecule RNA gene loci to form gene clusters, using Hela cell models and six-color microscopes detection systems (Wang et al., 2016). Moreover, we found that this gene cluster is not formed randomly but is a specific spatial structure of long-distance interactions. U1, U2, U3, U4, U5, and U11 are small non-coding RNA genes enriched in Cajal bodies, U87 scaRNA is small Cajal body-specific RNA gene, Histone cluster 2 and Histone H3F3 are histone small RNA genes. Among them, U1, U2, U3, U4, U5, and U11 are small non-coding RNA genes enriched in Cajal bodies. Using chromatin spatial conformation capture technology, it was found that small nuclear RNA mediates the formation of long-distance chromatin interactions (Wang et al., 2016). U11 (RNU11) is probably the most significant small nuclear RNA because its expression is extremely high expressed in rapidly growing tumor cells and very low expressed in slow-growing primary cells. Thus, the above studies provided the evidence that U11 small nuclear RNA plays a role in regulating the spatial structure of chromatin and may be of great significance in the development of tumors.

Interestingly, we found Cajal bodies were aberrantly activated in T24 bladder cancer cell lines and the increased numbers and sizes of Cajal bodies were displayed in two highly invasive and metastatic T24-SLT and T24-FL cell lines. Given that U11 is one of the most crucial snRNAs located in Cajal bodies, we speculated that U11 might play an important role in the gene expression of bladder cancer cells. In this study, the *in vitro* cell models with U11 knockout and overexpression were firstly constructed in T24 bladder cancer cell lines and further used to analyze the gene expression changes using RNA-sequencing technology. The differentially expressed genes were found to be mainly enriched in tumor-related pathways both in the U11 knockout and overexpression groups. Notably, alternative splicing events were innovatively proposed to participate in the potential regulatory mechanism induced by U11 knockout or overexpression. Taken together, our study innovatively elucidated that U11 may play the critical role in the regulation of gene expression in bladder cancer cells, which may provide a potentially new biomarker for clinical diagnosis and treatment of bladder cancer.

RESULTS

Cajal Body-Related snRNA U11 Affects the Occurrence and Development of Bladder Cancer by Regulating Differently Expressed Genes

By using immunofluorescence (IF) staining, we unexpectedly found Cajal bodies were aberrantly activated in T24 bladder cancer cell lines. More interestingly, the increased numbers and sizes of Cajal bodies were displayed in two highly invasive and metastatic T24-SLT and T24-FL cell lines (Nicholson et al., 2004; Jeppesen et al., 2014) (**Figure 1A**). Given that U11 is one of the most crucial snRNAs located in Cajal bodies, the *in vitro* cell models with U11 knockout and overexpression were successfully established in T24 cell lines. The knockout and overexpression efficiency of U11 in T24 cell lines were confirmed by real-time quantitative PCR, the expression level of U11 in U11-KO cell line was significantly decreased, and that of U11-KI cell line was significantly overexpressed compared to T24 WT cell line (**Figure 1B**). Moreover, MTT assay revealed that the cell proliferation ability of T24 WT cell line was significantly higher than that of the U11-KO cell line ($P < 0.001$, **Figure 1C**). The U11 knockout and U11 overexpression groups were used as experimental groups, and gene differences were analyzed by comparing with the control group. A total of 2,756 differentially expressed genes in the U11 knockout group were obtained, including 1,464 up-regulated and 1,292 down-regulated

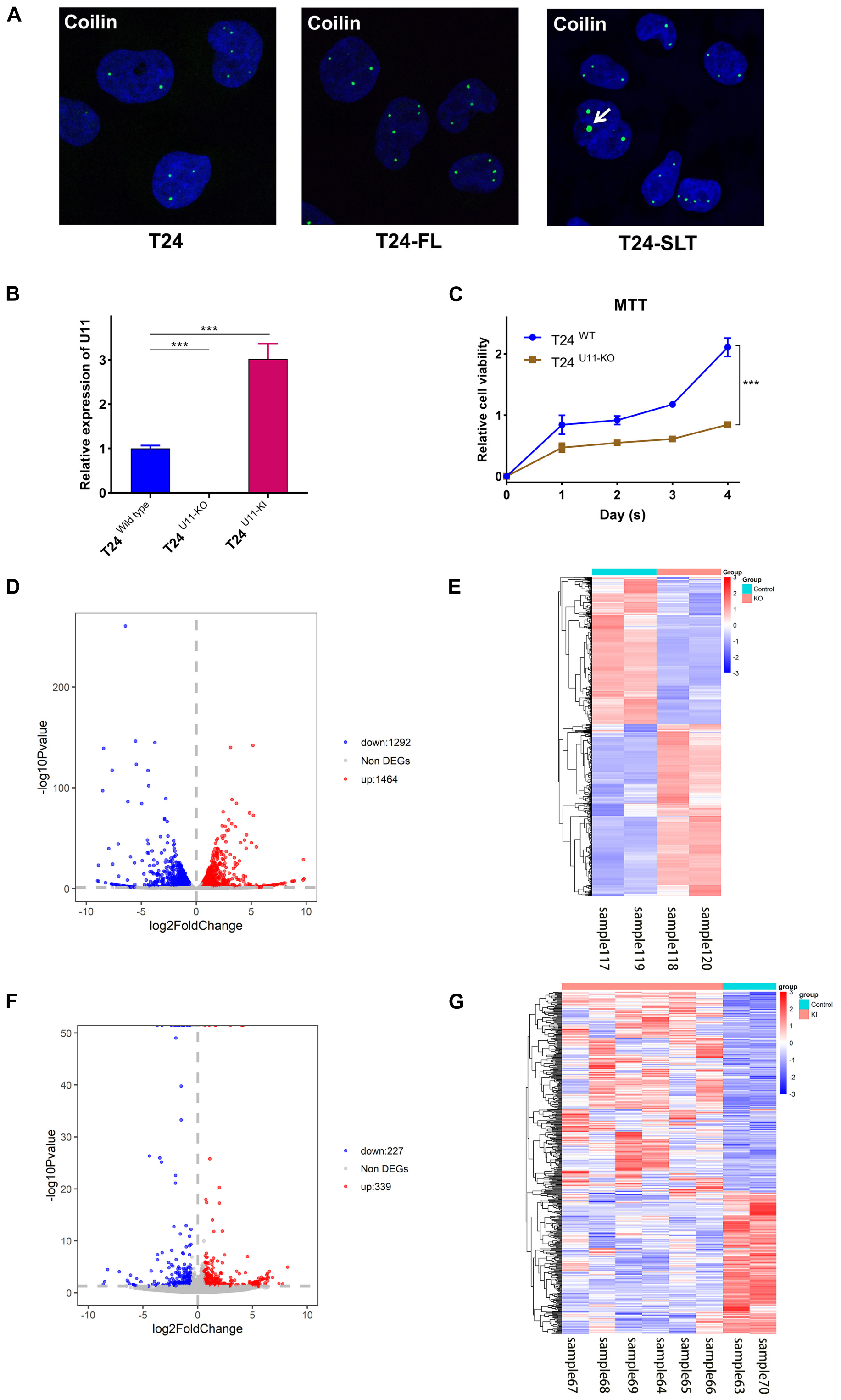


FIGURE 1 | Construction of T24^{U11-KO} and T24^{U11-KI} cell lines and DEGs analysis of T24-related cell lines. **(A)** Immunofluorescence of Coilin in T24, T24-FT, T24-SLT cell lines. **(B)** The expression levels of U11 in T24^{U11-KO} and T24^{WT} cell lines. **(C)** Cell viability of T24^{U11-KO} and T24^{WT} cell lines. **(D)** Volcano plot for differentially expressed genes between T24^{U11-KO} and T24^{WT} cell lines. **(E)** Heatmap for differentially expressed genes between T24^{U11-KO} and T24^{WT} cell lines. **(F)** Volcano plot for differentially expressed genes of between T24^{U11-KI} and T24^{WT} cell lines. **(G)** Heatmap for differentially expressed genes of between T24^{U11-KI} and T24^{WT} cell lines. Differential expressed genes (DEGs): *P*-value < 0.05 and | Fold change| ≥ 1.5. ***$P < 0.001$.

(**Figures 1D,E**); In addition, there were 566 differentially expressed genes in the U11 overexpression group, including 339 up-regulated and 227 down-regulated (**Figures 1F,G**). Pathway enrichment analysis by clusterProfiler R package showed that the up-regulated genes obtained by overexpressing U11 were mainly enriched in regulation of mast cell degranulation, chemokine activity, NF-kappa B signaling pathway, TNF signaling pathway, and Bladder cancer, etc. (**Figures 2A,B**). The down-regulated genes obtained by overexpressing U11 were mainly enriched in integral component of lumenal side of endoplasmic reticulum membrane, cellular response to type I interferon, defense response to virus, Allograft rejection and Antigen processing and presentation, etc. (**Figures 2C,D**). Interestingly, we found that the enrichment pathways of down-regulated genes in the U11 knockout group were similar to those of up-regulated genes in the U11 overexpression group. The pathways of down-regulated genes in U11 knockout group were mainly enriched in T cell apoptotic process, cytokine receptor binding, TNF signaling pathway, and NF-kappa B signaling pathway, etc. (**Figures 2G,H**). The pathways of up-regulated genes in U11 knockout group were mainly enriched in laminin binding, fibronectin binding, cell-substrate adhesion, Proteoglycans in cancer, and PI3K-Akt signaling pathway (**Figures 2E,F**). Among them, **Figure 3** showed the PI3K-Akt signaling pathway and the genes involved in this pathway.

Subsequently, we used the up-regulated genes in the U11 overexpression group and the down-regulated genes in the U11 knockout group for intersection analysis, as well as the down-regulated genes in the U11 overexpression group and the up-regulated genes in the U11 knockout group for intersection analysis. We found that there were 93 intersecting genes in the two intersecting groups (**Figures 4A,B**). Using these genes for protein interaction network analysis, we found that the hub genes were mainly CXCL2, CXCL3, CXCL6, CXCL8, and other CXCL gene families (**Figure 4C**). Interestingly, we found that the expression of CXCL2, CXCL3, CXCL6, CXCL8, and other CXCL gene families was significantly up-regulated in the U11 overexpression group, while the expression of these genes was significantly down-regulated in the U11 knockout group (**Figures 4D,E**). The intersecting genes of up-regulated genes in U11 knockout group and down-regulated genes in U11 overexpression group were mainly enriched in regulation of cell adhesion mediated by integrin, fibroblast migration and regulation of lipolysis in adipocytes (**Figure 4F**), and the intersecting genes of down-regulated genes in U11 knockout group and up-regulated genes in U11 overexpression group were mainly enriched in chemokine activity, response to lipopolysaccharide, neutrophil migration, NF-kappa B signaling pathway, TNF signaling pathway, and transcriptional mis-regulation in cancer (**Figure 4G**).

U11 Alters Gene-Splicing Events and Gene Expression

We further performed alternative splicing analysis and found that a total of 4,023 genes in the U11 overexpression group had significant differential alternative splicing events. Among them, exon skipping (SE) was the most frequent event, while intron retention (RI) was the least frequent event, and 316 genes were simultaneously exposed to five alternative splicing events (**Figure 5A**). Similarly, 4,774 genes were found to have significant differential alternative splicing events in the U11 knockout group, with exon skipping events occurring most frequently (**Figure 5B**).

Next, we intersected genes with differentially alternative splicing events and differentially expressed genes (**Figure 6A**). Seventy-one intersecting genes were obtained in the U11 overexpression group. Among them, murine double minute 2 (MDM2) gene had one exon skip and one mutually exclusive exon event, and the gene expression increased about 3.5-fold, and TGFB2 gene had one exon skip, and the gene expression decreased 2.1-fold (**Figures 6B,E**). Notably, 648 intersecting genes in the U11 knockout group, which were mainly enriched in pathways such as NF-kappa B signaling pathway and TNF signaling pathway (**Figure 6C**). The results of protein interaction network analysis of these intersecting genes also showed that the hub genes mainly included genes such as TIMP1, FN1, and RPL22L1 (**Figure 6D**). Intriguingly, FN1 gene had multiple alternative 3′ splice site (A3SS) events, one mutually exclusive exon event, four exon skipping events, and the level of mRNA expression increased 2.9-fold. TIMP1 gene had only one exon skipping event, and the level of mRNA expression increased 3.5-fold. RPL22L1 gene had one exon skipping events, and the level of expression decreased 1.7-fold (**Figures 6B,E**).

Given that the alternative splicing events of TIMP1, FN1, and RPL22L1 have been widely reported to participate in several biological processes (Usher et al., 2007; Lopez-Mejia et al., 2013; Liang et al., 2019), we therefore validated the AS events of the genes described above, using PCR and gel electrophoresis. We initially examined three typical exons skipping of FN1 (EDA, EDB, and IIICS) as previously reported (Lopez-Mejia et al., 2013), but no remarkably alternative splice events were detected in FN1 gene. The full-length TIMP1 transcript was then detected by the forward primer located in exon 1 combined with the reverse primer located in exon 6. Intriguingly, the band of full-length TIMP1 in T24^{U11-KO} cell line was observed to shift down a weak distance less than an exon, compared with T24^{WT} cell line. What' more, as the gel picture shown and the gray arrow indicated in **Figure 6F**, one indistinctly visible band appeared below the major band in T24^{U11-KO} cell line, but not in T24^{WT} cell line, suggesting a potential AS event of TIMP1 after U11 knockout. Three exons of TIMP1 gene were further examined the alternative splicing events, respectively. However, no significant alteration of splicing patterns was observed in the TIMP1-1, TIMP1-2, and TIMP1-3 segments because of the rare frequency and low abundance of the exon skipping (**Figure 6F**). All these results indicated that a link between alternative splicing regulated by U11 and bladder carcinogenesis.

FN1 and RPL22L1 May Be a Prognostic Marker for Bladder Cancer

To further investigate whether hub genes have an impact on the prognosis of bladder cancer, we performed survival analysis of CXCL8, MDM2, TGFB2, FN1, TIMP1, and RPL22L1, and

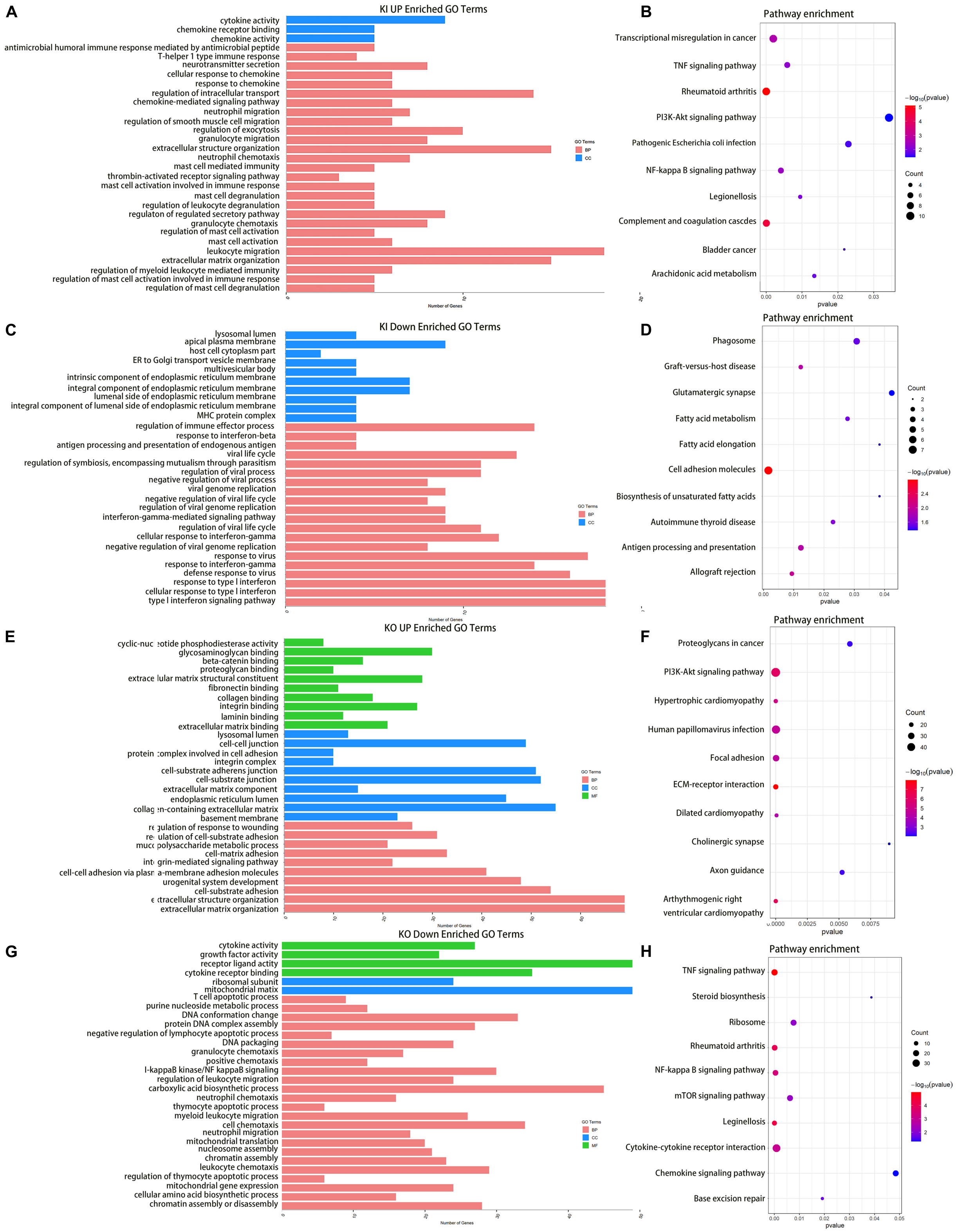


FIGURE 2 | Enrichment Analysis for DEGs upon knocking out and overexpressing U11 in T24 cell lines. **(A,C)** GO pathways of upregulated and downregulated genes in T24^{U11-KI} cell line, respectively. **(B,D)** KEGG pathways of upregulated and downregulated genes in T24^{U11-KI} cell line, respectively. Panels **(E,G)** are GO pathways of upregulated and downregulated genes in T24^{U11-KO} cell line, respectively. Panels **(F,H)** are KEGG pathways of upregulated and downregulated genes in T24^{U11-KO} cell line, respectively. GO: gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes.

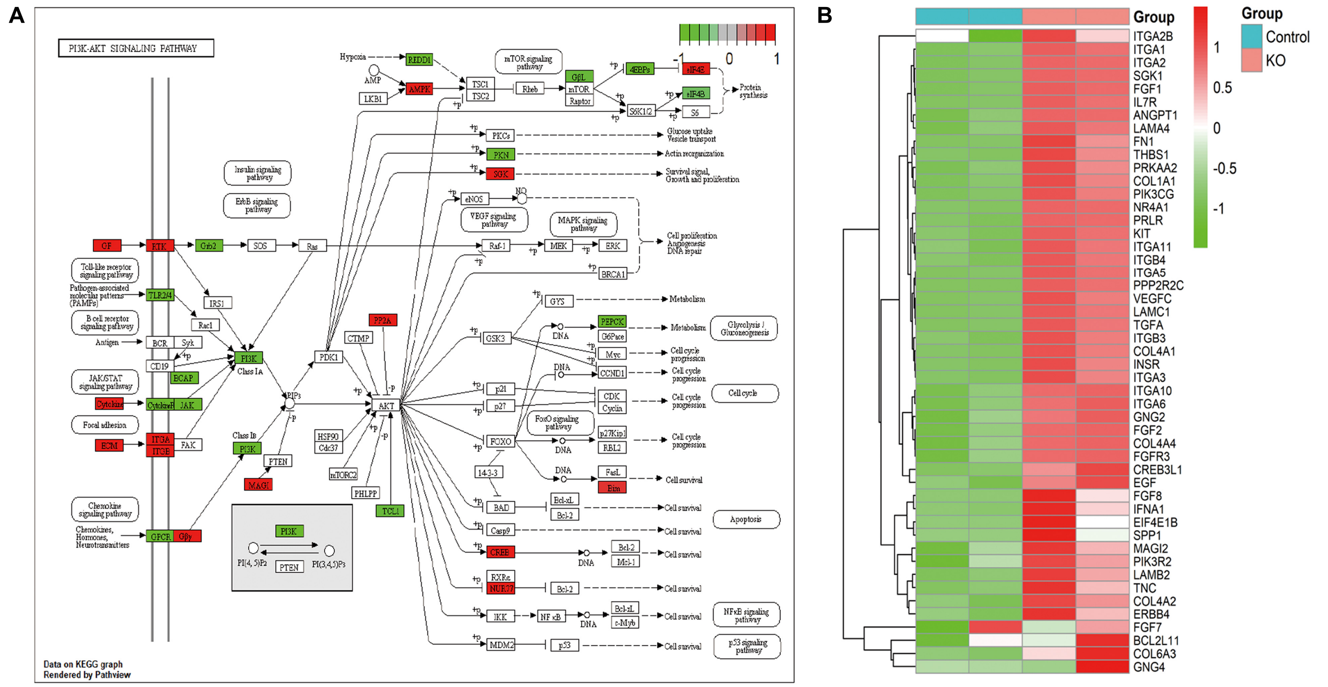


FIGURE 3 | PI3K-Akt signaling pathway for DEGs from knocking out U11. **(A)** PI3K-Akt signaling pathway diagram. **(B)** Heatmap from DEGs in this pathway.

their impact on tumor stage using TCGA database. Interestingly, overall survival (OS) of patients with bladder cancer with high FN1 expression was significantly lower than that with low expression (P = 0.012), while the OS of bladder cancer patients with low RPL22L1 expression was significantly lower than that of patients with high expression (P = 0.034) (**Figure 7A**). The other four genes had no significant effect on the prognosis of bladder cancer patients. The six hub genes also have an impact on bladder cancer stage. The expression of FN1, TIMP1, TGFB2, and RPL22L1 in Stage II was significantly lower than that in Stage III ($P < 0.05$), while the other two genes were not statistically different in stage (**Figure 7B**).

DISCUSSION

SnRNAs are a class of non-coding RNAs whose length ranges from 100 to 215 nucleotides in mammals, mainly including U1, U2, U3, U4, U5, U6, and U11 genes. SnRNAs have been present in the nucleus of mammalian cells, and together with more than 40 intranuclear proteins form RNA spliceosomes (Dvinge et al., 2019; Suzuki et al., 2019), which catalyze the maturation of precursor mRNAs in mammals, thereby affecting gene expression and leading to proliferation or apoptosis of cancer cells. We found that small nuclear RNAs mediate the formation of long-range chromatin interactions, of which U11 (RNU11) may be the most significant small nuclear RNA. To explore the gene expression changes and its potential effects mediated by U11 snRNA in bladder cancer cells, U11 snRNA knockout and overexpressed cell lines were constructed and further used to analyze the gene expression changes by RNA sequencing. Interestingly, we found that both up-regulated genes in the U11 overexpression group and down-regulated genes in the U11 knockout group were mainly enriched in cancer-related pathways, such as NF-kappa B signaling pathway, TNF signaling pathway and Bladder cancer. Protein interaction network analysis predicted that CXC chemokine family (CXCL2, CXCL3, CXCL6, and CXCL8) were hub genes. Further alternative splicing analysis also found that both the U11 knockout group and the U11 overexpression group caused alternative splicing events in genes with different expression, including some genes in the PI3K-Akt signaling pathway, such as FN1 and FGF1 genes, and other oncogenes, such as TGFB2, TIMP1, and MDM2. Our results suggest that U11 may affect the expression of cancer-related genes and be implicated in bladder carcinogenesis by affecting alternative splicing.

NF-kappa B is a heterodimer composed mainly of p65 and P50 proteins, and its function is to induce the transcription factors of inflammatory cytokines and anti-apoptotic proteins. In most cells, NF-kappa B mediates cell survival signals and protects cells from apoptosis (Jung and Dritschilo, 2001). Increasing evidence suggests that activation of NF-kappa B is associated with apoptosis, expression of angiogenic proteins, and carcinogenesis due to its fundamental effects on the dedifferentiation and proliferation of malignant tumor cells (Dorai and Aggarwal, 2004; Umezawa, 2006). Related studies have found that NF-kappa B activation is associated with urogenital cancers, such as prostate cancer and renal cell carcinoma (Oya et al., 2003; Ross et al., 2004; Domingo-Domenech et al., 2005). Similarly, in bladder tumors, the effect of NF-kappa B activation on tumorigenesis has also been reported (Levidou et al., 2008) and in our study, both

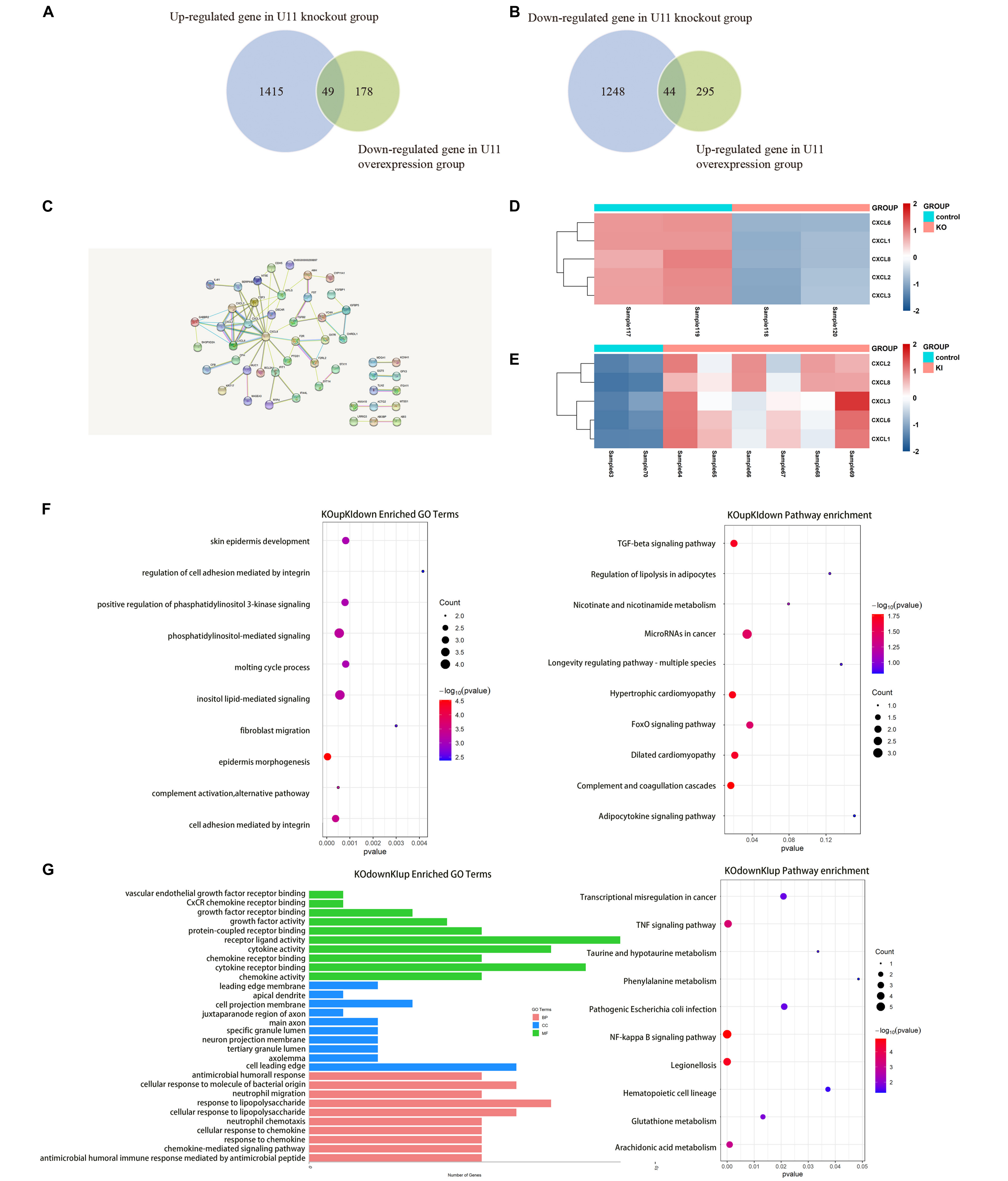


FIGURE 4 | Comprehensive analysis of DEGs upon knocking out and overexpressing U11 in T24 cell lines. **(A,B)** Venn diagrams of the intersection of differentially expressed genes between U11 overexpression group and knockout group. **(C)** Protein–protein interactions (PPI) of intersection gene. **(D)** The heatmap of CXCL family genes in T24^{U11-KO} and T24^{WT} cell lines. **(E)** The heatmap of CXCL family genes in T24^{U11-KI} and T24^{WT} cell lines. **(F)** Pathways from intersecting genes of up-regulated genes in T24^{U11-KO} cell line and down-regulated genes in T24^{U11-KI} cell line. **(G)** Pathways from intersecting genes of down-regulated genes in T24^{WT} cell line and up-regulated genes in T24^{U11-KI} cell line.

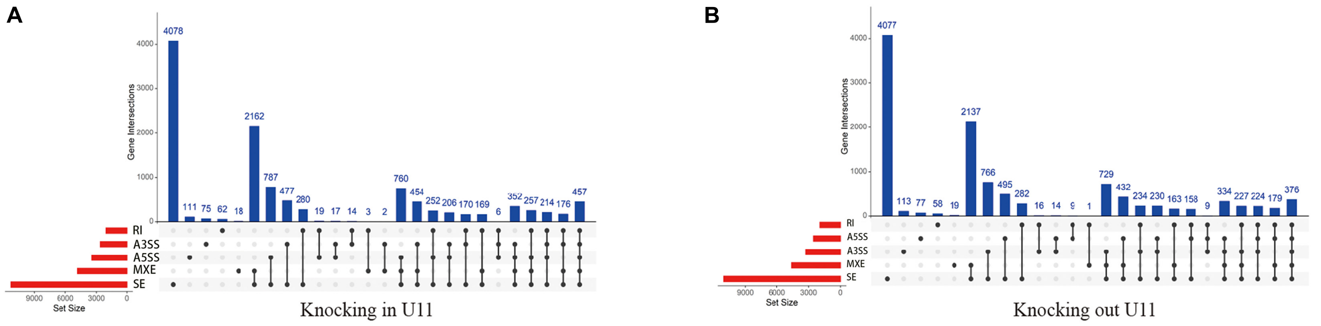


FIGURE 5 | The identification of alternative splicing events. **(A,B)** Up-set plots of significant alternative splicing events upon knockout and overexpressed U11 in T24 cells. Skipped exon (SE), alternative 5′ splice site (A5SS), alternative 3′ splice site (A3SS), mutually exclusive exons (MXE), and retained intron (RI).

up-regulated genes in the U11 overexpression group and down-regulated genes in the U11 knockout group were mainly enriched in the NF-kappa B signaling pathway.

PI3K-Akt pathway is a downstream signal transduction pathway of various cytokines and growth factors, which is involved in regulating cell proliferation and apoptosis (Lim and Counter, 2005; Bleau et al., 2009). PI3K belongs to the phospholipid kinase family and can be activated by many extracellular factors to participate in the cellular response. Activated PI3K phosphorylates PIP2–PIP3, thereby activating its downstream target kinase Akt. Activated Akt is ectopic from the cell membrane to the nucleus and cytoplasm. It activates or inhibits downstream target proteins, further promotes cell proliferation, apoptosis and energy metabolism, and is closely related to tumorigenesis and development (Franke et al., 2003). It was found that PI3K-Akt signaling pathway plays an important role in the occurrence, development and progression of malignant tumors, and the activation of Akt is closely related to the proliferation, migration and invasion of tumor cells (Roncolato et al., 2019; Ma et al., 2020). In this study, after knocking out U11, intersecting genes with significant differential alternative splicing and abnormal expression were mainly enriched in the PI3K-Akt signaling pathway.

Among them, fibronectin (FN) on the PI3K-Akt signaling pathway is a high molecular weight extracellular matrix glycoprotein with a molecular weight of 440 kDa. Its molecular structure contains a variety of domains, which can selectively bind to a variety of macromolecules in the extracellular matrix, such as collagen, heparin, fibrin and cell surface receptors, and play a crucial role in cell adhesion, migration, growth, differentiation and other cell overgrowth (Oya et al., 2003; Li et al., 2019). FN1 is a member of the FN family and plays a variety of biological functions in tumors, atherosclerosis, arthritis, and other diseases (Castelletti et al., 2008). Recent studies have found that FN1 is an important regulator promoting the formation and development of a variety of cancer cells, such as glioblastoma, laryngeal cancer and cutaneous squamous cell carcinoma (Jerhammar et al., 2010; Liao et al., 2018). In breast cancer, FN1 activates specific matrix metalloproteinases to promote breast cancer cell invasion and metastasis (Qian et al., 2011). It has been reported that the combination of miR-200c and FN1 can effectively inhibit the development of endometrial cancer in terms of FN1 expression in endometrial cancer cells (Howe et al., 2011). MiR-200c inhibits the expression of FNI, significantly reduces cell proliferation, and inhibits migration and invasion, suggesting that the expression of FN1 is a good indicator of the state of cancer cells. FN1 affects proliferation, senescence and apoptosis of human glioma cells through PI3K-AKT signaling pathway (Liao et al., 2018). Down-regulation of FN1 can inhibit proliferation, migration and invasion, thereby inhibiting the occurrence of colorectal cancer (Cai et al., 2018). Interestingly, we found that the expression of FN1 increased significantly after the knockout of U11, and FN1 underwent several meaningful alternative splicing events, including three alternative 3′ splicing site (A3SS) events, one mutually exclusive exon event, and four exon skipping events. Although we examined three typical exons skipping of FN1 (EDA, EDB, and IIICS) as previously reported (Lopez-Mejia et al., 2013), no remarkably alternative splice events were detected in FN1 gene. Due to the numerous exons of FN1 and thus generating multiple alternative splicing events, prominent bands of these AS events were extremely challenged to detect using conventional PCR. More advanced testing technologies and further versus nested PCR experiments should be conducted to detect the multiple alternative splicing events.

Chemokines essentially belong to a class of small molecule proteins, whose initial role is mainly to participate in the directional chemotaxis of leukocytes to inflammatory sites. The role of chemokines and their receptors in the process of tumorigenesis and development cannot be ignored increasingly. A large number of studies have shown that the regularity of malignant tumor cell metastasis is similar to that of chemokine migration during inflammatory cell metastasis (Ha et al., 2017). CXCL8 is an important member of the chemokine family. It was first discovered by Yoshimura in 1987 in the culture supernatant of human peripheral blood mononuclear cells stimulated with lipopolysaccharide (Yoshimura et al., 1987) and formally named IL-8/NAP (IL-8 NAP neutrophil active peptide) in 1988. At present, studies have confirmed that CXCL8 is highly expressed in thyroid tumors, ovarian cancer, liver cancer, prostate cancer and many other tumors, and its role is mainly reflected in: accelerating the growth of tumor cells, enhancing the motility

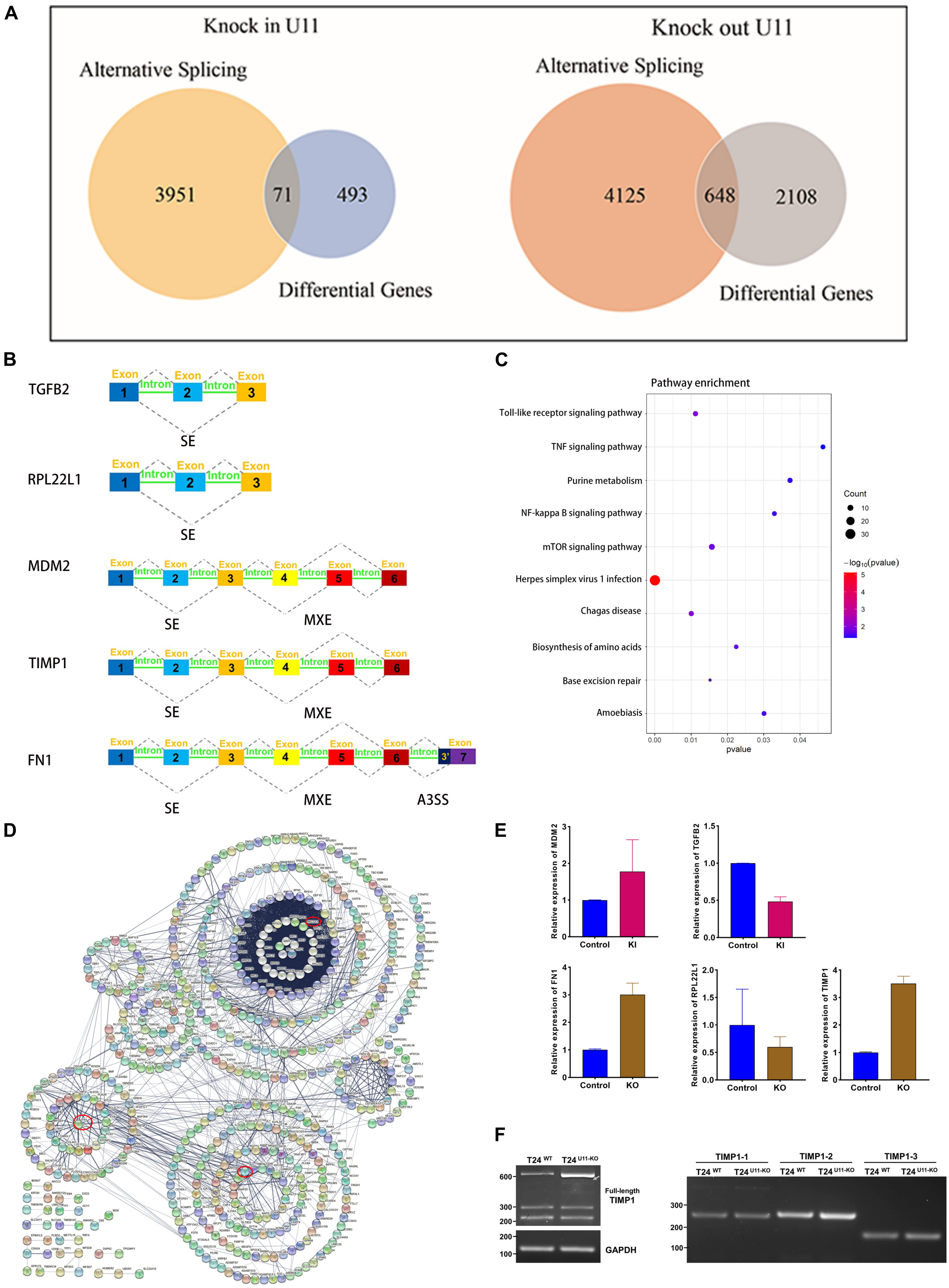


FIGURE 6 | Comprehensive analysis of alternative splicing events and differential genes upon knocking out and overexpressing U11 in T24 cell lines. **(A)** Venn diagrams of DEGs and genes with significant alternative splicing events. **(B)** Diagrams of alternative splicing events in MDM2, TGFB2, RPL22L1, FN1, and TIMP1. Mutually exclusive exons (MXE) and skipped exon (SE) in MDM2; SE in TGFB2; SE in RPL22L1; MXE and SE in TIMP1; MXE, SE and alternative 3′ splice site (A3SS) in FN1. **(C)** Pathways from intersecting genes between $T24^{U11-KO}$ and $T24^{WT}$ cell lines. **(D)** Venn diagram of Protein–protein interactions (PPI) of intersecting genes. **(E)** The relative expression levels of MDM2 and TGFB2 between $T24^{U11-KI}$ and $T24^{WT}$ cell lines and the relative expression levels of FN1, RPL22L1, and TIMP1 between $T24^{U11-KO}$ and $T24^{WT}$ cell lines. **(F)** Alternative splice identification of TIMP1-full length, TIMP1-1, TIMP1-2, and TIMP1-3 in $T24^{U11-KO}$ and $T24^{WT}$ cell lines.

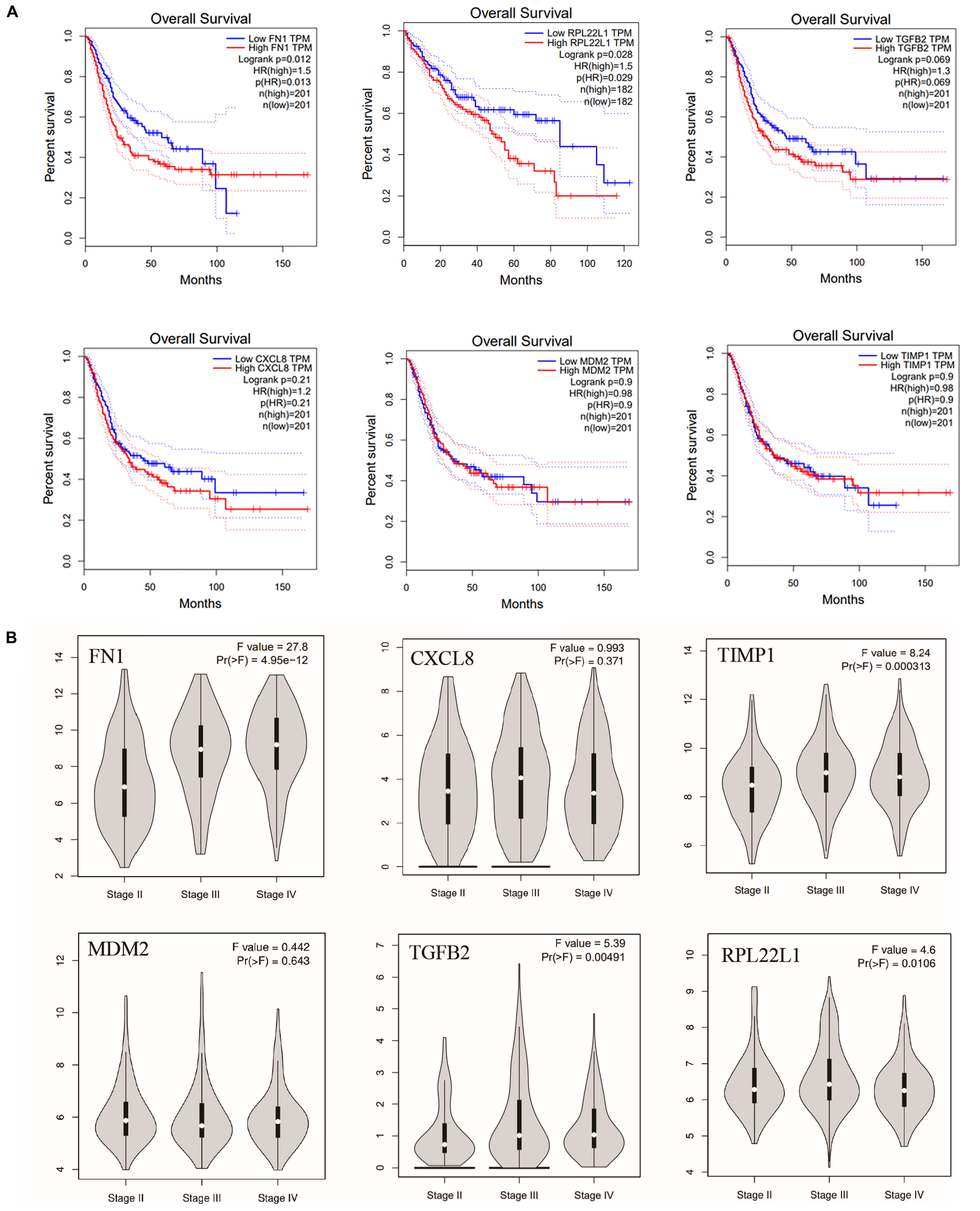


FIGURE 7 | (A) Overall survival of the six hub genes in bladder cancer based on TCGA database, including FN1, MDM2, TGFB2, CXCL8, TIMP1, and RPL22L1. OS: overall survival. **(B)** The effect of six hub genes on bladder cancer stage.

of tumor cells, changing the local environment of tumors and inhibiting the immune system to play a role, and ultimately making tumor cells invade and metastasize in distant areas (Liu J. et al., 2016). In this study, we found that CXCL8 expression increased significantly after overexpression of U11, while decreased significantly after knockout of U11. Taking the intersection of the differential genes in the knockout and overexpression groups and predicting by protein interaction network analysis, we found that the chemokine family was the hub gene group, especially CXCL8. It can be seen that the expression of U11 affects the expression of the chemokine family, especially CXCL8. These results showed that the overexpression of U11 could promote the expression of chemokines, thereby promoting cell proliferation and tumor metastasis (Liu Q. et al., 2016). However, in alternative splicing analysis after knockout or overexpression of U11, we did not find significant splicing events in CXCL8, suggesting that the change in expression of U11 to CXCL8 may not be through the regulation of gene splicing. In summary, our results predict that U11 plays an important role in the regulation of chemokine expression in bladder cancer cells, but the specific mechanism is unknown.

Murine double minute 2 is a tumor protein that is highly expressed in tumors. In cancer cells, MDM2 proteins help to modify biological programs, enhance growth-promoting signals, and reduce apoptotic signals. P53 protein is a very important tumor suppressor and plays an important role in regulating cell cycle, apoptosis, DNA damage repair, angiogenesis, cell metabolism and aging (Prives, 1998; Gupta et al., 2019). In more than half of human tumors, the p53 gene is mutated or deleted, while in the remaining human tumors, there is wild-type p53, whose function is also effectively inhibited by MDM2. E3 ubiquitin ligase MDM2 is an important inhibitor of p53, which can block the transcriptional function of p53, promote the transfer of p53 from the nucleus to the cytoplasm, and ubiquitinate and degrade p53 (Brooks and Gu, 2006). In this study, we found that after overexpression of U11, MDM2 expression increased significantly, and MDM2 had a meaningful mutually exclusive exon and an exon skipping event. MDM2 was well-known as the most critical negative regulator of p53 pathway, whether the alternative splicing events of MDM2 directly or indirectly regulated by U11 deserved to be further investigated.

In summary, we found that U11 may alter gene expression by affecting the PI3K-Akt signaling pathway and NF-kappa B signaling pathway. U11 may be involved in the regulation of gene expression in bladder cancer cells, which may provide a novel biomarker for clinical diagnosis and treatment of bladder cancer.

MATERIALS AND METHODS

Cell Lines and Cell Culture

T24 bladder cancer cell line was purchased from Cell Bank of Shanghai Institute of Cell Biology (Shanghai, China). T24-FL and T24-SLT cell lines were gifts from Dr. Gordon Hager (NIH, United States). All cells were cultured in F12 Medium (Gibco, China) supplemented with penicillin, streptomycin and 10% FBS (Gibco, Australia). All cell lines were maintained at 37°C in a humidified atmosphere containing 5% CO_2.

Design of sgRNA and Construction of Its Expression Vector

Using an online CRISPR design tool, two sgRNAs targeting the U11 region were designed by selecting the sgRNA sequences with high scores. The sequence is as follows:

SG1-F: 5′-CACC GCTGTCGTGAGTGGCACACGT-3′
SG1-R: 5′-AAAC ACGTGTGCCACTCACGACAGC-3′
SG2-F: 5′-CACC GCAGCTGGTGATCGTTGGTCC-3′
SG2-F: 5′-AAAC GGACCAACGATCACCAGCTGC-3′

Sequencing primers were designed with the location of sgRNA as the center. The sgRNA expression vector was cloned into pX330 all in one vector by the *Bbs*I digestion site so that the vector could express CAS9 protein and corresponding sgRNA.

Cell Transfection

1.5×10^5 T24 cells/well were plated on 24-well plates and transfected after 12–16 h (500 ng for each of the two sgRNA vectors; sgRNA vector was used as the control group). After 6–8 h, the liquid was changed. 48 h after transfection, 2 ug/ml Puro was added into the fresh medium for screening for 48 h. Cells were then grown in a fresh medium at 37°C in a humidified incubator containing 5c/o CO_2 and collected 24 h later. The cells were subjected to genome extraction, and the other part of the cells were cloned in 96-well plates.

Knockout of Small Nuclear RNA U11 in T24 Bladder Cancer Cells Using CRISPR/Cas9 Gene-Editing Technology

Genomic DNA was extracted with Tiangen Genome Extraction Kit and PCR was performed with sequencing primers. Three monoclonal cell lines were selected, and total RNA was extracted by the Tiangen RNA extraction kit for reverse transcription. Subsequently, the expression of U11 in wild-type and U11-knockout cells was detected by fluorescence quantitative PCR, and the knockout efficiency of U11 was identified. The monoclonal cell lines with the highest knockout efficiency were selected to construct the U11 knockout model.

Construction of Stable Overexpressing U11 Bladder Cancer T24 Cells by pcDNA-U11 Recombinant Plasmid Transfection

RNA from T24 cells was extracted and reverse transcribed into cDNA. Using this cDNA as a template, the U11 gene sequence was amplified with specific PCR primers. The U11 fragment and pcDNA3 were digested by *Hin*dIII and *Kpn*I restriction endonucleases. The product was ligated by T4 DNA ligase and transformed into E. coli DH5α cells. The plasmid identified by sequencing was named pcDNA-U11. After transfection, identification, and monoclonal selection, the culture was expanded.

Cell Proliferation Assay

T24^{WT} and T24^{U11-KO} cells in logarithmic growth phase were digested and seeded into 96-well plates at the concentration of 3,500 cells per well. For the MTT assay, cells were cultured for 0, 24, 48, 72, and 96 h, respectively and then 10 ul MTT (5 mg/mL) was added into each well for another 4 h at 37°C. MTT solution was then removed, and MTT formazan dissolved in 100 μL dimethyl sulfoxide (DMSO) for detection of the absorbance at 490 nm.

Immunofluorescence

To detect expression of CBs at cellular levels, IF localization was conducted according to standard procedures. First, cells were fixed with 4% paraformaldehyde for 20 min at room temperature and permeabilized with 0.2% Triton X-100 for 10 min on ice. Then, cells were washed three times with PBS. Subsequently, cells were incubated with anti-coilin antibody (Cat# 10967-1-AP, Proteintech, United States) for 1 h, and followed by incubation with secondary antibody (Goat Anti-Mouse IgG, DyLight 488). Cells were then counterstained with DAPI after washing three times with PBS. Multicolor imaging was performed and captured utilizing an IX70 microscope at 20× magnification (Olympus, Japan).

mRNA Library Construction and Sequencing

A total of 12 samples were taken for RNA sequencing, including 6 overexpressed U11 samples (sample64-smaple69), 2 knockout U11 samples (sample118, sample120), and 4 control samples (sample63, sample70, sample117, and sample119). Total RNA was extracted from the control group, U11 knockout bladder cancer cells, and U11 overexpressing bladder cancer cells cultured *in vitro*. The ribosome RNA was removed by ribosome RNA depletion kit and then reverse transcribed into cDNA for second-strand synthesis. dsDNA is interrupted by ultrasound to grow uniform fragments. The fragments were flattened, 3′A bases were generated, and the adaptor was ligated to complete the construction of the RNA-seq high-throughput sequencing library. High-throughput sequencing via Illumina HiSeq2000 platform. All operations were performed by Shanghai WUXI NEXTCODE. Sequencing was performed using the Illumina system, following the protocol provided by Illumina, with 2 × 150 paired-end sequencing.

The Analysis of Alternative Splicing

After comparing the data, the file is converted to bam format using Samtools. Then alternative splice analysis was performed using rMATS 4.0.2. rMATS is a software for differential alternative splice analysis of RNA-seq data. The rMATS statistical model was used to quantify the expression of alternative splice events in different samples, and then the *P*-value was calculated by the Likelihood Ratio Test to represent the differences in LncLevel (Inclusion Level) between the two groups of samples. There are five kinds of alternative splice events recognized by rMATS, respectively skipped exon (SE), alternative 5 splice site (A5SS), A3SS, mutually exclusive exons (MXE), retained intron (RI) (Shen et al., 2014). The detailed results of alternative splicing in T24 U11-KI and T24 U11-KO cell lines are presented in **Supplementary Materials**.

RNA Extraction, PCR, and DNA Agarose GEL Electrophoresis

Total RNA was extracted from the control group, U11 knockout bladder cancer cells, and U11 overexpressing bladder cancer cells cultured *in vitro* using Trizol Reagent (Invitrogen), and reverse-transcribed to cDNA using PrimeSciptTM RT reagent Kit with Gdna Eraser (Takara, China) following the manufacturer's instructions. Primer sequence were displayed in **Table 1**.

PCR was performed in 20 ul reactions containing 10 ul 2× ES Taq MasterMix (CW BIO, China), 3.4 ul H_2O, 0.8 ul of each gene-specific primer and 5 ul cDNA. Reaction conditions were 30 cycles of 94°C for 2 min, 60°C for 30s and 72°C for 30s. PCR products were separated by 3c/o gel electrophoresis. Then Image Quant LAS 500 was used for exposure.

Data Processing and Bioinformatics Analysis

Sequencing data is Illumina raw data of RNA-seq. Fastqc is used to evaluate the quality of raw data. Fastp is used for data pre-processing, including removing adapter components and effectively correcting lower quality bases. After fastp treatment, Fastqc detects data quality again and obtains qualified clean data. Clean data were aligned to the reference genome hg38 using Hisat2, gene expression was obtained by Stringtie, and differential genes were finally obtained by the Deseq2 R package ($|\text{Fold change}| \geq 1.5$ and $P < 0.05$). Gene ontology (GO) and The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were used for DEGs using the clusterProfiler R package, and significant enrichment was defined as $P < 0.05$. Cytoscape was used to construct the U11 regulatory network. The prognostic analysis of core genes was performed using GEPIA based on the TCGA database.

TABLE 1 | Primers used for PCR.

Gene	Forward primer 5′∼3′	Reverse primer 5′∼3′
TIMP1 Full-length	CCCTAGCGTGGACATTTATC	AAGGTGACGGGACTGGAAG
TIMP1-1	CCCTAGCGTGGACATTTATC	GGTATAAGGTGGTCTGGTTG
TIMP1-2	ACTTCCACAGGTCCCACAAC	AAGGTGACGGGACTGGAAG
TIMP1-3	CTTCTGGCATCCTGTTGTTG	GGTATAAGGTGGTCTGGTTG
GAPDH	GTGAACCATGAGAAGTATGACAAC	CATGAGTCCTTCCACGATACC
snRNA U11	AGATAGGTAATACGACTCACTATAG	TTAACCCTCACTAAAGGGAAGAA
	GGAAAAAGGGCTTCTGTCGTGAGTG	AGGGCGCCGGGACC

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the Gene Expression Omnibus database (GEO) under accession number: GSE171744.

AUTHOR CONTRIBUTIONS

QW and ZT conceived the study. YX, ZW, and XW wrote the manuscript. YW, ST, CF, and LP analyzed the results. QL and YT edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (82060512, 82060460, and 31471271), the Guangxi Natural Science Fund for Innovation Research Team (2016GXNSFGA380006), and Guangxi Science and Technology Base and Talent Special Fund (2019AC17009 and AD17195090).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.695597/full#supplementary-material

REFERENCES

Andrade, L. E., Chan, E. K., Raska, I., Peebles, C. L., Roos, G., and Tan, E. M. (1991). Human autoantibody to a novel protein of the nuclear coiled body: immunological characterization and cDNA cloning of p80-coilin. *J. Exp. Med.* 173, 1407–1419. doi: 10.1084/jem.173.6.1407

Bleau, A. M., Hambardzumyan, D., Ozawa, T., Fomchenko, E. I., Huse, J. T., Brennan, C. W., et al. (2009). PTEN/PI3K/Akt pathway regulates the side population phenotype and ABCG2 activity in glioma tumor stem-like cells. *Cell Stem Cell* 4, 226–235. doi: 10.1016/j.stem.2009.01.007

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Brooks, C. L., and Gu, W. (2006). p53 ubiquitination: Mdm2 and beyond. *Mol. Cell* 21, 307–315. doi: 10.1016/j.molcel.2006.01.020

Cai, X., Liu, C., Zhang, T. N., Zhu, Y. W., Dong, X., and Xue, P. (2018). Down-regulation of FN1 inhibits colorectal carcinogenesis by suppressing proliferation, migration, and invasion. *J. Cell Biochem.* 119, 4717–4728. doi: 10.1002/jcb.26651

Castelletti, F., Donadelli, R., Banterla, F., Hildebrandt, F., Zipfel, P. F., Bresin, E., et al. (2008). Mutations in FN1 cause glomerulopathy with fibronectin deposits. *Proc. Natl. Acad. Sci. U.S.A.* 105, 2538–2543. doi: 10.1073/pnas.0707730105

Domingo-Domenech, J., Mellado, B., Ferrer, B., Truan, D., Codony-Servat, J., Sauleda, S., et al. (2005). Activation of nuclear factor-kappaB in human prostate carcinogenesis and association to biochemical relapse. *Br. J. Cancer* 93, 1285–1294. doi: 10.1038/sj.bjc.6602851

Dorai, T., and Aggarwal, B. B. (2004). Role of chemopreventive agents in cancer therapy. *Cancer Lett.* 215, 129–140. doi: 10.1016/j.canlet.2004.07.013

Dvinge, H., Guenthoer, J., Porter, P. L., and Bradley, R. K. (2019). RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing. *Genome Res.* 29, 1591–1604. doi: 10.1101/gr.246678.118

Feng, R. M., Zong, Y. N., Cao, S. M., and Xu, R. H. (2019). Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? *Cancer Commun.* 39:22. doi: 10.1186/s40880-019-0368-6

Franke, T. F., Hornik, C. P., Segev, L., Shostak, G. A., and Sugimoto, C. (2003). PI3K/Akt and apoptosis: size matters. *Oncogene* 22, 8983–8998. doi: 10.1038/sj.onc.1207115

Gupta, A., Behl, T., Heer, H. R., Deshmukh, R., and Sharma, P. L. (2019). Mdm2-P53 interaction inhibitor with cisplatin enhances apoptosis in colon and prostate cancer cells In-Vitro. *Asian Pac. J. Cancer Prev.* 20, 3341–3351. doi: 10.31557/APJCP.2019.20.11.3341

Ha, H., Debnath, B., and Neamati, N. (2017). Role of the CXCL8-CXCR1/2 Axis in Cancer and inflammatory diseases. *Theranostics* 7, 1543–1588. doi: 10.7150/thno.15625

Hearst, S. M., Gilder, A. S., Negi, S. S., Davis, M. D., George, E. M., Whittom, A. A., et al. (2009). Cajal-body formation correlates with differential coilin phosphorylation in primary and transformed cell lines. *J. Cell Sci.* 122(Pt 11), 1872–1881. doi: 10.1242/jcs.044040

Hebert, M. D. (2010). Phosphorylation and the Cajal body: modification in search of function. *Arch. Biochem. Biophys.* 496, 69–76. doi: 10.1016/j.abb.2010.02.012

Hebert, M. D. (2013). Signals controlling Cajal body assembly and function. *Int. J. Biochem. Cell Biol.* 45, 1314–1317. doi: 10.1016/j.biocel.2013.03.019

Howe, E. N., Cochrane, D. R., and Richer, J. K. (2011). Targets of miR-200c mediate suppression of cell motility and anoikis resistance. *Breast Cancer Res.* 13:R45. doi: 10.1186/bcr2867

Jeppesen, D. K., Nawrocki, A., Jensen, S. G., Thorsen, K., Whitehead, B., Howard, K. A., et al. (2014). Quantitative proteomics of fractionated membrane and lumen exosome proteins from isogenic metastatic and nonmetastatic bladder cancer cells reveal differential expression of EMT factors. *Proteomics* 14, 699–712. doi: 10.1002/pmic.201300452

Jerhammar, F., Ceder, R., Garvin, S., Grénman, R., Grafström, R. C., and Roberg, K. (2010). Fibronectin 1 is a potential biomarker for radioresistance in head and neck squamous cell carcinoma. *Cancer Biol. Ther.* 10, 1244–1251. doi: 10.4161/cbt.10.12.13432

Jung, M., and Dritschilo, A. (2001). NF-kappa B signaling pathway as a target for human tumor radiosensitization. *Semin. Radiat. Oncol.* 11, 346–351. doi: 10.1053/srao.2001.26034

Levidou, G., Saetta, A. A., Korkolopoulou, P., Papanastasiou, P., Gioti, K., Pavlopoulos, P., et al. (2008). Clinical significance of nuclear factor (NF)-kappaB levels in urothelial carcinoma of the urinary bladder. *Virchows. Arch.* 452, 295–304. doi: 10.1007/s00428-007-0560-y

Li, B., Shen, W., Peng, H., Li, Y., Chen, F., Zheng, L., et al. (2019). Fibronectin 1 promotes melanoma proliferation and metastasis by inhibiting apoptosis and regulating EMT. *Onco Targets Ther.* 12, 3207–3221. doi: 10.2147/OTT.S195703

Liang, Z., Mou, Q., Pan, Z., Zhang, Q., Gao, G., Cao, Y., et al. (2019). Identification of candidate diagnostic and prognostic biomarkers for human prostate cancer: RPL22L1 and RPS21. *Med. Oncol.* 36:56. doi: 10.1007/s12032-019-1283-z

Liao, Y. X., Zhang, Z. P., Zhao, J., and Liu, J. P. (2018). Effects of fibronectin 1 on cell proliferation, senescence and apoptosis of human glioma cells through the PI3K/AKT signaling pathway. *Cell Physiol. Biochem.* 48, 1382–1396. doi: 10.1159/000492096

Lim, K. H., and Counter, C. M. (2005). Reduction in the requirement of oncogenic Ras signaling to activation of PI3K/AKT pathway during tumor maintenance. *Cancer Cell* 8, 381–392. doi: 10.1016/j.ccr.2005.10.014

Liu, J., Xu, R., and Zhao, X. (2016). [Mechanisms for effect of osthole on inhibiting the growth and invasion of bladder cancer cells]. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* 41, 345–352. doi: 10.11817/j.issn.1672-7347.2016.04.002

Liu, Q., Li, A., Tian, Y., Wu, J. D., Liu, Y., Li, T., et al. (2016). The CXCL8-CXCR1/2 pathways in cancer. *Cytokine Growth Factor Rev.* 31, 61–71. doi: 10.1016/j.cytogfr.2016.08.002

Lopez-Mejia, I. C., De Toledo, M., Della Seta, F., Fafet, P., Rebouissou, C., Deleuze, V., et al. (2013). Tissue-specific and SRSF1-dependent splicing of fibronectin, a matrix protein that controls host cell invasion. *Mol. Biol. Cell* 24, 3164–3176. doi: 10.1091/mbc.E13-03-0142

Lui, L., and Lowe, T. (2013). Small nucleolar RNAs and RNA-guided post-transcriptional modification. *Essays Biochem.* 54, 53–77. doi: 10.1042/bse0540053

Ma, Z., Yu, R., Zhu, Q., Sun, L., Jian, L., Wang, X., et al. (2020). CXCL16/CXCR6 axis promotes bleomycin-induced fibrotic process in MRC-5 cells via the PI3K/AKT/FOXO3a pathway. *Int. Immunopharmacol.* 81:106035. doi: 10.1016/j.intimp.2019.106035

Machyna, M., Heyn, P., and Neugebauer, K. M. (2013). Cajal bodies: where form meets function. *Wiley Interdiscip. Rev. RNA* 4, 17–34. doi: 10.1002/wrna.1139

Nicholson, B. E., Frierson, H. F., Conaway, M. R., Seraj, J. M., Harding, M. A., Hampton, G. M., et al. (2004). Profiling the evolution of human metastatic bladder cancer. *Cancer Res.* 64, 7813–7821. doi: 10.1158/0008-5472.Can-04-0826

Oya, M., Takayanagi, A., Horiguchi, A., Mizuno, R., Ohtsubo, M., Marumo, K., et al. (2003). Increased nuclear factor-kappa B activation is related to the tumor development of renal cell carcinoma. *Carcinogenesis* 24, 377–384. doi: 10.1093/carcin/24.3.377

Ploeg, M., Aben, K. K., and Kiemeney, L. A. (2009). The present and future burden of urinary bladder cancer in the world. *World J. Urol.* 27, 289–293. doi: 10.1007/s00345-009-0383-3

Prives, C. (1998). Signaling to p53: breaking the MDM2-p53 circuit. *Cell* 95, 5–8. doi: 10.1016/s0092-8674(00)81774-2

Qian, P., Zuo, Z., Wu, Z., Meng, X., Li, G., Wu, Z., et al. (2011). Pivotal role of reduced let-7g expression in breast cancer invasion and metastasis. *Cancer Res.* 71, 6463–6474. doi: 10.1158/0008-5472.CAN-11-1322

Roncolato, F., Lindemann, K., Willson, M. L., Martyn, J., and Mileshkin, L. (2019). PI3K/AKT/mTOR inhibitors for advanced or recurrent endometrial cancer. *Cochrane Database Syst. Rev.* 10:CD012160. doi: 10.1002/14651858.CD012160.pub2

Ross, J. S., Kallakury, B. V., Sheehan, C. E., Fisher, H. A., Kaufman, R. P. Jr., Kaur, P., et al. (2004). Expression of nuclear factor-kappa B and I kappa B alpha proteins in prostatic adenocarcinomas: correlation of nuclear factor-kappa B immunoreactivity with disease recurrence. *Clin. Cancer Res.* 10, 2466–2472. doi: 10.1158/1078-0432.ccr-0543-3

Salim, E. I., Moore, M. A., Bener, A., Habib, O. S., Seif-Eldin, I. A., and Sobue, T. (2010). Cancer epidemiology in South-West Asia - past, present and future. *Asian Pac. J. Cancer Prev.* 11(Suppl. 2), 33–48.

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., et al. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5593–E5601. doi: 10.1073/pnas.1419161111

Strzelecka, M., Trowitzsch, S., Weber, G., Lührmann, R., Oates, A. C., and Neugebauer, K. M. (2010). Coilin-dependent snRNP assembly is essential for zebrafish embryogenesis. *Nat. Struct. Mol. Biol.* 17, 403–409. doi: 10.1038/nsmb.1783

Suzuki, H., Kumar, S. A., Shuai, S., Diaz-Navarro, A., Gutierrez-Fernandez, A., De Antonellis, P., et al. (2019). Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature* 574, 707–711. doi: 10.1038/s41586-019-1650-0

Umezawa, K. (2006). Inhibition of tumor growth by NF-kappaB inhibitors. *Cancer Sci.* 97, 990–995. doi: 10.1111/j.1349-7006.2006.00285.x

Usher, P. A., Sieuwerts, A. M., Bartels, A., Lademann, U., Nielsen, H. J., Holten-Andersen, L., et al. (2007). Identification of alternatively spliced TIMP-1 mRNA in cancer cell lines and colon cancer tissue. *Mol. Oncol.* 1, 205–215. doi: 10.1016/j.molonc.2007.05.002

Wang, Q., Sawyer, I. A., Sung, M. H., Sturgill, D., Shevtsov, S. P., Pegoraro, G., et al. (2016). Cajal bodies are linked to genome conformation. *Nat. Commun.* 7:10966. doi: 10.1038/ncomms10966

Yoshimura, T., Matsushima, K., Oppenheim, J. J., and Leonard, E. J. (1987). Neutrophil chemotactic factor produced by lipopolysaccharide (LPS)-stimulated human blood mononuclear leukocytes: partial characterization and separation from interleukin 1 (IL 1). *J. Immunol.* 139, 788–793.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 27 July 2021
doi: 10.3389/fgene.2021.665843

RFCell: A Gene Selection Approach for scRNA-seq Clustering Based on Permutation and Random Forest

*Yuan Zhao[1], Zhao-Yu Fang[2], Cui-Xiang Lin[1], Chao Deng[1], Yun-Pei Xu[1] and Hong-Dong Li[1]**

[1] Hunan Provincial Key Laboratory on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China, [2] School of Mathematics and Statistics, Central South University, Changsha, China

In recent years, the application of single cell RNA-seq (scRNA-seq) has become more and more popular in fields such as biology and medical research. Analyzing scRNA-seq data can discover complex cell populations and infer single-cell trajectories in cell development. Clustering is one of the most important methods to analyze scRNA-seq data. In this paper, we focus on improving scRNA-seq clustering through gene selection, which also reduces the dimensionality of scRNA-seq data. Studies have shown that gene selection for scRNA-seq data can improve clustering accuracy. Therefore, it is important to select genes with cell type specificity. Gene selection not only helps to reduce the dimensionality of scRNA-seq data, but also can improve cell type identification in combination with clustering methods. Here, we proposed RFCell, a supervised gene selection method, which is based on permutation and random forest classification. We first use RFCell and three existing gene selection methods to select gene sets on 10 scRNA-seq data sets. Then, three classical clustering algorithms are used to cluster the cells obtained by these gene selection methods. We found that the gene selection performance of RFCell was better than other gene selection methods.

Keywords: single-cell RNA sequencing, gene selection, permutation, random forest, clustering

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Junwei Luo,
Henan Polytechnic University, China
Yannan Bin,
Anhui University, China

****Correspondence:***
Hong-Dong Li
hongdong@csu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 February 2021*
Accepted: *01 April 2021*
Published: *27 July 2021*

Citation:
Zhao Y, Fang Z-Y, Lin C-X, Deng C, Xu Y-P and Li H-D (2021) RFCell: A Gene Selection Approach for scRNA-seq Clustering Based on Permutation and Random Forest. Front. Genet. 12:665843. doi: 10.3389/fgene.2021.665843

INTRODUCTION

Single cell RNA-Seq (scRNA-Seq) provides unprecedented insight into biological concerns at the level of individual cells (Hwang et al., 2018). Bulk RNA sequencing analysis, based on the average expression of large populations of cells, is difficult to reveal the expression heterogeneity between different cells. However, scRNA-Seq only studies the expression of single-cell level, so scRNA-Seq improves cell resolution across global transcriptome profile (Pouyan and Kostka, 2018). In recent years, scRNA-seq has been widely used in many aspects of biological and medical research (Hedlund and Deng, 2018), for example, discovering the new cell states and tracing the origin of its development (Trapnell, 2015), cell type identification (Xu and Su, 2015), heterogeneity of cell responses (Pollen et al., 2014), understanding of cell-specific biological characteristics (Poirion et al., 2016), building gene regulatory networks across the entire gene expression profiles (Zheng et al., 2019), tracking of different cell lineage trajectories (Shao and Hofer, 2017), and cell fate decisions (Goolam et al., 2016). In addition, scRNA-seq data is useful to study cellular immunity, drug and antibiotic resistance (Patel et al., 2014).

Genome-wide transcriptome analysis is usually used to study the expression of tissue, disease and cell type-specific genes, but generating expression profiles at single-cell resolution is technically challenging. Therefore, researchers have proposed many sequencing technologies, such as: a robust mRNA-Seq protocol that is applicable to a single cell level; and a scalable method to characterize many cell types and states under various conditions and disturbances Drop-seq protocol for complex organizations (Ramskold et al., 2012; Macosko et al., 2015)). From the perspective of scRNA-Seq technology, the scRNA-seq capture efficiency and dropout rate have limitations due to the small amount of starting materials. At the same time, due to the uncertainty of cell separation protocol, library preparation methods, sequencing methods, reagent usage methods, and various types of samples, batch effects may be introduced, which leads to the high noise characteristics of scRNA-seq data (Chen et al., 2019). From the perspective of gene expression, gene expression in scRNA-Seq data is specific (Aevermann et al., 2018), only a small part of the genes are biologically meaningful. So, scRNA-Seq research is challenging due to its high noise, high dimensionality and sparsity (Schnable et al., 2009). Considering that scRNA-seq data play an important role in the effectiveness and accuracy of downstream analysis, the most important goal of scRNA-seq is to select highly variable genes in the single cell transcriptome profiling.

scRNA-seq data usually has the problems of high noise, high dimensionality and sparseness. Therefore, before downstream analysis, researchers usually use certain feature selection methods to extract scRNA-seq data. A common gene selection strategy for high-dimensional gene expression analysis is by projecting data points from a high-dimensional gene expression space into a low-dimensional space. Single cell expression data in low-dimensional space is expected to be an important feature in high-dimensional space. In recent years, there have been many methods to analyze and study scRNA-seq data from the angles of reduce dimension. Principal component analysis (PCA) (Lever et al., 2017) is a method of converting scRNA-seq data into fewer features to achieve data dimensionality reduction. By generating two-dimensional embedding of high-dimensional data, t-distributed stochastic neighborhood embedding (t-SNE) (Linderman and Steinerberger, 2019) is an effective non-linear dimensionality reduction technology that has attracted more and more scientific attention. Recently, it has been widely popular in the field of scRNA-seq data research.

Andrews and Hemberg (2019) proposed a gene selection method called M3Drop. Wang et al. (2019) proposed a new marker selection strategy SCMarker to accurately delineate cell types in scRNA-seq data by identifying genes that have bi/multi-modally distributed expression levels and are co-or mutually-exclusively expressed with some other genes. In addition, Expr is also a gene selection method based on scRNA-Seq sequencing data. This method only retains the genes with the highest average expression (logarithmic normalized count) value in all cells.

We propose RFCell, a gene selection strategy based on permutation and random forest, which uses supervised classification in pattern recognition to determine the best subset of genes for cell type recognition without referring to any known transcriptome profile or cell related information. The central idea of our method is that random forests based on ensemble method can not only process scRNA-seq data with high-dimensional features, but also evaluate the importance of each gene in gene expression data through information gain. Our main goal is to identify marker genes from scRNA-seq data that can not only judge cell types but also have biological significance. After using RFCell for gene selection on 10 scRNA-seq data sets, we found that the accuracy of the average results is higher than that of using conventional gene selection strategies.

MATERIALS AND METHODS

The pipeline of our proposed RFCell is depicted in **Figure 1**. In the following section, we describe this pipeline in detail.

Method

Pouyan and Kostka (2018) proposed RAFSIL, a random forest-based method that can learn the similarity between cells from scRNA-seq data. RAFSIL consists of two steps: feature construction based on scRNA-seq data and similarity learning. RAFSIL has strong adaptability and scalability, and the similarity can be used for typical exploratory scRNA-seq data research, such as dimensionality reduction, visualization and clustering. Considering that RAFSIL uses permutation to generate similarity, we propose to use permutation to generate negative samples. We develop RFCell, a supervised gene selection strategy based on permutation and random forest. RFCell evaluates the importance of each gene through random forest classification. RFCell works in two steps: generation of negative samples and evaluation of gene importance using Random Forest.

Generation of Negative Samples

It is well known that scRNA-seq data is complex and diverse, so it is particularly important for scRNA-seq data gene selection. First, to generate a random negative sample matrix of gene expression data, we input the gene expression matrix **X** (**X** consists of *m* rows and *n* columns) obtained after data preprocessing as a positive sample. After that, the gene in each column of the positive sample matrix **X** is randomly permutated to form a new gene expression matrix **Z** (**Z** consists of m rows and n columns). We define each row of cells in the new gene expression matrix **Z** as a negative sample.

Next, we create the vector **y**. First, we define the label of the positive sample matrix **X** as a vector *p*, and *p* are all 1, where the number of 1 is the number of rows (m) of the positive sample matrix **X**. Second, the label of the negative sample matrix **Z** is defined as a vector *q*, and *q* is all 0, where the number of 0 is the number of rows (m) of the negative sample matrix **Z**. Here, we convert the *p* vector and *q* vector into data frame format respectively. Third, the vector *y* (*y* consists of 2 × m rows and one column) is generated by vertically merging the vector *p* and the vector *q*.

Finally, the positive sample matrix **X** and the negative sample matrix **Z** obtained from the above are merged vertically to obtain

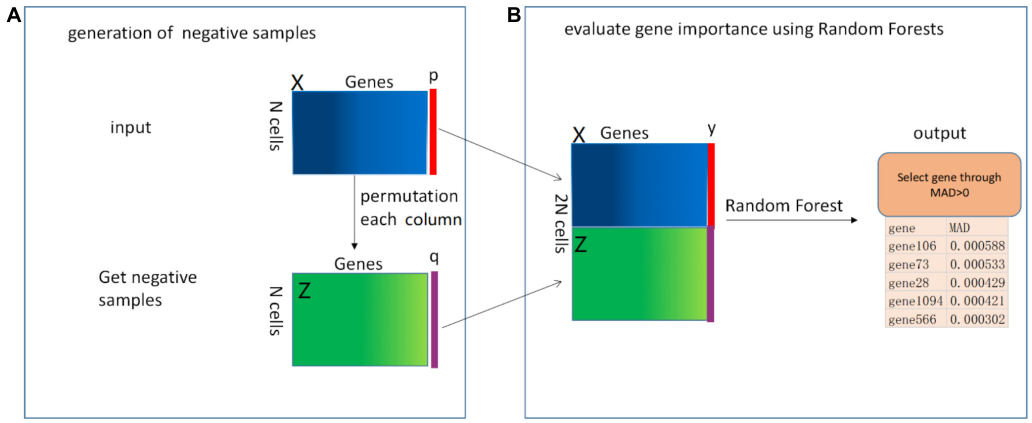


FIGURE 1 | The mechanism of RFCell (scRNA-seq gene selection based on permutation and Random Forest) algorithm. The input is a gene expression matrix. The RFCell algorithm includes two steps: **(A)** Generation of negative samples; **(B)** Evaluation of gene importance using Random Forest.

a new gene expression matrix **N** (**N** contains 2 × m rows and *n* columns).

Evaluation of Gene Importance Using Random Forest

We use the randomforest (Xin-Hai, 2013) package in R language to evaluate gene importance. First, in order to generate the random forest training data set, we horizontally merge the matrix *N* and the vector *y*. Through merging, we get the random forest training data set matrix *M* (*M* contains 2 × m rows and *n*+1 columns). Then, we call the random forest R language package. According to the usage of the randomforest package in R language, we use the vector **y** obtained above as the formula setting of the randomforest package, and use the matrix *M* as data setting of randomforest package. The importance parameter is set to True, and the remaining parameters are default values.

After calling the randomforest package, we use the importance function to calculate the importance of each gene, and obtain the importance of each gene through the mean decrease accuracy (MDA). MDA represents the degree of reduction in the accuracy of random forest prediction after one gene is permutated. The larger the value, the greater the importance of the gene. In our study, genes with MDA>0 are selected as genes that can identify cell types.

ScRNA-Seq Datasets

We tested 10 published scRNA-seq datasets and obtained results using gene selection methods. All these data sets have been used for performance research by several latest algorithms. For each data set, we use the expression unit provided by the author.

Darmanis dataset (Darmanis et al., 2015): In order to capture the cellular complexity of adult and fetal human brains at the entire transcriptome level, the authors performed single-cell RNA sequencing on 466 cells. This data set consists of oligodendrocytes, astrocytes, microglia, neuronal cells, endothelial cells, neural progenitor cells, quiescent newborn neurons, and two types of cells containing more than one different cell type Cells with characteristic genes are composed together.

Deng dataset (Deng et al., 2014): The authors used the Smart-seq or Smart-seq2 platform to perform RNA-Seq sequencing on Mus musculus cells from zygotic to late blastocysts of a single cell from the adult liver. The cells in this data set are separated from mouse embryonic oocytes to blastocyst stage, including four 1- cells (zygotes), eight early 2- cells, 12 metaphase 2- cells, 10 late 2- cells, and 14 4- cells, 28 8- cells, 50 16- cells, 43 early blast cells, 60 mid blast cells, and 30 late blast cells.

Engel dataset (Engel et al., 2016): The authors analyzed purified populations of thymic natural killer T cells (NKT cells) at the transcriptome level and epigenome level, as well as by single-cell RNA sequencing. The data consists of NKT1 cells, NKT2 cells, and NKT17 cells.

Grover dataset (Grover et al., 2016): Using single-cell RNA-seq technology, the authors systematically compared single hematopoietic stem cells (HSC) from young mice and old mice that were transgenic from Vwf-EGFP bacterial artificial chromosomes (BAC). By analyzing HSC transcriptome and HSC function at the single cell level, the authors found that molecular platelet priming and increased functional platelet bias are the main age-dependent changes in HSCs.

Pollen dataset (Pollen et al., 2014): Using microfluidic technology, the authors captured 301 single cells from 11 populations and analyzed the single-cell transcriptome within the down-sampling sequencing depth range. They proved that for unbiased cell type classification and biomarker identification, shallow scRNA-seq is indeed sufficient.

Sasagawa dataset (Sasagawa et al., 2013): The authors proposed a novel scRNA-seq method named Quartz-Seq. They applied this method to ES cells in different three cell-cycle phases (G1, S, and G2/M).

Ting dataset (Ting et al., 2014): The authors applied a microfluidic device to isolate Circulating tumor cells (CTCs) based on the model from a pancreatic cancer mouse to determine the heterogeneity of pancreatic CTCs. Then these CTCs were sequenced and compared to matched primary tumors, cell line controls.

Trapnell dataset (Trapnell et al., 2010): The author sequenced and analyzed more than 430 million paired 75 bp RNA-Seq reads from mouse myoblast cell lines on differentiation time series.

Treutlein dataset (Treutlein et al., 2014): The authors analyzed 198 single-cell transcriptomes from mouse lung epithelium in total. For time point E18.5, three individual experiments were performed using three different pregnant mices (3 biological replicates): 20 single cell transcriptomes yielded from pooled sibling lungs, 34 single cell transcriptomes yielded from one single embryonic lung, 26 single cell transcriptomes yielded from pooled sibling lungs. The authors used an unbiased genome-wide approach and classified these 80 cells into five populations: Clara (Scgblal), ciliate (Foxjl), AT1 (Pdpn, Ager), AT2 (Sftpc, Sftpb), and alveolar bipotential progenitor (BP) cells.

Zhou dataset (Zhou et al., 2016): The author used effective surface markers to capture the newborn pre-HSC with high purity, and then applied single-cell RNA sequencing to analyze endothelial cells, CD45- and CD45+ pre-HSC in the aorta-gonad-mesonephrine region, and fetus HSC of the liver.

The summary description of the scRNA-seq datasets we used is shown in **Table 1**.

Performance Evaluation

In order to compare the clustering results of RFCell and other gene selection methods, we used two commonly used clustering algorithm evaluation indicators: normalized mutual information (NMI) (Kiselev et al., 2017) and adjusted rand index (ARI) (Rand, 1971).

Mutual information (MI) measures the correlation between two sets of events. In information theory, a useful measure of information can be seen as the amount of information contained in a random variable about another random variable, or the uncertainty reduced by knowing another random variable. Formally, the MI of two discrete random variables X and Y can be defined as:

$$I(X:Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \qquad (1)$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y. NMI is to place MI between [0, 1] through information entropy, and its purpose is to evaluate the quality of the algorithm. For a random variable X, its information entropy can be calculated as:

$$H(X)=\sum_{i=1}^{n}p(x_i)I(x_i)=\sum_{i=1}^{n}p(x_i)\log\frac{1}{p(x_i)} \qquad (2)$$

The value of the random variable X = $\{x_1\ ,x_2,\ldots x_n\}$ and $p(x_i)$ represent the probability of event occurring, on the other hand, the value of random variable Y = $\{y_1,y_2\ ,\ldots y_n\}$ and $p(y_i)$ represents the probability of event occurring. NMI can be defined as:

$$U(X,Y)=2\frac{I(X;Y)}{H(X)+H(Y)} \qquad (3)$$

NMI is used to evaluate the consistency between the clustering results obtained and the true cell markers.

Rand Index (RI) is a measure of the similarity between clustering results and real categories. Mathematically, the RI is associated with accuracy. Given a set of S with n elements, then compare the two partitions M, N of S. The RI is calculated as follows:

$$RI=\frac{a+b}{a+b+c+d}=\frac{a+b}{C_n^2}=\frac{a+b}{n(n-1)/2} \qquad (4)$$

where a is the number of pairs of elements in S that are in the same subset in M and in the same subset in N; b is the number of pairs of elements in S that are in different subsets in M and in different subsets in N; c is the number of pairs of elements in S that are in the same subset in M and in different subsets in N; d is the number of pairs of elements in S that are in different subsets in M and in the same subset in *N*.

The RI is between [0, 1]. The greater the RI value, the more consistent the clustering result of the algorithm is with the known label, the higher the accuracy of the clustering effect, and the higher the purity in each category. The problem with the RI is that, when comparing multiple clustering results, RI values are usually high, resulting in a poor evaluation of the superiority

TABLE 1 | Summary description of the ten scRNA-seq datasets.

Datasets	#Samples	#Genes	#Classes	Unit
Darmanis (Darmanis et al., 2015)	466	22,088	9	CPM
Deng (Deng et al., 2014)	259	22,958	10	RPKM
Engel (Engel et al., 2016)	203	23,342	4	RPKM
Grover (Grover et al., 2016)	135	15,181	2	CPM
Pollen (Pollen et al., 2014)	249	14,805	11	TPM
Sasagawa (Sasagawa et al., 2013)	23	36,807	3	FPKM
Ting (Ting et al., 2014)	149	29,018	7	CPM
Trapnell (Trapnell et al., 2010)	372	47,192	4	FPKM
Treutlein (Treutlein et al., 2014)	80	23,271	5	FPKM
Zhou (Zhou et al., 2016)	181	23,624	8	FPKM

TABLE 2 | Comparison of SIMLR performance of gene sets obtained by four gene selection methods in terms of NMI.

DataSet	NMI			
	Expr	M3Drop	SCMarker	RFCell
Darmanis	0.720	0.687	**0.727**	0.724
Deng	0.676	**0.682**	0.650	**0.682**
Engel	0.528	0.609	**0.768**	0.670
Grover	0.004	0.043	0.002	**0.084**
Pollen	0.868	**0.944**	0.908	0.938
Sasagawa	0.592	**0.621**	NA	0.595
Ting	0.781	0.706	0.767	**0.829**
Trapnell	0.102	0.127	0.066	**0.222**
Treutlein	0.425	0.411	0.433	**0.531**
Zhou	0.631	0.619	0.590	**0.663**

NA:The number of genes selected by SCMarker is 0, so no results are obtained. The bold values mean the highest or equally-highest value among different methods.

TABLE 3 | Comparison of SIMLR performance of gene sets obtained by four gene selection methods in terms of ARI.

DataSet	ARI			
	Expr	**M3Drop**	**SCMarker**	**RFCell**
armanis	**0.549**	0.537	0.530	0.537
Deng	0.343	**0.412**	0.367	**0.412**
Engel	0.390	0.509	**0.710**	0.622
Grover	0.007	0.044	0.001	**0.109**
Pollen	0.798	**0.937**	0.832	0.917
Sasagawa	**0.561**	0.516	NA	0.555
Ting	0.540	0.532	0.491	**0.668**
Trapnell	0.010	0.062	0.010	**0.168**
Treutlein	0.237	0.239	0.285	**0.349**
Zhou	0.415	0.410	0.363	**0.483**

NA:The number of genes selected by SCMarker is 0, so no results are obtained. The bold values mean the highest or equally-highest value among different methods.

of the clustering algorithm. Therefore, ARI presented has better differentiation degree than RI. The range of ARI is (−1, 1). ARI can be defined as:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \tag{5}$$

where $E(RI)$ and $\max(RI)$ can be defined as:

$$E(RI) = E(\sum_{i,j}\binom{n_{i,j}}{2}) = [\sum_{i}\binom{n_i}{2}\sum_{j}\binom{n_j}{2}]/\binom{n}{2} \tag{6}$$

$$\max(RI) = \frac{1}{2}[\sum_{i}\binom{n_i}{2} + \sum_{j}\binom{n_j}{2}] \tag{7}$$

where $n_{i,j}$ are values from the contingency table, n_i is the sum of the i-th row of the contingency table, n_j is the sum of the j-th column of the contingency table.

Adjusted rand index is commonly used to assess the consistency between predicted clusters and true categories.

RESULTS

Comparison of RFCell With Benchmark Gene Selection Methods

To show the performance of RFCell over other gene selection methods, we used three classical clustering algorithms: Clustering method for single-cell interpretation through multikernel learning (SIMLR) (Wang et al., 2017), Single-cell consensus clustering (Wilkerson and Hayes, 2010) merges clustering results of multiple cells by consensus method (SC3) (Kiselev et al., 2017) and *k*-means (Kim et al., 2019). SIMLR is a software that learns the similarity measure between cells from the input single cell data, for SIMLR, we use the SIMLR package and igraph package in R language and apply their default parameters to get a good clustering effect. SC3 is a user-friendly tool for unsupervised clustering, which methods include gene filtering, similarity calculation, Transformations, k-means, consensus clustering, and finally hierarchical clustering of the results obtained by consensus clustering. We usually use SC3, SingleCellExperiment and scater package in R language to perform SC3 clustering. For hierarchical clustering, we use the hclust (Xu et al., 2019) function with default parameters in R to perform hierarchical clustering analysis on the similarity matrix of gene expression data to obtain the final clustering results. The parameter k of three methods was set to the true number of clusters. In addition to these three algorithms, gene selection based on scRNA-seq data can apply the RFCell

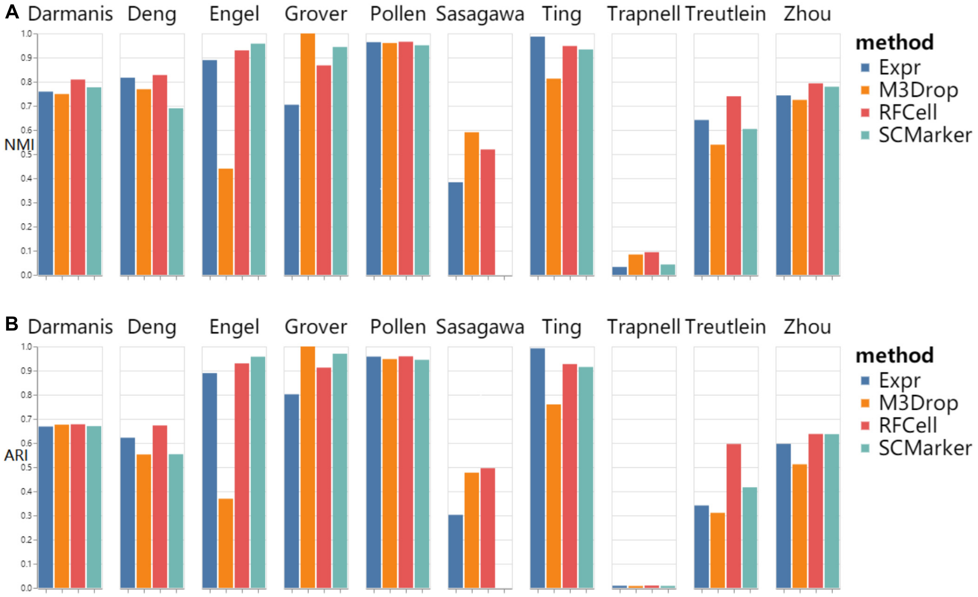


FIGURE 2 | SC3 clustering results based on RFCell and three gene selection methods including Expr, M3Drop, and SCMarker in terms of NMI **(A)** and ARI **(B)**.

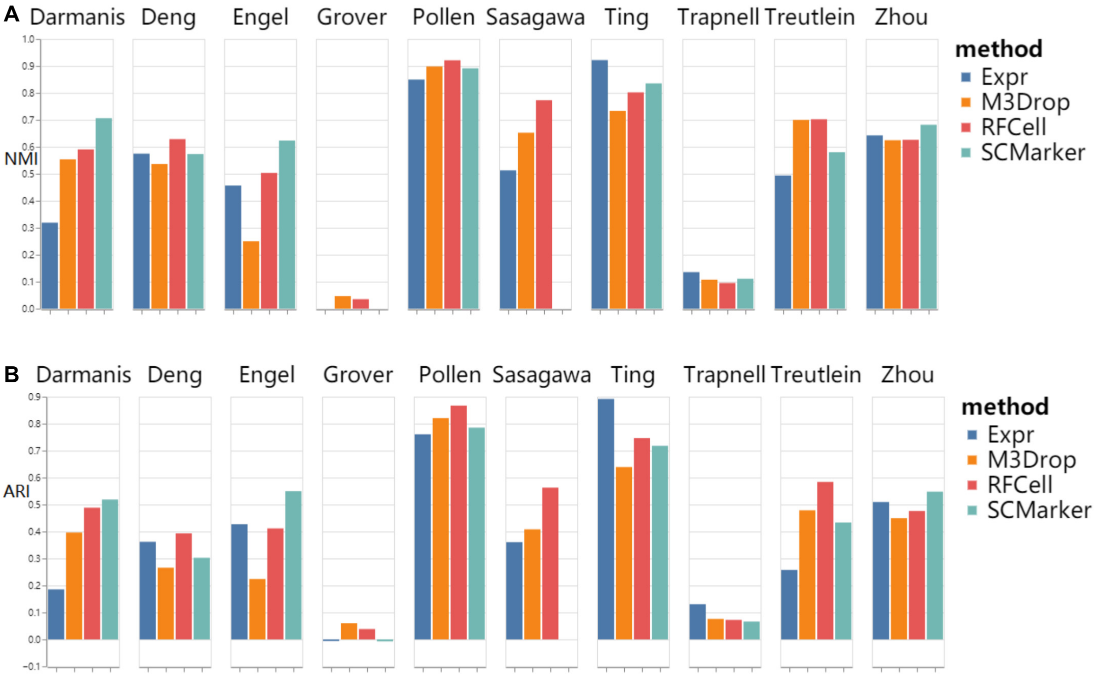


FIGURE 3 | k-means clustering results based on RFCell and three gene selection methods including Expr, M3Drop, and SCMarker in terms of NMI **(A)** and ARI **(B)**.

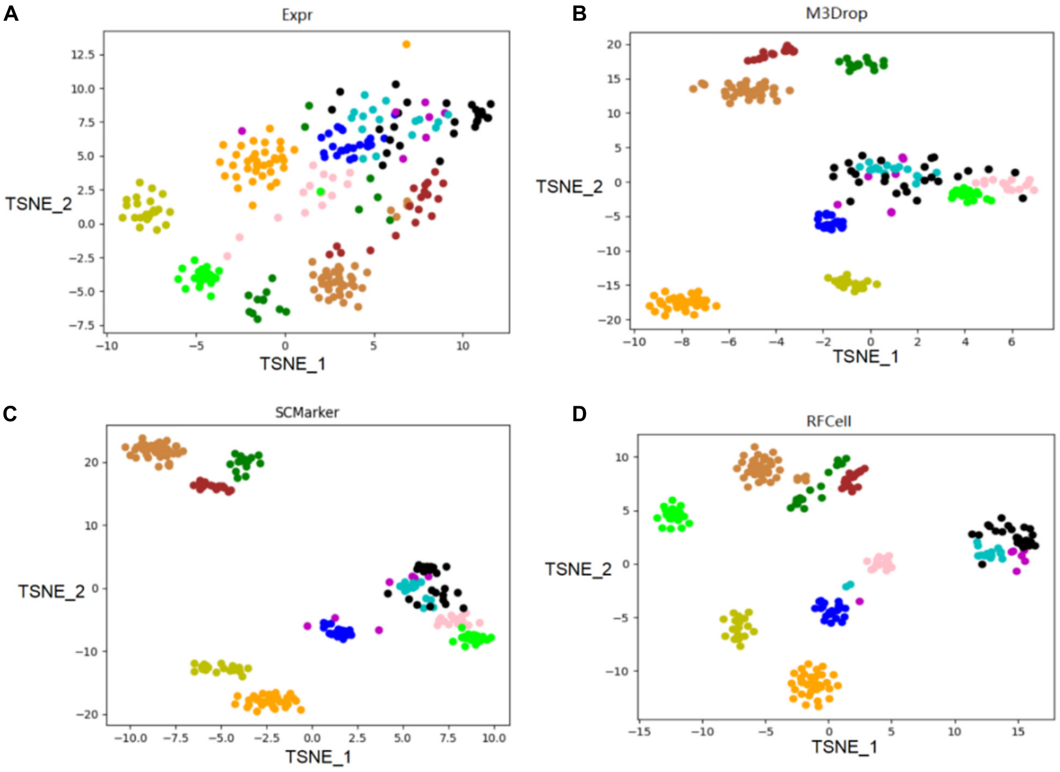


FIGURE 4 | On the published pollen data of the scRNA-seq data set, the gene sets obtained by the three gene selection methods (Expr, M3Drop, and SCMarker) and the gene sets obtained by the RFCell gene selection method were compared. The visualization diagrams respectively show the gene sets obtained by the four gene selection methods: **(A)** Visualization of the results of Expr gene selection; **(B)** Visualization of the results of M3Drop gene selection; **(C)** Visualization of the results of SCMarker gene selection; **(D)** Visualization of the results of RFCell gene selection.

feature selection results to any clustering method. In fact, the final gene selected by RFCell can be used not only for any clustering algorithms, but also for similarity calculation and building a cell network. The three feature selection methods specifically for scRNA-seq data are: Andrews and Hemberg (2019) proposed M3Drop, Wang et al. (2019) proposed SCMarker. The last method of selecting genes is to select the gene with the highest average expression value (Expr). For each scRNA-seq data, we first run RFCell 10 times, and then calculate the average of the NMI and ARI as the final result.

Based on SIMLR, **Table 2** clearly shows that, compared with other gene selection methods, RFCell can achieve better gene selection performance in more data in terms of NMI. For example, the average NMI of the data set clustering after RFCell gene selection is 0.593, the average NMI of the data set clustering after the Expr gene selection is 0.532, the average NMI of the data set clustering after the M3Drop gene selection is 0.544, and the average NMI of the data set clustering after SCMarker gene selection is 0.545. In more than half of all data sets, RFCell gene selection results are the best. **Table 3** also shows that, compared to other feature selection methods, in terms of ARI, RFCell achieve better gene selection performance in more datasets. For example, the average ARI of the data set clustering after RFCell gene selection is 0.482, the average ARI of the data set clustering after the Expr gene selection is 0.385, the average ARI of the data set clustering after the M3Drop gene selection is 0.419, and the average ARI of the data set clustering after SCMarker gene selection is 0.398. Considering both NMI and ARI, our method does perform better than other methods on a few datasets such as the Darmanis and Engel datasets, possibly because the characteristics of the genes that can distinguish cell types for these datasets could not be captured by RFCell.

As shown in **Figure 2**, we found that RFCell basically showed good results in SC3 clustering. The picture shows that compared with other gene selection methods, the scRNA-seq data set obtained by our proposed RFCell recognizes cell types more clearly. For Darmanis dataset, Deng dataset, pollen dataset, Trapnell dataset, Treutlein dataset and Zhou dataset, compared with other gene selection methods, the gene set obtained by

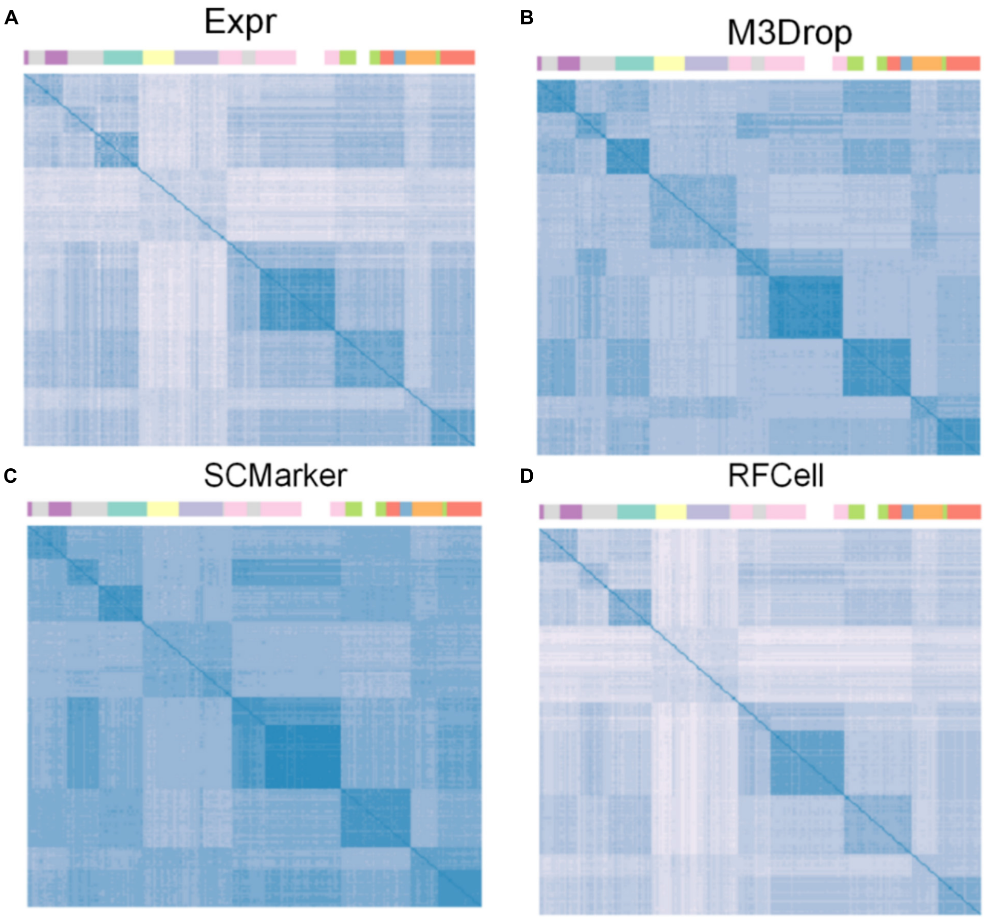


FIGURE 5 | The heat map of the result is derived from the spearman similarity measure of the gene set obtained after the gene selection of pollen data by four gene selection methods. The cells in the matrix are sorted by their true labels so that cells of the same type are adjacent. Cell clusters are clearly indicated by colored bars. **(A)** Heat map of the gene set obtained by the Expr gene selection; **(B)** Heat map of the gene set obtained by the M3Drop gene selection; **(C)** Heat map of the gene set obtained by the SCMarker gene selection; **(D)** Heat map of the gene set obtained by the RFCell gene selection.

RFCell has obvious advantages in distinguishing cell types. Both NMI and ARI have achieved the best gene selection performance, which shows that the gene set obtained with RFCell has biological significance. For Engle dataset, Grover dataset, Sasagawa dataset and Ting dataset, we found that through different gene selection methods to obtain different gene sets have their own advantages and disadvantages in distinguishing cell types. These results indicate that scRNA-Seq data is complex and diverse, and the gene set related to cell type recognition may have some unknown factors, which require further research.

As shown in **Figure 3**, we found that RFCell basically showed good results in k-means. The picture shows that compared with other gene selection methods, the scRNA-seq data set obtained by our proposed RFCell can significantly improve the clustering accuracy. For Deng dataset, pollen dataset, Sasagawa dataset and Treutlein dataset, compared with other gene selection methods, our proposed RFCell achieves satisfactory clustering performance, and more importantly, it can also provide potential biological explanations for clustering. This also shows that RFCell can identify the gene sets that contribute the most to the clusters.

Application of RFCell to Single Cell RNA-seq Data

We use the single-cell transcriptome data of 249 cells captured in 11 populations obtained using microfluidic technology as our original data, and visualize the different gene sets corresponding to the original data. Data visualization results show that RFCell separates cells more clearly. It is better than the results obtained by Expr, M3Drop and SCMarker (**Figures 4**, **5**).

As shown in **Figure 4**, the visualization results of the gene set selected by the Expr method show that only five cell types can be clearly distinguished, and the other cell types are scattered in confusion. The visualization results of the gene set selected by the M3Drop method also show that although there are eight cell types that can be effectively identified, the other three cell types (cell type 4, cell type 5, and cell type 6) are scattered and difficult to identify. The visualization results of the gene set selected by the SCMarker method are also difficult to effectively distinguish cell types. On the one hand, cell type 4 and cell type 5 are too widely dispersed; on the other hand, there is multiple cell types (cell type 3, cell type 4, cell type 5, and cell type 6) has a crossover, which makes the identification of cell type confused. The result of the visualization of the gene set obtained after gene selection by our proposed RFCell shows that all cell types can be clearly identified, and there is no crossover between cell types. This also shows that RFCell has superiority in cell type recognition. The heat map in **Figure 5** is derived from the spearman similarity measure of the gene set obtained after gene selection of pollen data by four gene selection methods. RFCell also showed better performance.

DISCUSSION AND CONCLUSION

In recent years, scRNA-seq technology has become a powerful tool for studying cell heterogeneity in tissues, advances in sequencing technology have enabled scientists to perform large-scale transcriptome profiling at single cell resolution in a high-throughput manner, clustering algorithms have passed unsupervised learning has become the main way to identify and characterize new cell types and gene expression patterns, however, on the one hand, differences in scRNA-seq technology can cause noise in scRNA-seq data, especially because it is impossible to repeat measurements on the same cell (Severson et al., 2018; Zhang et al., 2020). On the other hand, scRNA-seq data is noisier and more complex than traditional RNA-Seq data, and the high variability of the data also brings challenges to scRNA-seq data analysis (Chen et al., 2019). In order to analyze scRNA-seq data, feature selection methods can greatly reduce the dimensionality of the data and improve the results of cell type recognition. For analyzing specific data, especially gene expression data, many studies have shown that certain gene sets with correlation and functional synergy play an important role in analyzing scRNA-seq data and identifying specific cell types (Eisen, 1998; Young et al., 2010; Buettner et al., 2017).

In this study, we proposed a new feature selection method, RFCell, for gene selection of scRNA-seq data. Through feature selection based on permutation and random forest for each gene expression data. RFCell uses classic machine learning methods to perform supervised classification of scRNA-seq data to show its superiority compared with other feature selection methods. RFCell is characterized by a series of noteworthy functions. First, the negative samples are obtained by using scRNA-seq data permutation. Secondly, RFCell obtains the training data of the random forest by combining the original scRNA-seq and negative samples. Third, considering that the information contained in each genome and the ability to recognize cell types is different, we estimate the importance of each genome by calculating the importance function. Finally, RFCell selects genes with MDA>0 as the gene set that can identify cell types. This is done to make the results of RFCell robust to gene set mutations.

RFCell does have some limitations. First of all, the negative samples obtained from the original gene expression data using permutation are uncertain, so this means that for each data set, there may be some genes that can identify cell types are disrupted to the wrong cells. Therefore, in this process, some genes that are essential for classification are likely to be discarded, resulting in failure to obtain the best classification results. With this in mind, we have conducted many experiments to make RFCell stable to the results of gene selection. Experiments include visual analysis of gene sets obtained through different gene selection methods. The details are as follows. We use the single-cell transcriptome data of 249 cells captured in 11 populations obtained using microfluidic technology as our original data, use four gene selection methods to select the gene sets of the original data to obtain different gene sets, and visualize these sets of genes. In addition, we also do heat map analysis on gene sets. Corresponding experimental results show that RFCell shows superiority in the visualization map, but RFCell needs to be improved in the heat map analysis.

It is expected that biological information (such as labeled gene sets) will be used in the future to select genes related to cell types in scRNA-seq for further study. Incorporating information from different views may be helpful in improving gene selection (Liu et al., 2020a; Liu et al., 2020b; Lan et al., 2020). There are some differences among the results for scRNA-seq data based on different gene selection methods. Analyzing the preference performance of different gene selection methods for scRNA-seq data could improve the accuracy of cell type identification. Therefore, we believe that integrating different gene selection methods may benefit gene selection.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study and the references for the data are provided in this article.

AUTHOR CONTRIBUTIONS

H-DL conceived the study. YZ and Z-YF performed the experiments and wrote manuscripts. C-XL, CD, Y-PX, and H-DL wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work is supported by National Key R&D Program of China (No. 2018YFC0910504), National Natural Science Foundation of China (Nos. U1909208, 61972423, and 61772557), 111 Project (No. B18059), and Hunan Provincial Science and Technology Program (2018WK4001).

REFERENCES

Aevermann, B. D., Novotny, M., Bakken, T., Miller, J. A., Diehl, A. D., Osumi-Sutherland, D., et al. (2018). Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Hum. Mol. Genet.* 27, R40–R47. doi: 10.1093/hmg/ddy100

Andrews, T. S., and Hemberg, M. (2019). M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics* 35, 2865–2867. doi: 10.1093/bioinformatics/bty1044

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). F-sclvm: scalable and versatile factor analysis for single-cell Rna-seq. *Genome Biol.* 18:212. doi: 10.1186/s13059-017-1334-8

Chen, G., Ning, B., and Shi, T. (2019). Single-cell Rna-seq technologies and related computational data analysis. *Front. Genet.* 10:317. doi: 10.3389/fgene.2019.00317

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* 112, 7285–7290. doi: 10.1073/pnas.1507125112

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell Rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196. doi: 10.1126/science.1245316

Eisen, M. B. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868.

Engel, I. G., Seumois, L., Chavez, D., Samaniego-Castruita, B., White, A., Chawla, et al. (2016). Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nat. Immunol.* 17, 728–739. doi: 10.1038/ni.3437

Goolam, M., Scialdone, A., Graham, S. J. L., Macaulay, I. C., Jedrusik, A., Hupalowska, A., et al. (2016). Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* 165, 61–74. doi: 10.1016/j.cell.2016.01.047

Grover, A., Sanjuan-Pla, A., Thongjuea, S., Carrelha, J., Giustacchini, A., Gambardella, A., et al. (2016). Single-cell Rna sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nat. Commun.* 7:11075. doi: 10.1038/ncomms11075

Hedlund, E., and Deng, Q. (2018). Single-cell Rna sequencing: technical advancements and biological applications. *Mol. Aspects Med.* 59, 36–46. doi: 10.1016/j.mam.2017.07.003

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell Rna sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50, 1–14.

Kim, T. I, Chen, R., Lin, Y., Wang, A. Y. Y., Yang, J. Y. H., and Yang, P. (2019). Impact of similarity metrics on single-cell Rna-seq data clustering. *Brief. Bioinform.* 20, 2316–2326.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). Sc3: consensus clustering of single-cell Rna-seq data. *Nat. Methods* 14, 483–486.

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). Ldicdl: Lncrna-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.3034910

Lever, J., Krzywinski, M., and Altman, N. (2017). Principal component analysis. *Nat. Methods* 14, 641–642. doi: 10.1038/nmeth.4346

Linderman, G. C., and Steinerberger, S. (2019). Clustering with T-Sne, provably. *SIAM J. Math. Data Sci.* 1, 313–332. doi: 10.1137/18m1216134

Liu, J., Zeng, D., Guo, R., Lu, M., Wu, F.-X., and Wang, J. (2020a). Mmhge: Detecting mild cognitive impairment based on multi-atlas multi-view hybrid graph convolutional networks and ensemble learning. *Cluster Comput.* 24, 103–113. doi: 10.1007/s10586-020-03199-8

Liu, J., Pan, Y., Wu, F.-X., and Wang J. (2020b). Enhancing the feature representation of multi-modal MRI data by combining multi-view information for MCI Classification. *Neurocomputing* 400, 322–332. doi: 10.1016/j.neucom.2020.03.006

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell Rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401.

Poirion, O. B., Zhu, X., Ching, T., and Garmire, L. (2016). Single-cell transcriptomics bioinformatics and computational challenges. *Front. Genet.* 7:163. doi: 10.3389/fgene.2016.00163

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi: 10.1038/nbt.2967

Pouyan, M. B., and Kostka, D. (2018). Random forest based similarity learning for single cell Rna sequencing data. *Bioinformatics* 34, i79–i88. doi: 10.1093/bioinformatics/bty260

Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length Mrna-seq from single-cell levels of Rna and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi: 10.1038/nbt.2282

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* 66, 846–850. doi: 10.1080/01621459.1971.10482356

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., et al. (2013). Quartz-seq a highly reproducible and sensitive single-cell Rna sequencing method, reveals nongenetic gene-expression heterogeneity. *Genome Biol.* 4:17.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Severson, D. T., Owen, R. P., White, M. J., Lu, X., and Schuster-Böckler, B. (2018). Bearscc determines robustness of single-cell clusters using simulated technical replicates. *Nat. Commun.* 9:1187. doi: 10.1038/s41467-018-03608-y

Shao, C., and Hofer, T. (2017). Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* 33, 235–242. doi: 10.1093/bioinformatics/btw607

Ting, D. T., Wittner, B. S., Ligorio, M., Vincent Jordan, N., Shah, A. M., Miyamoto, D. T., et al. (2014). Single-cell Rna sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* 8, 1905–1918. doi: 10.1016/j.celrep.2014.08.029

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.* 25, 1491–1498. doi: 10.1101/gr.190595.115

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by Rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell Rna-seq. *Nature* 509, 371–375. doi: 10.1038/nature13173

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell Rna-seq data by kernel-based similarity learning. *Nat. Methods* 14:414.

Wang, F., Liang, S., Kumar, T., Navin, N., and Chen, K. (2019). Scmarker: Ab initio marker selection for single cell transcriptome profiling. *PLoS Comput. Biol.* 15:e1007445. doi: 10.1371/journal.pcbi.1007445

Wilkerson, M. D., and Hayes, D. N. (2010). Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572–1573. doi: 10.1093/bioinformatics/btq170

Xin-Hai, L. I. (2013). Using "random forest"for classification and regression. *Chin. J. Appl. Entomol.* 50, 1190–1197.

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980.

Xu, Y., Li, H., Pan, Y., Luo, F., and Wang, J. (2019). A gene rank based approach for single cell similarity assessment and clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2931582 [Epub ahead of print].

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for Rna-seq: accounting for selection bias. *Genome Biol.* 11:R14. doi: 10.1186/gb-2010-11-2-r14

Zhang, S., Li, X., Lin, Q., and Wong, K. C. (2020). *Review of Single-Cell Rna-Seq Data Clustering for Cell Type Identification and Characterization*.

Zheng, R., Li, M., Chen, X., Wu, F. X., Pan, Y., and Wang, J. (2019). Bixgboost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. *Bioinformatics* 35, 1893–1900. doi: 10.1093/bioinformatics/bty908

Zhou, F., Li, X., Wang, W., Zhu, P., Zhou, J., He, W., et al. (2016). Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* 533, 487–492. doi: 10.1038/nature17997

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 16 August 2021
doi: 10.3389/fgene.2021.695662

Integrative Transcriptomic, Lipidomic, and Metabolomic Analysis Reveals Potential Biomarkers of Basal and Luminal Muscle Invasive Bladder Cancer Subtypes

Chao Feng[2,4,5†], *Lixin Pan*[2,4,5†], *Shaomei Tang*[4,7†], *Liangyu He*[1,2,3,4], *Xi Wang*[2,4], *Yuting Tao*[2,4,5], *Yuanliang Xie*[2,4,8], *Zhiyong Lai*[2,4], *Zhong Tang*[6*], *Qiuyan Wang*[2,4*] *and Tianyu Li*[1,2,3,4*]

[1] Institute of Urology and Nephrology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [2] Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, Nanning, China, [3] Department of Urology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [4] Guangxi Key Laboratory for Genomic and Personalized Medicine, Guangxi Collaborative Innovation Center for Genomic and Personalized Medicine, Nanning, China, [5] Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Guangxi Medical University, Nanning, China, [6] School of Information and Management, Guangxi Medical University, Nanning, China, [7] Department of Gastroenterology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [8] Department of Urology, Affiliated Tumor Hospital of Guangxi Medical University, Nanning, China

OPEN ACCESS

Edited by:
Chen Qingfeng,
Guangxi University, China

Reviewed by:
Zeeshan Ahmed,
Rutgers, The State University of New Jersey, United States
Debo Cheng,
University of South Australia, Australia

***Correspondence:**
Qiuyan Wang
qiuyanwang510@yahoo.com
Tianyu Li
litianyu@gxmu.edu.cn
Zhong Tang
18978868502@qq.com

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 April 2021*
Accepted: *29 June 2021*
Published: *16 August 2021*

Citation:
Feng C, Pan L, Tang S, He L, Wang X, Tao Y, Xie Y, Lai Z, Tang Z, Wang Q and Li T (2021) Integrative Transcriptomic, Lipidomic, and Metabolomic Analysis Reveals Potential Biomarkers of Basal and Luminal Muscle Invasive Bladder Cancer Subtypes. Front. Genet. 12:695662. doi: 10.3389/fgene.2021.695662

Muscle invasive bladder cancer (MIBC) is a heterogeneous disease with a high recurrence rate and poor clinical outcomes. Molecular subtype provides a new framework for the study of MIBC heterogeneity. Clinically, MIBC can be classified as basal and luminal subtypes; they display different clinical and pathological characteristics, but the molecular mechanism is still unclear. Lipidomic and metabolomic molecules have recently been considered to play an important role in the genesis and development of tumors, especially as potential biomarkers. Their different expression profiles in basal and luminal subtypes provide clues for the molecular mechanism of basal and luminal subtypes and the discovery of new biomarkers. Herein, we stratified MIBC patients into basal and luminal subtypes using a MIBC classifier based on transcriptome expression profiles. We qualitatively and quantitatively analyzed the lipids and metabolites of basal and luminal MIBC subtypes and identified their differential lipid and metabolite profiles. Our results suggest that free fatty acids (FFAs) and sulfatides (SLs), which are closely associated with immune and stromal cell types, can contribute to the diagnosis of basal and luminal subtypes of MIBC. Moreover, we showed that glycerophosphocholine (GCP)/imidazoles and nucleosides/imidazoles ratios can accurately distinguish the basal and luminal tumors. Overall, by integrating transcriptomic, lipidomic, and metabolomic data, our study reveals specific biomarkers to differentially diagnose basal and luminal MIBC subtypes and may provide a basis for precision therapy of MIBC.

Keywords: MIBC, subtype, transcriptomic, lipidomics, metabolomic

INTRODUCTION

Bladder cancer (BC) is the 10th most common malignancy worldwide (Bray et al., 2018). BC can be classified into non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC) based on the depth of tumor cells invasion (Kamat et al., 2016). Approximately 25% of BC patients are diagnosed with MIBC, which has a higher rate of relapse and worse prognosis than NMBIC. Neoadjuvant cisplatin-based chemotherapy (NAC) before radical cystectomy is the standard treatment option for MIBC patients (Grossman et al., 2003; International Collaboration of Trialists et al., 2011). However, approximately 40% of MIBC patients benefit from NAC, and only a minority of patients with MIBC respond to immunotherapy (Zargar et al., 2015). Therefore, new MIBC diagnostic biomarkers and therapeutic strategies are urgently needed.

Accumulating evidence indicates that MIBC is a heterogeneous disease that can be divided into different molecular subtypes based on transcriptome profiles or specific genomic alterations (Sjodahl et al., 2012; Cancer Genome Atlas Research Network, 2014; Robertson et al., 2017; McConkey and Choi, 2018). MIBC can be grouped into basal and luminal subtypes with distinct classifiers or models, which are similar to the molecular subtypes used to stratify types of breast cancer (Damrauer et al., 2014; Sjodahl et al., 2017). Among these classifiers, a clinically significant panel of 47 genes (BASE47) is used as a classifier of high-grade MIBC. BASE47 accurately discriminates intrinsic MIBC subtypes and promotes an understanding of MIBC pathobiology (Damrauer et al., 2014). Typical urothelial basal cells markers, such as KRT6B, KRT14, and KRT5, are highly expressed in basal tumors, while luminal tumors express high levels of genes that mark terminal urothelial differentiation, such as those seen in umbrella cells (KRT20, UPK1B, UPK3A, and UPK2). MIBC subtypes not only demonstrate distinctive biological characteristics but also have prognostic and therapeutic value. Basal MIBC has a worse prognosis and a higher rate of metastasis than the luminal subtype (Choi et al., 2014a; Robertson et al., 2017). Moreover, basal MIBC subtype is more sensitive to anti-epidermal growth factor receptor (anti-EGFR) agents and cisplatin-based combination chemotherapy than the luminal subtype (Choi et al., 2014a,b). Given the complex heterogeneity of MIBC, there is an urgent need for the definition of subtype-specific biomarkers that can be applied for more precise management and therapeutic interventions for MIBC.

The reprogramming of metabolic patterns in tumor tissue facilitates the rapid proliferation of tumor cells in the absence of oxygen and nutrients and drives tumor progression (Putluri et al., 2011; Nuhn et al., 2012). The tumor metabolome originates from the interaction of genome, transcriptome, proteome, and a series of external influences. Metabolomic signatures mirror the dynamic biochemical activity of the tumor's pathobiology (Loras et al., 2018). Therefore, over the last decade, research had increasingly focused on the identification of novel biomarkers associated with metabolomics for the early detection of cancer (Alberice et al., 2013; Frantzi et al., 2016; Yumba Mpanga et al., 2018). Although previous research had concentrated on BC metabolism for screening and detection (Sahu et al., 2017), it has become evident that lipid metabolism is also an important component to be considered. Lipids are employed to store energy; they are also involved in cell membrane synthesis and act as messengers for molecular recognition and signal transduction (Larrouy-Maumus, 2019). Lipid metabolism is closely related to cancer progression (Munir et al., 2019). Thus, both lipidomics and metabolomics play vital roles in the occurrence and development of cancer. However, to date, differential lipid and metabolite profiles between basal and luminal MIBC subtypes have not been examined.

Herein, we integrated transcriptomic, lipidomic, and metabolomic analyses to identify the differential lipids and metabolites between basal and luminal MIBC subtypes, which will provide potential biomarkers for precision therapy of MIBC.

MATERIALS AND METHODS

Clinical Samples

The 12 MIBC tissues used in this study were obtained from The First Affiliated Hospital of Guangxi Medical University in China from June 2019 to June 2020. Patients undergoing chemotherapy or radiotherapy before surgical resection were excluded, and the diagnosis of MIBC was confirmed by two experienced pathologists.

RNA Sequencing

Total RNA was extracted from tissues using TRIzol® reagent (Invitrogen, Carlsbad, CA, United States) according to the manufacturer's protocol. Ribosomal RNA (rRNA) was removed from the total RNA using Ribo-Zero rRNA removal kits (Illumina, San Diego, CA, United States). Complementary DNA (cDNA) libraries were constructed by reverse transcription of the purified messenger RNAs (mRNAs). The libraries were amplified by PCR, followed by sequencing for 150 cycles on an Illumina HiSeq 4000 sequencer (Illumina). The quality of the raw sequencing data was assessed using FastQC software. Fastp was used to preprocess the raw data (Chen et al., 2018). The clean data were mapped to the human genome (hg19) using HISAT2 (Kim et al., 2019) and StringTie (Pertea et al., 2015, 2016), and Cufflinks was used to merge the data (Ghosh and Chan, 2016). The 47-gene panel was used to accurately separate MIBC samples into luminal and basal subtypes (Damrauer et al., 2014). Gene set enrichment analysis (GSEA) was conducted based on the default parameters, using mRNA expression profiles of samples. The xCell algorithm was used to specifically infer 64 immune and stromal cell types in each sample, based on mRNA expression profiles (Aran et al., 2017). The expression profiles of samples were prepared and uploaded to the xCell web[1]. Analysis was performed by xCell signature ($N = 64$) with 1,000 permutations, based on the parameter settings.

[1] http://xcell.ucsf.edu/

Tissue Metabolome Extraction

Extraction methods were performed as previously reported (Yuan et al., 2012). Tissues were ground using the Precellys evolution system (Bertin Technologies, Saint Quentin en Yvelines, French) under 1,600 × *g*, for 10 s, two cycles, and a 5-s pause. Samples were then incubated at 500 × *g* for 30 min at 4°C. The sample was centrifuged at 4,000 × *g* for 10 min at 4°C; then, the supernatant was isolated and dried by Genevac miVac (Tegent Scientific Ltd., Ipswich, United Kingdom). Precipitates were resuspended in 100 μl of 1% acetonitrile, and the supernatant was isolated for further analysis.

Metabonomics Data Acquisition

Ultrahigh performance liquid chromatography (UPLC, Agilent 1290 II, Agilent Technologies, Waldbronn, Germany) combined with tandem quadrupole time-of-flight (5600 Triple TOF Plus, AB Sciex, Singapore), and ACQUITY UPLC HSS T3 (1.8 μm, 2.1 mm × 100 mm, Waters, Dublin, Ireland) chromatographic column were used for the analysis. All analyses were performed in electrospray ionization mode. Instrument conditions were as previously reported (Song et al., 2020), including the following: curtain gas = 35; positive ion spray voltage = 5,500 V; negative ion spray voltage = -4,500 V; temperature = 450°C; ion source gas 1 = 50; and ion source gas 2 = 50. Data acquisition mode included a full scan of the primary mass spectrum and information-dependent acquisition of secondary mass spectrum data. MarkerView 1.3 (AB Sciex, Concord, ON, Canada) was used to extract the peak area, mass-to-charge ratio, and retention time of the primary mass spectrum data to generate a two-dimensional data array. Secondary mass spectrum data were extracted by PeakView 2.2 (AB Sciex), and metabolite IDs were identified after interrogation of a metabolite database, HMDB, and METLIN standards. Metabolite IDs were assigned to the corresponding ion of the two-dimensional data array.

Tissue Lipid Extraction

Lipid extraction was conducted according to a modified Bligh/Dyer extraction method (Song et al., 2020). Samples were redissolved in isotopic mixed standards and then analyzed via Exion UPLC-QTRAP 6500 Plus (Sciex) with the electrospray ionization mode under the following conditions: curtain gas = 20;

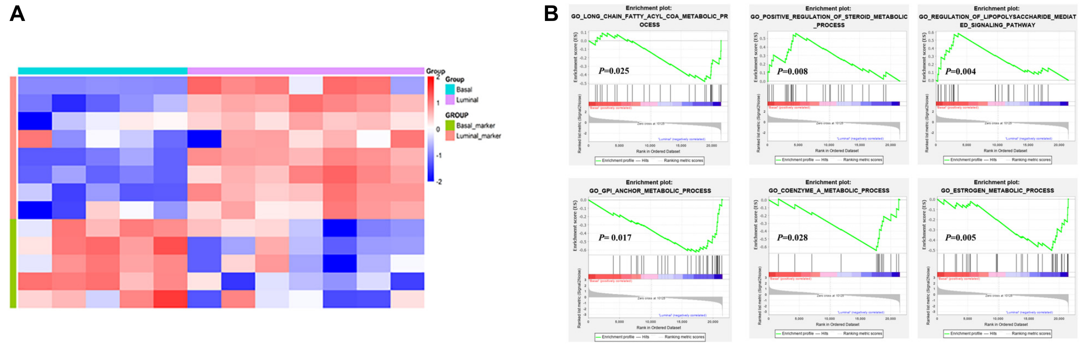


FIGURE 1 | Transcriptome analysis reveals changes in lipid and metabolic pathways. **(A)** Expression heatmap of specific MIBC basal and luminal markers. **(B)** GSEA analysis showed the activation pathways in basal and luminal MIBC subtypes.

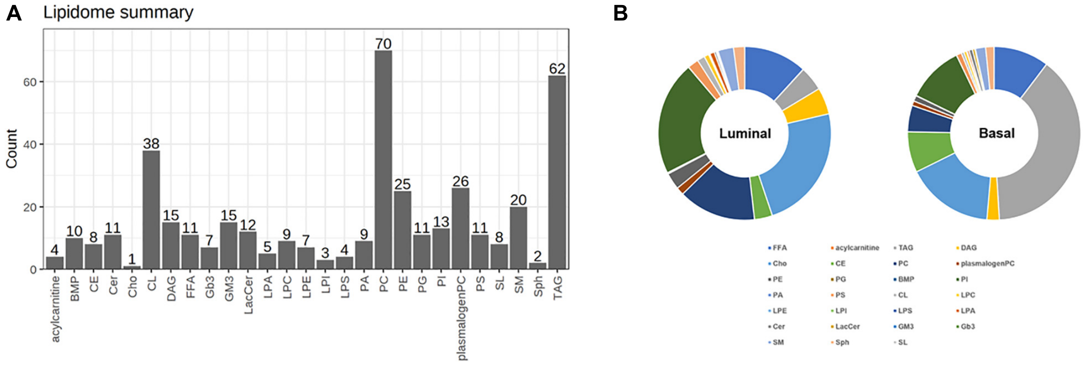


FIGURE 2 | Distinct lipid profiles in basal and luminal MIBC subtypes. **(A)** The lipid types and amounts tested. **(B)** The relative frequencies of lipids in basal and luminal MIBC subtypes. BMP, bis (monoglycerol) phosphate ester; CE, cholesteryl esters; Cer, ceramides; Cho, free cholesterols; CL, cardiolipins; DAG, diacylglycerols; FFA, free fatty acids; Gb3, Ceramide trihexoside; GM3, monosialogangliosides; LacCer, lactosylceramides; LPA, lyso-PA; LPC, lyso-PC; LPE, lyso-PE; LPI, lyso-PI; LPS, lyso-PS; PA, phosphatidic acids; PC, phosphatidylcholines; PE, phosphatidylethanolamines; PG, phosphatidylglycerols; PI, phosphatidylinositols; PS, phosphatidylserines; SL, sulfatides; SM, sphingomyelins; Sph, sphingosine; TAG, triacylglycerols.

ion spray voltage = 5,500 V; temperature = 400°C; ion source gas 1 = 35; and ion source gas 2 = 35.

Lipidomics Data Acquisition

Phenomenex Luna silica (3 μm, 1.5 mm × 200 mm) was selected as the chromatographic column. Lipids were extracted under A phase (chloroform/methanol/ammonia = 89.5:10:0.5) and B phase (chloroform/methanol/ammonia/water = 55:39:0.5:5.5). Extraction began with a 95% gradient of A phase from 0 to 5 min, then a linear decrease to 60% (in 7 min) for 4 min, a further decline to 30% for 15 min, and return to 95% for the last 5 min. Mass spectrometry multiple reaction monitoring was established for lipid identification and quantitative analysis (Lam et al., 2017, 2018).

Metabonomics and Lipidomics Data Analysis

Metabonomics and lipidomics data were prepared and uploaded to the MetaboAnalyst software 4.0[2] (Chong et al., 2019). Multivariate statistical analysis, cluster analysis, dimensionality reduction, and heatmaps were performed, based on the default parameters.

Statistical Analysis

Statistical analyses were performed using GraphPad Prism software (version 8.0, GraphPad, San Diego, CA, United States). Statistically significant differences between the two groups were evaluated by two-tailed Student's *t*-test. The relationships between lipid elements and cell types in the tumor microenvironment were analyzed by Pearson correlation analysis. A $p < 0.05$ was considered statistically significant. The area under the receiver operating characteristic (ROC) curve (AUC) was calculated to evaluate the accuracy of prediction.

RESULTS

Transcriptome Analysis Reveals Changes in Lipid and Metabolic Pathways

The establishment of tumor molecular subtypes has deepened our understanding of mutation gene profiles, tumor progression, and therapy responses (Robertson et al., 2018; Kamoun et al., 2020). Herein, we accurately classified 12 MIBC patients into basal and luminal subtypes using the BASE47 classifier based on transcriptome expression profiles (Damrauer et al., 2014). RNA-seq analysis revealed that basal and luminal MIBC subtype tumors displayed distinct gene expression patterns. Basal subtype had high levels of basal marker expression but low levels of luminal marker expression, while the luminal subtype displayed an opposite pattern (**Figure 1A**). GSEA analysis showed that activated long-chain fatty acyl-coA metabolic processes, positive regulation of steroid metabolic processes, and regulation of the lipopolysaccharide-mediated signaling pathway were associated with basal MIBC subtype, and glycosyl-phosphatidyl inositol (GPI) anchor metabolic process, coenzyme A metabolic process, and estrogen metabolic process were related to luminal

TABLE 1 | Differential lipids of basal and luminal subtype (basal vs. luminal).

Elevated lipids	Log2FC	*p*-value	Declined lipids	Log2FC	*p*-value
SL d18:1/24:1h	3.9567	0.014*	CL68:6(16:1)	−2.3181	0.018*
SL d18:1/22:0	3.9406	0.012*	CL68:5(16:1)	−2.1704	0.045*
SL d18:1/24:0h	3.107	0.011*	SM d(18:1/26:0)	−1.9878	0.020*
SL d18:1/22:1	2.8944	0.012*	PC34:2 (16:1/18:1)	−1.8854	0.011*
SL d18:1/22:0h	2.8846	0.003*	PC34:1 (16:1/18:0)	−1.5558	0.009*
LacCer d18:1/14:0	2.2354	0.042*	BMP36:2	−1.5464	0.012*
GM3 d18:1/22:1	1.8766	0.002*	PI 34:1	−1.4499	0.034*
SM d18:1/20:1	1.6788	0.033*	BMP36:1	−1.4439	0.004*
SM d18:1/18:1	1.537	0.018*	CL70:7(16:1)	−1.3534	0.004*
SM d18:1/22:1	1.5258	0.045*	BMP36:4	−1.3515	0.040*
SM d18:1/18:0	1.4422	0.002*	CL70:6(16:1)	−1.3464	0.007*
SL	1.3674	0.013*	BMP	−1.2687	0.008*
SM d18:1/20:0	1.1882	0.000*	BMP36:3	−1.2364	0.042*
DAG38:4(18:0/20:4)	1.1665	0.004*	PC34:3 (16:1/18:2)	−1.2302	0.012*
FFA16:0	0.99879	0.000*	PA32:1	−1.0937	0.032*
FFA18:0	0.98576	0.000*	CL70:6(18:2)	−1.0884	0.033*
FFA	0.91768	0.000*	PC34:3	−1.0569	0.010*
LPI20:4	0.86724	0.004*	PG38:6	−0.98387	0.043*
SM d18:1/22:0	0.82901	0.005*	PC36:2	−0.82249	0.045*
PI 38:4	0.77328	0.025*	BMP38:4	−0.80758	0.022*
FFA18:1	0.75857	0.047*	PE38:6	−0.76824	0.043*
TAG52:5(16:0)	3.0945	0.0527	PE40:6	−0.74456	0.033*
Cer d(18:1/20:0)	2.8726	0.065368	CL66:4(16:1)	−2.2132	0.081798
LacCer d18:1/18:0	2.6487	0.082395	PC32:2 (16:1/16:1)	−1.8787	0.088617
GM3 d18:1/1:80	1.9886	0.070227	GM3 d18:0/26:0	−1.8627	0.070982
Cer d(18:1/14:0)	1.6358	0.066787	BMP34:1	−1.7599	0.057363
Gb3 d18:1/18:0	1.4349	0.098536	PE32:1	−1.6519	0.086111
GM3 d18:1/22:0	1.3553	0.053135	BMP34:2	−1.5235	0.051358
LysoPC18:0	1.0642	0.055778	CL70:5(16:1)	−1.3076	0.070479
FFA22:4	1.0201	0.092333	PC32:2	−1.2708	0.079861
FFA22:5	0.92609	0.087524	SM d18:1/25:0	−1.2471	0.052022
PA(36:1)	0.78922	0.066101	PC32:1	−1.2189	0.091632
SM d18:1/24:1	0.78035	0.056402	GM3 d18:0/25:0	−1.1178	0.058641
FFA20:4	0.75017	0.090839	PC36:2 (18:1/18:1)	−1.0772	0.050403
			PC36:3 (18:1/18:2)	−0.86314	0.090513
			LysoPC16:1	−0.84867	0.094135
			PC32:1 (16:0/16:1)	−0.84791	0.069625
			PC36:3	−0.78299	0.050465
			PC40:7 (22:6/18:1)	−0.69562	0.062649

** indicates $p < 0.05$.*

[2] www.metaboanalyst.ca

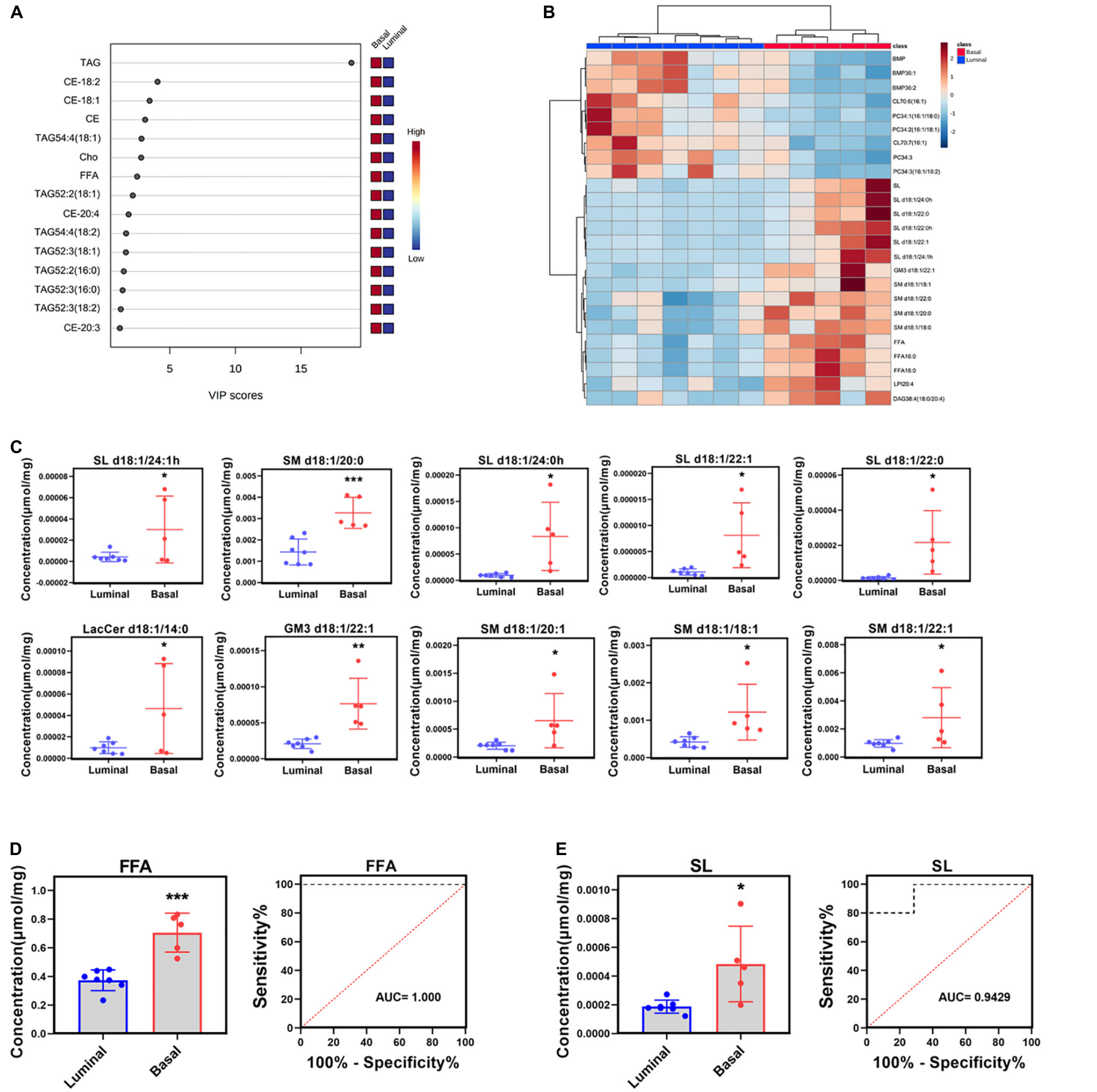


FIGURE 3 | Potential lipid biomarkers of basal and luminal MIBC subtypes. **(A)** VIP score of altered lipid elements. **(B)** Heatmap of the top 25 altered lipid elements in basal and luminal MIBC subtypes. **(C)** The levels of the top 10 significantly differential lipid constituents in basal and luminal MIBC subtypes. **(D,E)** FFA and SL levels and AUC values. * indicates $p < 0.05$; ** indicates $p < 0.01$; and *** indicates $p < 0.001$.

MIBC subtype (**Figure 1B**). These results indicated that the lipid and metabolic pathways of basal and luminal MIBC subtypes were different.

Distinct Lipid Profiles in Basal and Luminal MIBC Subtypes

To further explore the differential lipids between basal and luminal MIBC subtypes. A total of 417 lipid elements could be qualitatively and quantitatively detected (**Figure 2A**). The content of lipid elements was significantly different in basal and luminal MIBC subtypes (**Figure 2B**). The differential lipid elements are shown in **Table 1**.

Potential Lipid Biomarkers of Basal and Luminal MIBC Subtypes

Partial least squares discrimination analysis (PLS-DA) was performed to detect significant differential lipid elements between basal and luminal MIBC subtypes. By the variable import in project (VIP) score of each group, the top 15 lipid elements were identified (**Figure 3A**). The top 25

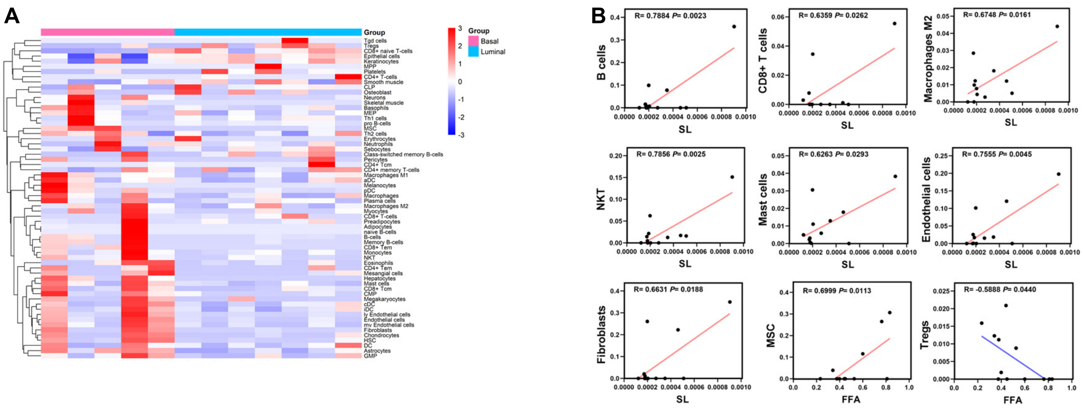


FIGURE 4 | Potential lipid biomarkers are associated with tumor microenvironment. **(A)** Heatmap of the relative frequency of immune cell and stromal cell types in basal and luminal MIBC samples as identified by the "xCell" algorithm. Red line represents the maximum expression level and blue line represents the minimum expression level. **(B)** Pearson correlation analysis revealed the relationship among FFA, SL, immune, and stromal cell types. Red line represents the maximum expression level and blue line represents the minimum expression level.

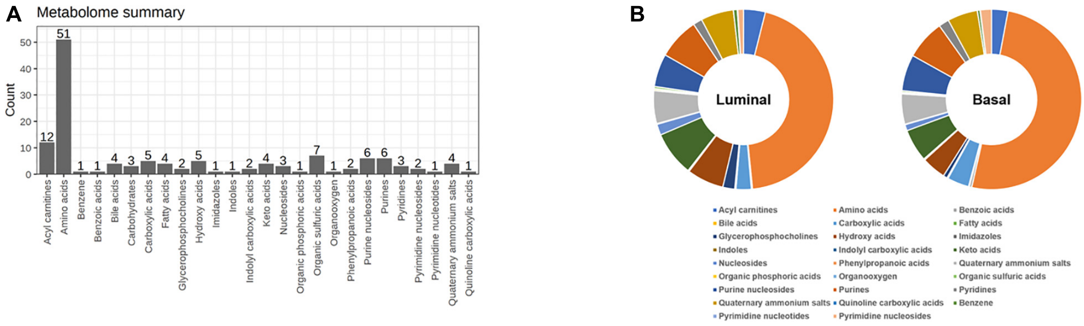


FIGURE 5 | Distinct metabolite profiles in basal and luminal MIBC subtypes. **(A)** The types and amounts of metabolites examined in this study. **(B)** Relative frequencies of metabolites in basal and luminal MIBC subtypes.

differential lipid elements between the basal and luminal MIBC subtypes are shown in **Figure 3B**. To explore the potential lipid biomarkers of basal and luminal MIBC subtypes, the following top 10 significantly differential lipid elements were analyzed: SL d18:1/24:1h, SM d18:1/20:0, SL d18:1/24:0h, SL d18:1/22:1, SL d18:1/22:0, LacCer d18:1/14:0, GM3 d18:1/22:1, SM d18:1/20:1, SM d18:1/18:1, and SM d18:1/22:1 (**Figure 3C**). Of these, SL d18:1/24:1h, SM d18:1/20:0, SL d18:1/24:0h, SL d18:1/22:1, SL d18:1/22:0, GM3 d18:1/22:1, SM d18:1/18:1, and SM d18:1/22:1 produced the highest AUC values (**Supplementary Figure 1**), indicating that these lipid elements could accurately separate basal and luminal MIBC subtypes, and these elements potentially to be targets of precision therapy in the future. In addition, the levels of total FFA and SL in the basal subtype were significantly higher than the luminal subtype, which displayed high AUC values (**Figures 3D,E**). These data indicated that SL d18:1/24:1h, SM d18:1/20:0, SL d18:1/24:0h, SL d18:1/22:1, SL d18:1/22:0, GM3 d18:1/22:1, SM d18:1/18:1, SM d18:1/22:1, FFA, and SL had potencies to be biomarkers for precisely distinguishing basal and luminal MIBC subtypes.

Potential Lipid Biomarkers Are Associated With Tumor Microenvironment

Tumor microenvironment is composed of numerous cell types and greatly influences tumor progression and therapy response (Pfannstiel et al., 2019). We measured the relative frequencies of immune and stromal cell types using a new algorithm based on transcriptome profiles called "xCell" (Aran et al., 2017). The analysis showed that the relative frequencies of cell types in basal and luminal MIBC subtypes greatly differed (**Figure 4A**). Pearson correlation analysis showed that SL levels of samples were strongly related to B cells, CD8 + T cell, macrophages M2, natural killer T (NKT) cells, mast cells, endothelial cells, and fibroblasts values, while FFA levels of samples were closely related to mesenchymal stem cell (MSC) and regulatory T cell (Treg) values (**Figure 4B**). These data suggested that SL and FFA were both strongly associated with tumor microenvironment and may play key roles in MIBC progression.

Distinct Metabolite Profiles in Basal and Luminal MIBC Subtypes

To map the differential metabolite profile between basal and luminal MIBC subtypes, 133 metabolites were measured (**Figure 5A**). The metabolite profiles of basal differed from luminal MIBC subtypes (**Figure 5B**), and the differential metabolites are shown in **Table 2**.

TABLE 2 | Differential metabolites of basal and luminal subtype (basal vs. luminal).

Elevated metabolites	Log2FC	*p*-value	Declined metabolites	Log2FC	*p*-value
Arabinonic acid	2.2107	0.000*	Glutathione	−1.964	0.025*
Allantoin	1.8753	0.032*	Oxidized glutathione	−1.6275	0.020*
Gamma-Glutamyl Glutamine	1.7797	0.000*	Glycerophosphocholine	−1.6203	0.014*
Pyroglutamic acid	1.3859	0.002*	Butyrylcarnitine	−1.5932	0.001*
Glyceric acid	1.0352	0.033*	L-Malic acid	−1.1582	0.006*
Uridine	0.96311	0.031*	R-3-Hydroxybutyric acid	−1.145	0.025*
Uric acid	0.86495	0.022*	Propionylcarnitine	−1.01	0.023*
Glutaric acid	0.7218	0.016*	3′-AMP	−1.2204	0.052
tert-Butyl 3-amino-1,4,6,7-tetrahydro-5H-pyrazolo4,3-cpyridine-5-carboxylate	0.58832	0.014*	Pivaloylcarnitine	−2.2577	0.086
5-methoxy-L-tryptophan	2.2235	0.097	Xanthine	−1.3516	0.054
Methionine sulfoxide	1.3653	0.061	S-Glutathionyl-L-cysteine	−0.96333	0.096
N-Acetylleucine	1.3234	0.075	Leucyl-Serine	−0.79759	0.080
Taurodeoxycholic acid	0.97946	0.062	N-Acetyl-L-alanine	−0.74714	0.064
8-Hydroxydeoxyguanosine	0.85958	0.065	Succinyladenosine	−0.62835	0.062
Guanine	0.68097	0.074			

** indicates $p < 0.05$.*

Potential Metabolite Biomarkers of Basal and Luminal MIBC Subtypes

To further reveal the potential metabolite biomarkers in basal and luminal MIBC subtypes, we employed PLS-DA analysis to evaluate metabolite VIP scores. Based on the VIP score rank, the top 10 metabolites were identified: tyrosyl-alanine, pyroglutamic acid, 5-methoxy-L-tryptophan, citric acid, uridine, and uric acid were increased in the basal subtype, while glutathione, pyruvic acid, oxidized glutathione, glycerophosphocholine, creatine, L-lactic acid, S-glutathionyl-L-cysteine, L-malic acid, and 3′-adenosine monophosphate (3′-AMP) were increased in the luminal subtype (**Figure 6A**). The top 25 differential metabolites are shown in **Figure 6B**. The peak intensities of the top 10 significantly different metabolites in basal and luminal MIBC subtypes are shown in **Figure 6C**. To further identify potential metabolite biomarkers in basal and luminal MIBC subtypes, we analyzed the levels of the main types of metabolites. It was found that the levels of glycerophosphocholine (GCP), hydroxy acids, and nucleosides increased in the luminal subtype, while the levels of imidazoles and pyrimidine nucleoside were higher in the basal than in the luminal subtype. These metabolites presented different AUC values (**Figure 6D**). Remarkably, the ratios of GCP/imidazoles (AUC = 1) and nucleosides/imidazoles (AUC = 0.9714) had higher AUC values than GCP (AUC = 0.8857), nucleosides (AUC = 0.8571), or imidazoles (AUC = 0.9143) levels alone (**Figure 6E**). The above results indicated that the ratios of GCP/imidazoles and nucleosides/imidazoles had a greater capacity to differentiate basal and luminal MIBC subtypes than the single metabolites; these ratios could be used as potential biomarkers to distinguish basal and luminal MIBC subtypes.

DISCUSSION

Muscle invasive bladder cancer is a molecularly heterogeneous disease with high recurrence rates and poor prognosis (Prasad et al., 2011; Meeks et al., 2020). The BASE47 classifier divides MIBC into basal and luminal subtypes based on transcriptome expression profiles. The differentiation pattern, histological characteristic, overall survival, and therapy response of basal and luminal MIBC subtypes are significantly different (Kamoun et al., 2020). This classifier provides a new framework for studying MIBC heterogeneity and has potential values for clinical application (Ochoa et al., 2016; Fong et al., 2020). Metabolic reprogramming of tumors drives tumor progression by many aspects (Pavlova and Thompson, 2016). Although previous studies have explored the metabolic profile and identified metabolites associated with recurrence and poor prognosis of BC (Armitage and Ciborowski, 2017; Loras et al., 2018; Zhang et al., 2018), the differential lipids and metabolites between basal and luminal MIBC subtypes remain unclear. Knowledge of these profiles may provide potential biomarkers and therapy targets for clinical application. In this study, we integrated transcriptomics, lipidomics, and metabolomics analysis to reveal the differential lipid and metabolite profiles between basal and luminal MIBC subtypes, providing potential lipid and metabolite biomarkers for precision therapy of MIBC.

According to the BASE47 classifier, we divided MIBC patients into basal and luminal subtypes based on transcriptomic expression profiles. RNA-sequencing analysis revealed that the lipid and metabolic pathways of basal and luminal MIBC subtypes differed significantly, which suggested that basal and luminal MIBC subtype potentially underwent lipid and metabolic reprogramming (Lee et al., 2018). To further explore the lipid profiles of basal and luminal MIBC subtypes, we evaluated 417 tissue lipid elements in basal and luminal MIBC subtypes. Results showed that there were significant differences in the lipid profiles of basal and luminal MIBC subtypes. The top 10 differential lipid

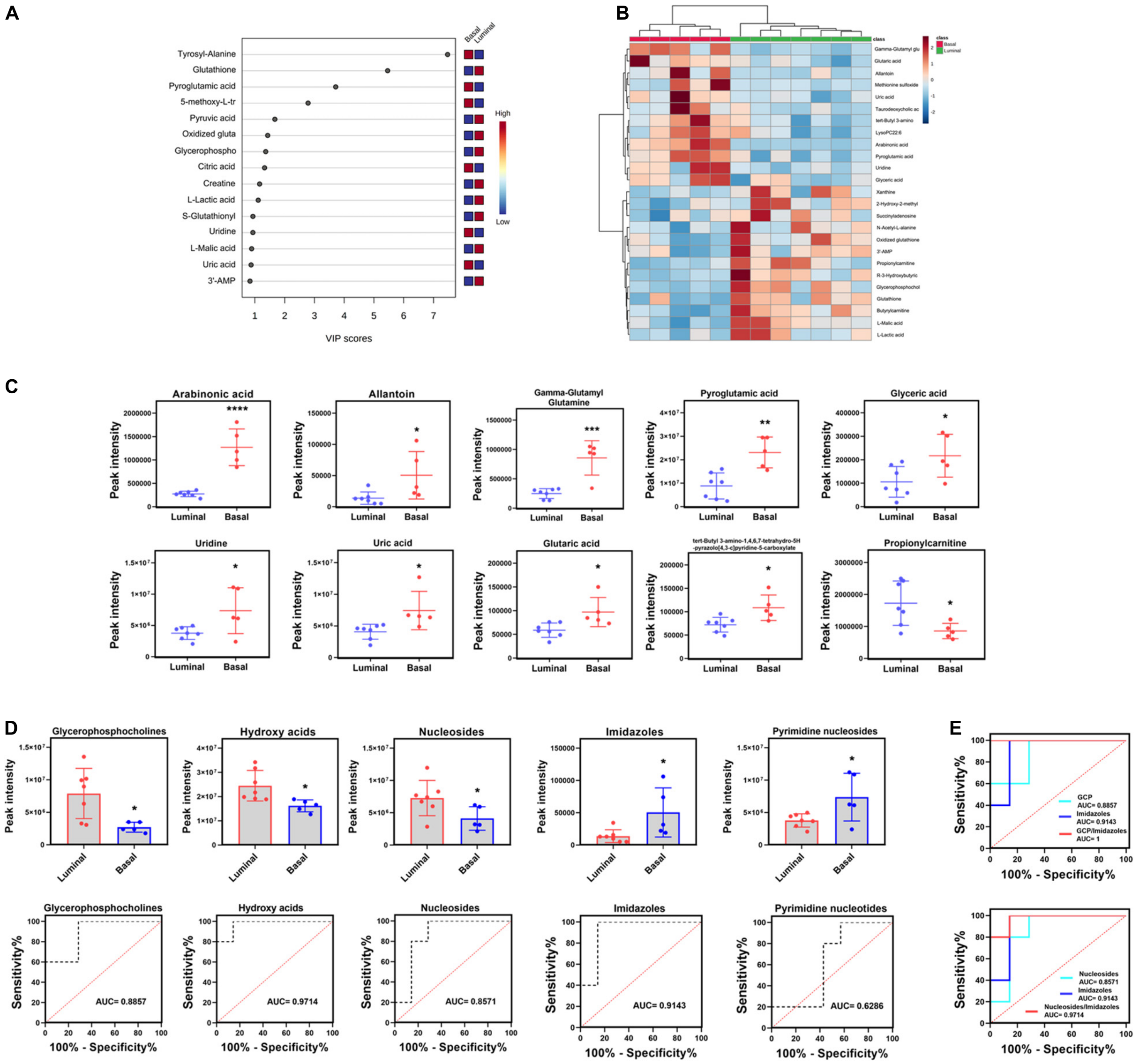


FIGURE 6 | Potential metabolite biomarkers of basal and luminal MIBC subtypes. **(A)** VIP score of altered metabolites. **(B)** Heatmap of the top 25 altered metabolites in basal and luminal MIBC subtypes. **(C)** The peak intensity of the top 10 significantly differential metabolites in basal and luminal MIBC subtypes. **(D)** The peak intensity and AUC values of GCP, hydroxy acids, nucleosides, imidazoles, and pyrimidine nucleosides. **(E)** The AUC values of GCP/imidazoles and nucleosides/imidazoles ratios. * indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$; **** indicates $p < 0.0001$.

elements were enriched in the basal subtype, eight of which exhibited maximum AUC values and could be considered as potential biomarkers: SL d18:1/24:1h, SM d18:1/20:0, SL d18:1/24:0h, SL d18:1/22:1, SL d18:1/22:0, GM3 d18:1/22:1, SM d18:1/18:1, and SM d18:1/22:1. Due to the small sample size, the differential lipids identified between basal and luminal MIBC subtypes may be limited. Thus, in our cohort, the top 10 differential lipids between basal and luminal MIBC subtypes were all enriched in basal MIBC subtype. Furthermore, examination of the main types of lipids revealed that the total FFA and SL levels of the basal samples were higher than that of the luminal samples. These lipids display strong potencies to be biomarkers. Additionally, according to a new algorithm, we inferred the relative frequencies of immune and stromal cells in samples based on their mRNA profiles. Pearson correlation analysis showed that FFA and SL were significantly related to specific immune and stromal cell types in the tumor microenvironment. Indeed, FFA drives tumor progression by stimulating cancer cell proliferation and promotes CD8 + TRM cells to persist in tumor tissue to mediate protective immunity (Iwamoto et al., 2018; Zhang et al., 2020). Meanwhile, SL is involved in cancer progression and improves sensitivity of tumor cells to microenvironmental stress factors including hypoxia and anticancer drugs (Suchanski and Ugorski, 2016; Suchanski et al., 2018). Therefore, FFA and SL may play important roles in MIBC progression and potentially used to be biomarkers of basal and luminal MIBC subtypes.

During tumor reprogramming, metabolic patterns of cancer cells are changed to adapt to the new microenvironments, which makes it important to deeply understand cancer metabolic profiles (Kim and DeBerardinis, 2019; La Vecchia and Sebastian, 2020). To reveal the differential metabolite profiles between basal and luminal MIBC subtypes, we evaluated a total of 133 metabolites. Our results suggested that GCP, hydroxy acids, nucleosides, imidazoles, and pyrimidine nucleosides could accurately distinguish the basal subtype from the luminal subtype. Furthermore, the AUCs of the GCP/imidazoles and nucleosides/imidazoles ratios were higher than those of GCP, nucleosides, and imidazoles alone, suggesting that these ratios were more sensitive for distinguishing basal from luminal MIBC subtypes. According to previous reports, GCP, nucleosides, and imidazoles drive cancer progression; they are associated with poor prognosis of several types of cancer (Moestue et al., 2012; Dolinar et al., 2018; Long and Wang, 2019). Therefore, the GCP/imidazoles and nucleosides/imidazoles ratios have potential clinical applications as biomarkers, while GCP, nucleosides, and imidazoles may be the targets of MIBC precision therapy.

The occurrence and development of tumor is a complex process, which is coregulated by genomics, epigenomics, transcriptomics, proteomics, metabolomics, microbiome, and other factors (Menyhart and Gyorffy, 2021). Single omics studies cannot fully reveal the characteristics of tumors and provide reliable biomarkers. In this study, the integration of transcriptomics, lipidomics, and metabonomics can be used to develop subtype-specific biomarkers and therapeutic targets and may provide more precise predictions for disease progression and prognosis. However, it should be noted that this study has some limitations. First, the sample size was small. Additional larger and independent cohorts should be analyzed to reveal more valuable lipidomic and metabonomic biomarkers. Second, this study only explored the differential lipid and metabolite profiles between basal and luminal MIBC subtypes. The accuracy and sensitivity of the potential biomarkers identified here needed to be confirmed in larger cohorts. Third, there is no strict exclusion to some potential conditions that influence lipid and metabolite profiles from our analysis, such as diabetes and hyperlipemia.

In conclusion, our study integrated transcriptomic, lipidomic, and metabolomic analysis to reveal the differential lipid and metabolite profiles between basal and luminal MIBC subtypes. It was also found that FFA, SL, the GCP/imidazoles, and nucleosides/imidazoles ratios have strong potencies to be biomarkers for distinguishing basal from luminal MIBC subtypes.

DATA AVAILABILITY STATEMENT

All the raw data used in this manuscript will be made available to any qualified researcher without reservation. The data presented in the study are deposited in the GEO repository, accession number: GSE179440.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of The First Affiliated Hospital of Guangxi Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TL, QW, and ZT designed the study. CF, LP, and ST wrote the manuscript. LH, XW, YT, YX, and ZL analyzed the results. All authors contributed to the manuscript and approved the submitted version.

FUNDING

This work was supported by the Guangxi Clinical Research Center for Urology and Nephrology (2020AC03006), Guangxi Science and Technology Base and Talent Special Fund (2019AC17009, AB18126055, and 2016GXNSFGA38006), the National Natural Science Foundation of China (81760454 and 82060460), and Innovation Project of Guangxi Graduate Education (YCBZ2021052).

ACKNOWLEDGMENTS

The authors would like to thank all participants for providing the samples used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.695662/full#supplementary-material

Supplementary Figure 1 | The AUC of eight lipids with differential distributions in basal and luminal MIBC subtypes. The AUC values of SL d18:1/24:1h, SM d18:1/20:0, SL d18:1/24:0h, SL d18:1/22:1, SL d18:1/22:0, GM3 d18:1/22:1, SM d18:1/18:1, and SM d18:1/22:1.

REFERENCES

Alberice, J. V., Amaral, A. F., Armitage, E. G., Lorente, J. A., Algaba, F., Carrilho, E., et al. (2013). Searching for urine biomarkers of bladder cancer recurrence using a liquid chromatography-mass spectrometry and capillary electrophoresis-mass spectrometry metabolomics approach. *J. Chromatogr. A* 1318, 163–170. doi: 10.1016/j.chroma.2013.10.002

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 18:220. doi: 10.1186/s13059-017-1349-1

Armitage, E. G., and Ciborowski, M. (2017). Applications of metabolomics in cancer studies. *Adv. Exp. Med. Biol.* 965, 209–234. doi: 10.1007/978-3-319-47656-8_9

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507, 315–322. doi: 10.1038/nature12965

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Choi, W., Czerniak, B., Ochoa, A., Su, X., Siefker-Radtke, A., Dinney, C., et al. (2014a). Intrinsic basal and luminal subtypes of muscle-invasive bladder cancer. *Nat. Rev. Urol.* 11, 400–410. doi: 10.1038/nrurol.2014.129

Choi, W., Porten, S., Kim, S., Willis, D., Plimack, E. R., Hoffman-Censits, J., et al. (2014b). Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell* 25, 152–165. doi: 10.1016/j.ccr.2014.01.009

Chong, J., Wishart, D. S., and Xia, J. (2019). Using metaboanalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Curr. Protoc. Bioinformatics* 68:e86. doi: 10.1002/cpbi.86

Damrauer, J. S., Hoadley, K. A., Chism, D. D., Fan, C., Tiganelli, C. J., Wobker, S. E., et al. (2014). Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3110–3115. doi: 10.1073/pnas.1318376111

Dolinar, K., Jan, V., Pavlin, M., Chibalin, A. V., and Pirkmajer, S. (2018). Nucleosides block AICAR-stimulated activation of AMPK in skeletal muscle and cancer cells. *Am. J. Physiol. Cell Physiol.* 315, C803–C817. doi: 10.1152/ajpcell.00311.2017

Fong, M. H. Y., Feng, M., McConkey, D. J., and Choi, W. (2020). Update on bladder cancer molecular subtypes. *Transl. Androl. Urol.* 9, 2881–2889. doi: 10.21037/tau-2019-mibc-12

Frantzi, M., van Kessel, K. E., Zwarthoff, E. C., Marquez, M., Rava, M., Malats, N., et al. (2016). Development and validation of urine-based peptide biomarker panels for detecting bladder cancer in a multi-center study. *Clin. Cancer Res.* 22, 4077–4086. doi: 10.1158/1078-0432.CCR-15-2715

Ghosh, S., and Chan, C. K. (2016). Analysis of RNA-Seq data using TopHat and Cufflinks. *Methods Mol. Biol.* 1374, 339–361. doi: 10.1007/978-1-4939-3167-5_18

Grossman, H. B., Natale, R. B., Tangen, C. M., Speights, V. O., Vogelzang, N. J., Trump, D. L., et al. (2003). Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer. *N. Engl. J. Med.* 349, 859–866. doi: 10.1056/NEJMoa022148

International Collaboration of Trialists, Medical Research Council Advanced Bladder Cancer Working Party, European Organisation for Research and Treatment of Cancer Genito-Urinary Tract Cancer Group, Australian Bladder Cancer Study Group, National Cancer Institute of Canada Clinical Trials Group, Finnbladder, et al. (2011). International phase III trial assessing neoadjuvant cisplatin, methotrexate, and vinblastine chemotherapy for muscle-invasive bladder cancer: long-term results of the BA06 30894 trial. *J. Clin. Oncol.* 29, 2171–2177. doi: 10.1200/JCO.2010.32.3139

Iwamoto, H., Abe, M., Yang, Y., Cui, D., Seki, T., Nakamura, M., et al. (2018). Cancer lipid metabolism confers antiangiogenic drug resistance. *Cell Metab.* 28, 104–117.e5. doi: 10.1016/j.cmet.2018.05.005

Kamat, A. M., Hahn, N. M., Efstathiou, J. A., Lerner, S. P., Malmstrom, P. U., Choi, W., et al. (2016). Bladder cancer. *Lancet* 388, 2796–2810. doi: 10.1016/S0140-6736(16)30512-8

Kamoun, A., de Reynies, A., Allory, Y., Sjodahl, G., Robertson, A. G., Seiler, R., et al. (2020). A Consensus molecular classification of muscle-invasive bladder cancer. *Eur. Urol.* 77, 420–433. doi: 10.1016/j.eururo.2019.09.006

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kim, J., and DeBerardinis, R. J. (2019). Mechanisms and implications of metabolic heterogeneity in cancer. *Cell Metab.* 30, 434–446. doi: 10.1016/j.cmet.2019.08.013

La Vecchia, S., and Sebastian, C. (2020). Metabolic pathways regulating colorectal cancer initiation and progression. *Semin. Cell Dev. Biol.* 98, 63–70. doi: 10.1016/j.semcdb.2019.05.018

Lam, S. M., Wang, R., Miao, H., Li, B., and Shui, G. (2018). An integrated method for direct interrogation of sphingolipid homeostasis in the heart and brain tissues of mice through postnatal development up to reproductive senescence. *Anal. Chim. Acta* 1037, 152–158. doi: 10.1016/j.aca.2018.01.015

Lam, S. M., Wang, Z., Li, J., Huang, X., and Shui, G. (2017). Sequestration of polyunsaturated fatty acids in membrane phospholipids of *Caenorhabditis elegans* dauer larva attenuates eicosanoid biosynthesis for prolonged survival. *Redox Biol.* 12, 967–977. doi: 10.1016/j.redox.2017.05.002

Larrouy-Maumus, G. (2019). Lipids as biomarkers of cancer and bacterial infections. *Curr. Med. Chem.* 26, 1924–1932. doi: 10.2174/0929867325666180904120029

Lee, M. Y., Yeon, A., Shahid, M., Cho, E., Sairam, V., Figlin, R., et al. (2018). Reprogrammed lipid metabolism in bladder cancer with cisplatin resistance. *Oncotarget* 9, 13231–13243. doi: 10.18632/oncotarget.24229

Long, Y., and Wang, D. (2019). Inhibition of colon cancer cell growth by imidazole through activation of apoptotic pathway. *Med. Sci. Monit.* 25, 7597–7604. doi: 10.12659/MSM.917779

Loras, A., Trassierra, M., Sanjuan-Herraez, D., Martinez-Bisbal, M. C., Castell, J. V., Quintas, G., et al. (2018). Bladder cancer recurrence surveillance by urine metabolomics analysis. *Sci. Rep.* 8:9172. doi: 10.1038/s41598-018-27538-3

McConkey, D. J., and Choi, W. (2018). Molecular subtypes of bladder cancer. *Curr. Oncol. Rep.* 20:77. doi: 10.1007/s11912-018-0727-5

Meeks, J. J., Al-Ahmadie, H., Faltas, B. M., Taylor, J. A. III, Flaig, T. W., DeGraff, D. J., et al. (2020). Genomic heterogeneity in bladder cancer: challenges and possible solutions to improve outcomes. *Nat. Rev. Urol.* 17, 259–270. doi: 10.1038/s41585-020-0304-1

Menyhart, O., and Gyorffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.* 19, 949–960. doi: 10.1016/j.csbj.2021.01.009

Moestue, S. A., Giskeodegard, G. F., Cao, M. D., Bathen, T. F., and Gribbestad, I. S. (2012). Glycerophosphocholine (GPC) is a poorly understood biomarker in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 109:E2506; author reply E2507. doi: 10.1073/pnas.1208226109

Munir, R., Lisec, J., Swinnen, J. V., and Zaidi, N. (2019). Lipid metabolism in cancer cells under metabolic stress. *Br. J. Cancer* 120, 1090–1098. doi: 10.1038/s41416-019-0451-4

Nuhn, P., May, M., Sun, M., Fritsche, H. M., Brookman-May, S., Buchner, A., et al. (2012). External validation of postoperative nomograms for prediction of all-cause mortality, cancer-specific mortality, and recurrence in patients with urothelial carcinoma of the bladder. *Eur. Urol.* 61, 58–64. doi: 10.1016/j.eururo.2011.07.066

Ochoa, A. E., Choi, W., Su, X., Siefker-Radtke, A., Czerniak, B., Dinney, C., et al. (2016). Specific micro-RNA expression patterns distinguish the basal and luminal subtypes of muscle-invasive bladder cancer. *Oncotarget* 7, 80164–80174. doi: 10.18632/oncotarget.13284

Pavlova, N. N., and Thompson, C. B. (2016). The emerging hallmarks of cancer metabolism. *Cell Metab.* 23, 27–47. doi: 10.1016/j.cmet.2015.12.006

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Pfannstiel, C., Strissel, P. L., Chiappinelli, K. B., Sikic, D., Wach, S., Wirtz, R. M., et al. (2019). The tumor immune microenvironment drives a prognostic relevance that correlates with bladder cancer subtypes. *Cancer Immunol. Res.* 7, 923–938. doi: 10.1158/2326-6066.CIR-18-0758

Prasad, S. M., Decastro, G. J., Steinberg, G. D., and Medscape. (2011). Urothelial carcinoma of the bladder: definition, treatment and future efforts. *Nat. Rev. Urol.* 8, 631–642. doi: 10.1038/nrurol.2011.144

Putluri, N., Shojaie, A., Vasu, V. T., Vareed, S. K., Nalluri, S., Putluri, V., et al. (2011). Metabolomic profiling reveals potential markers and bioprocesses altered in bladder cancer progression. *Cancer Res.* 71, 7376–7386. doi: 10.1158/0008-5472.CAN-11-1154

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., et al. (2017). Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* 171, 540–556.e25. doi: 10.1016/j.cell.2017.09.007

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., et al. (2018). Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* 174:1033. doi: 10.1016/j.cell.2018.07.036

Sahu, D., Lotan, Y., Wittmann, B., Neri, B., and Hansel, D. E. (2017). Metabolomics analysis reveals distinct profiles of nonmuscle-invasive and muscle-invasive bladder cancer. *Cancer Med.* 6, 2106–2120. doi: 10.1002/cam4.1109

Sjodahl, G., Eriksson, P., Liedberg, F., and Hoglund, M. (2017). Molecular classification of urothelial carcinoma: global mRNA classification versus tumour-cell phenotype classification. *J. Pathol.* 242, 113–125. doi: 10.1002/path.4886

Sjodahl, G., Lauss, M., Lovgren, K., Chebil, G., Gudjonsson, S., Veerla, S., et al. (2012). A molecular taxonomy for urothelial carcinoma. *Clin. Cancer Res.* 18, 3377–3386. doi: 10.1158/1078-0432.CCR-12-0077-T

Song, J. W., Lam, S. M., Fan, X., Cao, W. J., Wang, S. Y., Tian, H., et al. (2020). Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. *Cell Metab.* 32, 188–202.e5. doi: 10.1016/j.cmet.2020.06.016

Suchanski, J., and Ugorski, M. (2016). [The biological role of sulfatides]. *Postepy Hig. Med. Dosw.* 70, 489–504. doi: 10.5604/17322693.1201720

Suchanski, J., Grzegrzolka, J., Owczarek, T., Pasikowski, P., Piotrowska, A., Kocbach, B., et al. (2018). Sulfatide decreases the resistance to stress-induced apoptosis and increases P-selectin-mediated adhesion: a two-edged sword in breast cancer progression. *Breast Cancer Res.* 20:133. doi: 10.1186/s13058-018-1058-z

Yuan, M., Breitkopf, S. B., Yang, X., and Asara, J. M. (2012). A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. *Nat. Protoc.* 7, 872–881. doi: 10.1038/nprot.2012.024

Yumba Mpanga, A., Siluk, D., Jacyna, J., Szerkus, O., Wawrzyniak, R., Markuszewski, M., et al. (2018). Targeted metabolomics in bladder cancer: from analytical methods development and validation towards application to clinical samples. *Anal. Chim. Acta* 1037, 188–199. doi: 10.1016/j.aca.2018.01.055

Zargar, H., Espiritu, P. N., Fairey, A. S., Mertens, L. S., Dinney, C. P., Mir, M. C., et al. (2015). Multicenter assessment of neoadjuvant chemotherapy for muscle-invasive bladder cancer. *Eur. Urol.* 67, 241–249. doi: 10.1016/j.eururo.2014.09.007

Zhang, L., Han, L., He, J., Lv, J., Pan, R., and Lv, T. (2020). A high serum-free fatty acid level is associated with cancer. *J. Cancer Res. Clin. Oncol.* 146, 705–710. doi: 10.1007/s00432-019-03095-8

Zhang, W. T., Zhang, Z. W., Guo, Y. D., Wang, L. S., Mao, S. Y., Zhang, J. F., et al. (2018). Discovering biomarkers in bladder cancer by metabolomics. *Biomark. Med.* 12, 1347–1359. doi: 10.2217/bmm-2018-0229

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 August 2021
doi: 10.3389/fgene.2021.702259

Cascade Deep Forest With Heterogeneous Similarity Measures for Drug–Target Interaction Prediction

Ying Zheng and Zheng Wu*

School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha, China

Drug repositioning is a method of systematically identifying potential molecular targets that known drugs may act on. Compared with traditional methods, drug repositioning has been extensively studied due to the development of multi-omics technology and system biology methods. Because of its biological network properties, it is possible to apply machine learning related algorithms for prediction. Based on various heterogeneous network model, this paper proposes a method named THNCDF for predicting drug–target interactions. Various heterogeneous networks are integrated to build a tripartite network, and similarity calculation methods are used to obtain similarity matrix. Then, the cascade deep forest method is used to make prediction. Results indicate that THNCDF outperforms the previously reported methods based on the 10-fold cross-validation on the benchmark data sets proposed by Y. Yamanishi. The area under Precision Recall curve (AUPR) value on the Enzyme, GPCR, Ion Channel, and Nuclear Receptor data sets is 0.988, 0.980, 0.938, and 0.906 separately. The experimental results well illustrate the feasibility of this method.

Keywords: drug repositioning, drug discovery, drug–target interaction, heterogeneous similarity measures, cascade deep forest

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Yijie Ding,
Suzhou University of Science and Technology, China
Shichao Liu,
Huazhong Agricultural University, China

****Correspondence:***
Ying Zheng
zhengying@csust.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *29 April 2021*
Accepted: *24 May 2021*
Published: *24 August 2021*

Citation:
Zheng Y and Wu Z (2021) Cascade Deep Forest With Heterogeneous Similarity Measures for Drug–Target Interaction Prediction. Front. Genet. 12:702259. doi: 10.3389/fgene.2021.702259

INTRODUCTION

In the past few decades, investment in drug research and development has grown rapidly, but most drugs have failed in the first phase of clinical trials. Moreover, it normally costs billions of dollars and consumes 10 years for any drug to be put on the market completely (Roessler et al., 2021). At present, drug repositioning has a wide prospect and provides evidence for further drug discovery, whose purpose is to determine potential therapeutic targets for existing drugs, thereby saving time and minimizing risks of conventional drug development (Stein et al., 2021).

The key of drug repositioning hinges on identifying drug–target interaction (DTI), which exerts a vital role in drug research and development (Badkas et al., 2020). Currently, traditional experimental approaches are either time consuming or high costly. Despite that potential drug indications can be directly detected by target or cell screening of thousands of drugs in synthetic databases, there are still hurdles to massively relocate drugs owing to the needs of collecting existing drugs, specialized equipment, and screening tests (Turanli et al., 2018).

In general, the traditional methods for calculating drug target interactions mainly consist of ligand method and structure method (Huang et al., 2020; Yang et al., 2020; Zhang et al., 2020). The ligand-based methods predict potential DTI *via* contrasting candidate ligands with known ligands capable of binding to them, but it does not perform well in the absence of ligand information for potential targets (Juárez-Saldivar et al., 2020). The structure-based method mainly uses the docking simulation technology to predict the potential DTI on the basis of known three-dimensional structure. In the same way, this method that relies on simulated docking's reliability often consumes a plenty of time and requires all drugs and targets to provide accurate and reliable three-dimensional structure (Vivarelli et al., 2020).

Along with sustainable innovative developments of biological data, and high-speed improvements of machine learning technology in recent years, a variety of methods for computational drug repositioning have been put forward correspondingly and achieved some achievements in practical applications (Lan et al., 2016, 2020; Chen et al., 2021; Li et al., 2019; Liu et al., 2019; Zeng et al., 2019; Fahimian et al., 2020; Rauschenbach et al., 2020; Zhou et al., 2020; Jarada et al., 2021; Meng et al., 2021). Machine learning is a beneficial complement to ligand-based and structure-based methods. It has been widely developed and applied as an effective method for pinpointing drug–targets as well as predicting drug-diseases. Machine learning is able to systematically integrate biological databases, with the purpose of predicting potential DTI and drug–disease interactions.

The method of similarity constrained probabilistic matrix factorization (SCPMF) is used for drug repositioning through recognizing novel drug–virus coactions (Meng et al., 2021). Moreover, SCPMF innovatively reconstructs the drug–virus interaction matrix, by dexterously projecting the drug–virus interaction matrix into two potential feature matrices for viruses and drugs. A new framework named Similarity Network Fusion and Neural Networks (SNF-NN) on the basis of deep learning was proposed and elaborated, which predicts new drug–disease interactions though using similarity selection relevant to drugs and diseases, similarity network fusion, and a novel neural network model with superior tuning (Jarada et al., 2021). By comparison of the performance of SNF-NN with that of nine benchmark machine learning methods, the robustness of SNF-NN is calculated. The values of AUC and AUPR are 0.867 and 0.876, respectively. Besides, a previous study has shown that a method based on network called RepCOOL is utilized for drug repositioning (Fahimian et al., 2020). The eventual model of drug repositioning is constructed on account of a random forest classifier. RepCOOL recommends four novel drugs for the treatment of breast cancer at stage II, namely, paclitaxel, doxorubicin, tamoxifen, and trastuzumab. In addition, a network embedding based method for predicting drug–disease interactions (NEDD) is raised (Zhou et al., 2020). Initially, through constructing a heterogeneous network and utilizing meta-paths of various lengths, NEDD accurately obtains the indirect associations between drugs and diseases or their strong proximity, thereby acquiring representation vectors of drugs and diseases with low dimensions. NEDD estimates novel relationships between diseases and drugs by utilizing a random forest classifier. A recent study has reported that a network-based method about deep learning for drug repositioning (deepDR) recognizes advanced characteristics of drugs from heterogeneous networks through a multi-mode autoencoder. Then, through a variational autoencoder, the obtained low-dimensional representation of the drug as well as clinically reported drug–disease pairs are uniformly encoded and decoded to infer candidates for approved drugs that were actually without initial approval (Zeng et al., 2019).

The main contributions of this paper are summarized as follows:

We study various calculation methods based on the tripartite heterogeneous network, and finally adopt the Gaussian kernel between each layer, and the Tanimoto's coefficient is used in the drug layer to calculate the chemical structure similarity matrix. Besides, the similarity matrix is fitted by all matrices;

We improve and adjust the parameters according to the gcForest (Zhou and Feng, 2019) method. We use 10-fold cross-validation to check the final prediction (termed THNCDF, Tripartite Heterogeneous Network Cascade Deep Forest).

We compare the results of THNCDF with four types of methods (Cao et al., 2014; Hao et al., 2016; Rayhan et al., 2017; Thafar et al., 2020). The experimental results show that the THNCDF method has good performance, and the area under Precision Recall curve (AUPR) values on the four benchmark data sets reach 0.988, 0.980, 0.938, and 0.906.

The rest of this paper is organized as follows. In Section 2, we introduce the data sets used for similarity measurement, and then we present the general framework and cascade deep forest methods with details in Section 3. In Section 4, the performance of our proposed THNCDF method is evaluated through extensive experiments. At the end, some discussions are provided in Section 5.

RELATED WORK

Data Sets

In our experiments, we use the data sets listed in **Table 1** to build a tripartite heterogeneous network model. **Table 1**

TABLE 1 | Sources and verification of databases.

Resource	Description	Url
DrugBank	Free accessible drug database	www.drugbank.ca/
DisGeNET	Free accessible human disease database	www.disgenet.org/
ChEMBL	Free accessible drug and target database	www.ebi.ac.uk/chembl/
Kegg	Free accessible database for molecular-level information	www.kegg.jp/
Uniprot	Free accessible protein sequence and annotation database	www.uniprot.org
OMIM	Free accessible compendium for Mendelian disorder	www.omim.org/

shows the exactly biologic data sets we used during the experiments (Yamanishi et al., 2008; Zheng and Wu, 2021). Especially, the main resource of the data set for the disease layer is from DisGeNET. This paper also uses a data set called DisGeNET approved, which contains FDA-approved drugs and their corresponding protein targets in the DisGeNET.

We will evaluate the performance of THNCDF on benchmark data sets. The benchmark data sets used in many DTI predictions were originally proposed by Y. Yamanishi, which have been considered as the golden data sets for comparing various DTI prediction methods. The benchmark data sets are listed in **Table 2**, which are downloaded from http://web.kuicr.kyotou.ac.jp/supp/yoshi/drugtarget/. The data sets include four subsets grouped by target classification: Enzyme, ion channel, GPCR (G protein-coupled receptor), and nuclear receptor. The largest subset, Enzyme, includes 445 drugs and 664 targets with 2,926 known DTI between them. Another NR, the smallest subset includes only 54 drugs and 26 targets with 90 known interactions. The other two subsets, IC and GPCR, consist of 210 and 223 drugs, 204 and 95 targets, and 1,476 and 635 known interactions, respectively.

Tripartite Heterogeneous Network

Based on the related ideas of pharmacology, the therapeutic effect of a single drug is relatively limited for diseases that are complex multiple pathological (Zamami et al., 2017; Zhu et al., 2020). Recently, the development of high-throughput biotechnology has produced a large amount of data. However, one of the main difficulties is how to collect and analyze the required biomedical data because they are heterogeneous and the data generated from different experiments include different types of information, such as nucleotide sequences and protein–protein interactions (Luo et al., 2020).

In this paper, we integrate the composition of many different heterogeneous networks and construct our novel tripartite heterogeneous network model according to different types of data. **Figure 1A** is the part of visualization of the Enzyme in benchmark data sets, in which the red nodes are drugs and the green nodes are targets. **Figure 1B** is the bipartite graph model of a part of **Figure 1A**; the red nodes are drugs, and the green nodes are targets in the same.

We construct a tripartite network that includes three layers: drugs, targets, and diseases. Correspondingly, two types of interactions, drug–target interactions and target–disease interactions, are interpreted as edges to connect nodes in these layers. We mainly focus on constructing the similarity matrix and feature information of the tripartite heterogeneous network.

TABLE 2 | Benchmark data sets.

Data sets	Drugs	Targets	n_d/n_t	Interactions
Enzyme	445	664	0.667	2,926
Ion channel	210	204	1.03	1,476
GPCR	223	95	2.35	635
Nuclear receptor	54	26	2.08	90

MATERIALS AND METHODS

In this study, we propose THNCDF, a new computational approach for molecular target identification from known drug–target centered DTI prediction. It utilizes low-dimensional but informative matrix representations of features for both drugs and targets through a cascade deep forest classifier in prediction of DTI (Zheng and Wu, 2021).

As shown in **Figure 2**, THNCDF mainly includes three steps: (1) Data integration and complete heterogeneous network is obtained, which contains diverse cheminformatics and bioinformatics profiles; (2) Similarity matrix calculation and parameter setting; (3) Application of cascade deep forest classifier and verification of the results.

Similarity of Medicinal Chemical Structures

To ensure that the features in the network model are distinguishable, the similarity of the medicinal chemical structure is a relatively objective feature (Zheng and Wu, 2021). In particular, the chemical structure of various drugs under the same standard can be obtained through the simplified molecular-input line-entry system (SIMILES), and then converted into 166-bits string of a certain length fingerprint. Thus, each fingerprint represents a unique drug. Through the calculation of Tanimoto's coefficient, the similarity matrix of medicinal chemical structure among all drugs is obtained. The formula for calculating Tanimoto's coefficient is shown in Equation (1).

$$SIM_{chem} = \frac{|f(dx)\times f(dy)|}{|f(dx)+f(dy)|-|f(dx)\times f(dy)|} \tag{1}$$

where $f_{(dx)}$ is the binary chemical fingerprint of drug x. According to Equation (1), a matrix of chemical structure similarity is constructed.

Gaussian Kernel Similarity

The Gaussian kernel is defined as the unimodal of the Euclidean distance between any two points in the network (Zheng and Wu, 2021). In THNCDF method, the Gaussian kernel is mainly used to calculate the feature of the connection between two layers, like the edge between the drug layer and the target layer or between the target layer and the disease layer. Also, for drug–drug interactions and target–target interactions, the Gaussian Kernel can calculate the edges in the same layer. Therefore, the calculation formula is commonly used to construct various types of matrices, such as the drug–drug interactions similarity matrix, target–target interactions similarity matrix, and target–disease interactions similarity matrix. The calculation formula is as follows:

$$K_{GIP,d}(D_i, D_j) = exp(-\gamma_d||yd_i - yd_j||^2) \tag{2}$$

$$\gamma_d = \gamma_d^{'}/(\frac{1}{m}\sum_{i=1}^{m}||yd_i||^2) \tag{3}$$

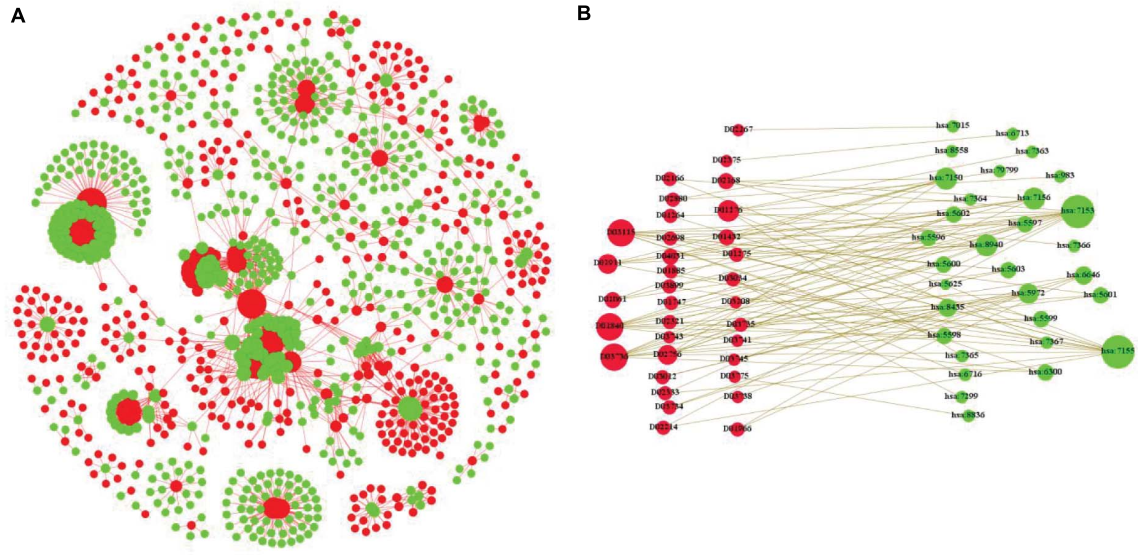


FIGURE 1 | An example of bipartite graph for drug–target interactions. **(A)** Is the part of visualization of the Enzyme in benchmark data sets, in which the red nodes are drugs and the green nodes are targets. **(B)** Is the bipartite graph model of a part of **Figure 1A**.

where D_i is defined as the *i-th* drug in the drug set, T_i represents the *i-th* target in the target set, while ts_i represents the *i-th* target in the target–disease interactions set. m is the size of drug set, while n and k represent the size of target set and the size of target–disease interactions set, respectively. The adjacency matrix $Y \in m \quad n$ represents the known drug–target interactions. If the drug and the target have an existing interaction, the value is 1; otherwise, the value is 0. $yd_i \quad \{y_{i1}, y_{i2}, ..., y_{in}\}$ is defined as the correlation vector between the drug d_i and all targets; meanwhile, $yts_i \quad \{y_{i1}, y_{i2}, ..., y_{in}\}$ is defined as the correlation vector between the target ts_i and all diseases. γ_d, γ_t, and γ_{ts} are adjustment parameters that control the width of the kernel, where$\gamma_d^{'}$, $\gamma_t^{'}$, and $\gamma_{ts}^{'}$ are set to 1 by using Gaussian kernels.

Similarity Matrix Fusion

According to the above multiple similarity matrices, we construct a kernel containing the spatial information of drugs and targets (Ding et al., 2018, 2020a,b; Zheng and Wu, 2021). Since the similarity matrix is not a positive definite matrix, predictions are ultimately required. We linearly fit the similarity matrix of drug chemical structure, the drug Gaussian kernel, the target Gaussian kernel, and the disease Gaussian kernel. We also set the weighted factors in the following equations empirically.

$$SIM_{drug}(d_x, d_y) = (1-\alpha) \times K_{GIP,d}(d_x, d_y) + \alpha \times SIM_{chem}(d_x, d_y) \tag{4}$$

$$SIM_{tar}(t_x, t_y) = (1-\alpha) \times K_{GIP,t}(t_x, t_y) + \alpha \times K_{GIP,S}(ts_x, ts_y) \tag{5}$$

The result of similarity matrices is used as the original input of the next step. In the latter experiments, in order to balance the constructed similarity matrix, the ratio of 0.5:0.5 is used with parameter setting.

Cascade Deep Forest

Random forest, developed by Bermain and Culter (Breiman, 2001), is widely used due to its excellent stability and resistance to overfitting. Nowadays, random forest has been successfully applied to the analysis of multiple biological and pharmacological contexts, such as Diabetic Retinopathy screening procedure (Alabdulwahhab et al., 2021) and detection of copy number variations for uncovering genetic factors (Zhuang et al., 2020). But in novel review by Zhou et al., deep learning based on non-differentiable modules exhibits the possibility of constructing deep models without using backpropagation. They have proposed the gcForest approach, which has generated three characteristics: layer-by-layer processing, in-model feature transformation, and sufficient model complexity. It provides an alternative methods to deep neural networks (DNNs) to learn hyper-level representations at a low computational cost. gcForest is a novel decision tree ensemble, with a cascade structure. It has much fewer hyper-parameters than DNNs, which the training process does not rely on backpropagation. In fact, the most important value of gcForest approach is it may open a door for non-NN style deep learning, or deep models based on non-differentiable modules. An extended depiction and the study of the theory on random forest or gcForest can be referred to the Web site of Bremain or the paper of Zhou et al.

Based on the advantage of random forest and characteristics of gcForest, we construct the THNCDF method, which includes the similarity matrices described above and utilizes improved gcForest approach for prediction. First, the fusion similarity matrix is the origin input for cascade structure of deep forest. Each level of cascade receives the feature information processed by its previous level and outputs its processing result to the next level. All level is an ensemble of decision tree forest. For example, each forest will count the percentages of different classes

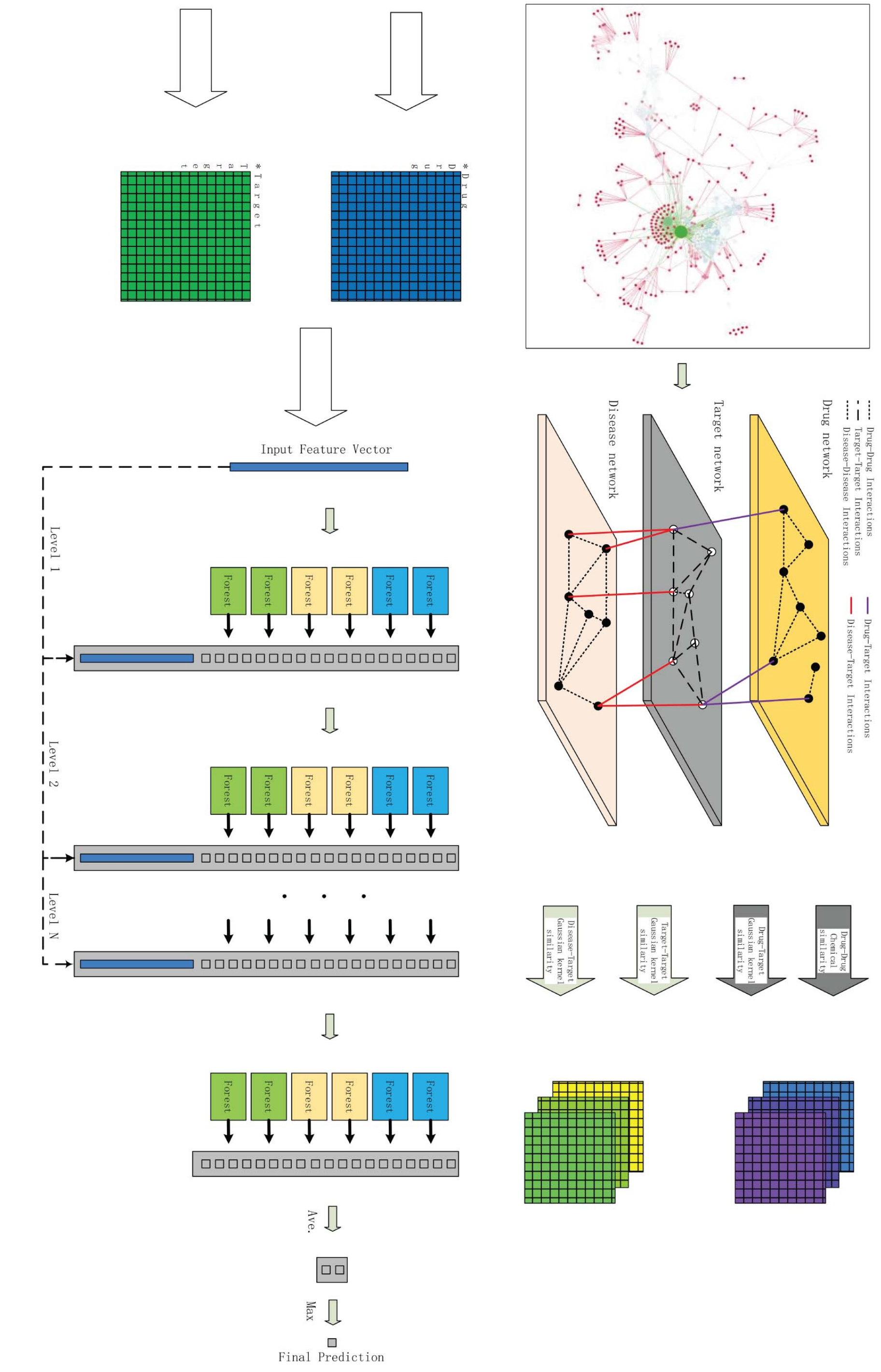

*Target
*Drug
Input Feature Vector
Level 1
Level 2
Level N
Forest
Ave.
Max
Final Prediction
Drug-Drug Interactions
Target-Target Interactions
Disease-Disease Interactions
Drug-Target Interactions
Disease-Target Interactions
Drug network
Target network
Disease network
Disease-Target Gaussian kernel similarity
Target-Target Gaussian kernel similarity
Drug-Target Gaussian kernel similarity
Drug-Drug Chemical similarity

FIGURE 2 | The overview of the proposed work THNCDF.

of training examples at the leaf node, and then average all trees in the same forest to obtain an estimate of the class distribution.

Secondly, we use three random forests: (a) two completely random tree forests, (b) two gradient boosting tree forests, and (c) two extra randomized tree forests. Each forest contains 1,000 trees, and there are 6,000 trees in total. Each node selects a feature randomly as the judgment condition and generates leaf nodes according to the condition. Stop until each leaf node contains only instances of the same class.

To compare with other results, we use 10-fold cross validation (Liu et al., 2016). It means that class vectors produced by each forest are generated by 10-fold cross validation to reduce the risk of overfitting. Finally, if there is no significant performance gain, the training process will terminate. The number of cascade levels is automatically determined.

EXPERIMENTAL RESULTS AND ANALYSIS

Baseline Methods

In order to evaluate the performance of our method, we mainly introduce DTI prediction results compared with baseline methods on the benchmark data sets that are proposed by Y. Yamanishi. The following are the state-of-the-art methods made in comparison with the same standard criteria:

RLS-KF (Hao et al., 2016): A regularized least squares combining with nonlinear kernel fusion method is developed.

RF (Cao et al., 2014): A computational method integrated the information from network, chemical, and biological properties. This method is developed based on the random forest combining with integrated features.

DTiGEMS (Thafar et al., 2020): A computational method using graph embedding, graph mining, and similarity properties techniques. DTiGEMS firstly applies a similarity selection procedure and a similarity fusion algorithm. Then, it integrates multiple drug–drug similarities and target–target similarities into the final heterogeneous graph structure after.

iDTI-ESBoost (Rayhan et al., 2017): A prediction model uses evolutionary and structural features. The method uses a new data balancing and boosting technique to make prediction.

Evaluation Criteria

Two quality measures are commonly used to evaluate the performance of these methods: AUC and AUPR. Specifically, we calculate the receiver operating characteristic curve (ROC) of true positive as a function of false positive, and use the area under the ROC curve (AUC) value as a quality measure. In addition, we also calculate the precision–recall curve (P–R), which is the chart of true positive rate between all positive predictions of each given recall rate. The area under the P–R curve (AUPR) provides a quantitative assessment. These two kinds of quality measures have become the standard criteria for evaluating methods.

Prediction Ability

To provide a fair comparison of DTI prediction performances, we apply these methods on the same benchmark data sets. We also use 10-fold cross-validation random setting, the same evaluation criteria, and optimal parameters of each method.

From the results reported in **Table 3** and **Figure 3**, THNCDF algorithm still maintains a high performance, especially for the AUPR values. For example, in the enzyme data set (**Figure 3A**), the ion channel data set (**Figure 3B**), and the GPCR data set (**Figure 3C**), THNCDF outperforms all other methods by achieving the best performance for AUPR values. On the other hand, for the AUC values, THNCDF still maintains the high performance. It is well known that the training of DNN usually requires a large amount of training data; hence, its implementation on tasks with small-scale data is not suitable. This is the inherently unavoidable characteristic of the method we use. Thus, it is reflected in the correlation between the size of the benchmark data sets and the AUC values obtained.

In addition, the number of positive samples and negative samples in each data set is highly imbalanced. The fact that few positive samples make THNCDF cannot exert its advantages, which is based on a large amount of training data. For benchmark data set, the feature dimension used in this method is low, and cascade deep forest has great advantages in the representation learning of ultra-high-dimensional data.

As shown in **Figure 3**, it is found that the prediction accuracy is approximately equal to each other. It also shows that the THNCDF preserves the best performance on all data sets so that it can be migrated to other predictions. The experiment procedure shows that THNCDF is not very sensitive to parameter settings. Therefore, it does not need large-scale parameter adjustment, especially the selection of the optimal combination of base classifiers. Comparing with DNN, THNCDF is more stable and easier.

It is worth mentioning that for the two commonly used evaluation metrics, more and more authors think that AUPR provides more informative assessment than AUC for highly imbalanced data sets. They argue in favor of AUPR values as a key standard of evaluating the performance for skewed data sets, especially the data sets with more negative samples than positive samples. In fact, all of the four subsets in the benchmark data

TABLE 3 | The results of the baseline methods and the THNCDF method.

Data sets	Methods	RLS-KF	RF	DTiGEMS	iDTI-ESBoost	THNCDF
Enzyme	AUC	0.990*	0.978	0.990	0.960	0.987
	AUPR	0.915	0.935	0.970	0.680	0.988*
Ion channel	AUC	0.987	0.924	0.990*	0.905	0.982
	AUPR	0.901	0.948	0.960	0.480	0.980*
GPCR	AUC	0.981	0.951	0.990*	0.932	0.937
	AUPR	0.806	0.896	0.860	0.480	0.938*
Nuclear receptor	AUC	0.987	0.987	0.990*	0.928	0.963
	AUPR	0.911*	0.847	0.880	0.790	0.906

*For each methods, * indicates the highest value.*

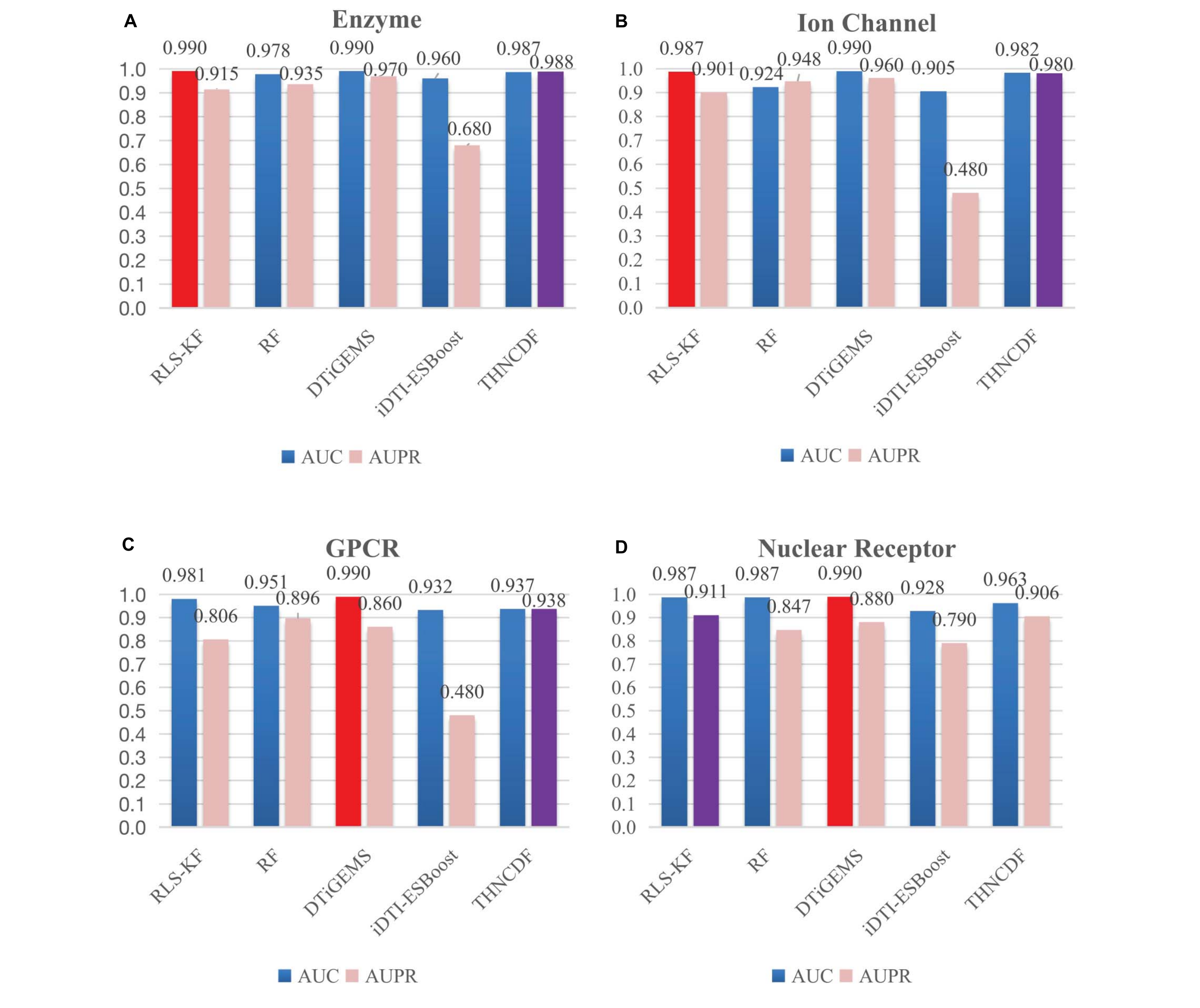


FIGURE 3 | Comparison results for THNCDF and other methods in terms of AUC and AUPR values on the benchmark data sets. The best AUC values are indicated in red, and the best AUPR values are in purple.

sets possess the imbalanced characteristic, which means that the number of known drug–target interactions is far less than the number of pairs with no interaction evidence. So a more sensitive AUPR metric is generally preferred for assessing the prediction results for those imbalanced datasets. From this perspective, the result clearly shows that THNCDF outperforms the prediction in terms of AUPR as well.

DISCUSSION

In this paper, we present a new multi-kernel computational approach combined with an improved cascade deep forest, which leads to good predictive performance on the task of predicting DTI. The values of AUPR on four benchmark data sets are improved to 0.988, 0.980, 0.938, and 0.906, respectively. Theoretically, THNCDF can process various high dimensional features by utilizing heterogeneous networks. However, we still have some problems to be solved in the future. First, even though studies have discussed multiple similarity calculation methods, they have not escaped the research scope on the network interactions. We are more looking forward to the introduction of new biochemical similarity calculation methods or data sets. Secondly, we suggest applying different embedding techniques, integrating more similarity measures from more sources, and generating more graph-based features. It can also be found that various data sets, such as chemical structure, side effect, therapeutic effect, gene expression, drug binding site, and semantic data, have been utilized in former studies. However, the disadvantages of these biomedical data sets are also obvious, which include high data noise, incompleteness, and inaccuracy. Thirdly, some potential extensions of our work include applying THNCDF to different networks formulated as an interaction prediction problem. Popular examples of interaction prediction in the bioinformatics field include but are not limited to drug–drug interactions prediction, drug–disease interactions prediction, and gene–disease association prediction.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YZ and ZW: conceptualization and validation. YZ: methodology, writing—review and editing, and supervision. ZW: software and writing—original draft preparation. Both authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China, grant number 61402054.

REFERENCES

Alabdulwahhab, K., Sami, W., Mehmood, T., Meo, S., Alasbali, T., and Alwadani, F. (2021). Automated detection of diabetic retinopathy using machine learning classifiers. *Riv. Eur. Sci. Med. Farmacol* 25, 583–590. doi: 10.26355/eurrev_202101_24615

Badkas, A., De Landtsheer, S., and Sauter, T. (2020). Topological network measures for drug repositioning. *Brief. Bioinform.* 1–13. doi: 10.1093/bib/bbaa357

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cao, D. S., Zhang, L. X., Tan, G. S., Xiang, Z., Zeng, W. B., Xu, Q. S., et al. (2014). Computational prediction of drugtarget interactions using chemical, biological, and network features. *Mol. Inform.* 33, 669–681. doi: 10.1002/minf.201400009

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y. P., et al. (2021). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1106–1112. doi: 10.1109/TCBB.2019.2936476

Ding, Y., Tang, J., and Guo, F. (2018). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Ding, Y., Tang, J., and Guo, F. (2020b). Identification of drug–target interactions via fuzzy bipartite local model. *Neural Comput. Applic.* 32, 10303–10319. doi: 10.1007/s00521-019-04569-z

Ding, Y., Tang, J., and Guo, F. (2020a). Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion. *Knowl. Based Syst.* 204:106254. doi: 10.1016/j.knosys.2020.106254

Fahimian, G., Zahiri, J., Arab, S. S., and Sajedi, R. (2020). RepCOOL: computational drug repositioning via integrating heterogeneous biological networks. *J. Transl. Med.* 18:375. doi: 10.1186/s12967-020-02541-3

Hao, M., Wang, Y. L., and Bryant, S. (2016). Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique. *Anal. Chim. Acta* 909, 41–50. doi: 10.1016/j.aca.2016.01.014

Huang, L., Luo, H. M., Li, S. N., Wu, F. X., and Wang, J. X. (2020). Drug–drug similarity measure and its applications. *Brief. Bioinform.* 1–20. doi: 10.1093/bib/bbaa265

Jarada, T., Rokne, J., and Alhajj, R. (2021). SNF-NN: computational method to predict drug-disease interactions using similarity network fusion and neural networks. *BMC Bioinformatics* 22:28. doi: 10.1186/s12859-020-03950-3

Juárez-Saldivar, A., Schroeder, M., Salentin, S., Haupt, V., Saavedra, E., Vázquez, C., et al. (2020). Computational drug repositioning for chagas disease using protein-ligand interaction profiling. *Int. J. Mol. Sci.* 21:4270. doi: 10.3390/ijms21124270

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). LDICDL: LncRNA-disease association identification based on Collaborative Deep Learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1–1. doi: 10.1109/TCBB.2020.3034910

Lan, W., Wang, J., Li, M., Liu, J., Li, Y., Wu, F., et al. (2016). Predicting drug-target interaction using positive-unlabeled learning. *Neurocomputing* 206, 50–57. doi: 10.1016/j.neucom.2016.03.080

Li, W. J., Xu, H. Y., Li, H. X., Yang, Y. J., Sharma, P., Wang, J., et al. (2019). Complexity and algorithms for superposed data uploading problem in networks with smart devices. *IEEE Intern.Things J.* 7, 5882–5891. doi: 10.1109/JIOT.2019.2949352

Liu, J., Wang, W. T., Chen, J., Sun, G. Z., and Yang, A. (2019). Classification and research of skin lesions based on machine learning. *Comput. Mater. Cont.* 61, 1187–1200. doi: 10.32604/cmc.2020.05883

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X. (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.* 12:e1004760. doi: 10.1371/journal.pcbi.1004760

Luo, H. M., Li, M., Yang, M. Y., Wu, F. X., Li, Y. H., and Wang, J. X. (2020). Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief. Bioinform.* 22, 1604–1619. doi: 10.1093/bib/bbz176

Meng, Y. J., Jin, M., Tang, X. F., and Xu, J. L. (2021). Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. *Appl. Soft Comput.* 103:107135. doi: 10.1016/j.asoc.2021.107135

Rauschenbach, L., Wieland, A., Reinartz, R., Kebir, S., Till, A., Darkwah Oppong, M., et al. (2020). Drug repositioning of antiretroviral ritonavir for combinatorial therapy in glioblastoma. *Eur. J. Cancer* 140, 130–139. doi: 10.1016/j.ejca.2020.09.017

Rayhan, F., Ahmed, S., Shatabda, S., Farid, D., Mousavian, Z., Dehzangi, I., et al. (2017). IDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. *Sci. Rep.* 7:17731. doi: 10.1038/s41598-017-18025-2

Roessler, H., Knoers, N., van haelst, M., and Haaften, G. (2021). Drug repurposing for rare diseases. *Trends Pharmacol. Sci.* 75, 157–160. doi: 10.1016/j.tips.2021.01.003

Stein, M., Levey, D., Cheng, Z. S., Wendt, F., Harrington, K., Pathak, G., et al. (2021). Genome-wide association analyses of post-traumatic stress disorder and its symptom subdomains in the Million Veteran Program. *Nat. Genet.* 53, 174–184. doi: 10.1038/s41588-020-00767-x

Thafar, M., Olayan, R., Ashoor, H., Albaradei, S., Bajic, V., Gao, X., et al. (2020). DTiGEMS+: drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *J. Cheminform.* 12:44. doi: 10.1186/s13321-020-00447-2

Turanli, B., Grotli, M., Boren, J., Nielsen, J., Uhlen, M., Arga, K., et al. (2018). Drug repositioning for effective prostate cancer treatment. *Front. Physiol.* 9:500. doi: 10.3389/fphys.2018.00500

Vivarelli, S., Candido, S., Caruso, G., Falzone, L., and Libra, M. (2020). Patient-derived tumor organoids for drug repositioning in cancer care: a promising approach in the era of tailored treatment. *Cancers.* 12:3636. doi: 10.3390/cancers12123636

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics (Oxf. Engl.)* 24, i232–i240. doi: 10.1093/bioinformatics/btn162

Yang, J. B., Zhang, D. N., Liu, L., Li, G. Q., Cai, Y. Y., Zhang, Y., et al. (2020). Computational drug repositioning based on the relationships between substructure–indication. *Brief. Bioinform.* 1–11. doi: 10.1093/bib/bbaa348

Zamami, Y., Imanishi, M., Takechi, K., and Ishizawa, K. (2017). Pharmacological approach for drug repositioning against cardiorenal diseases. *J. Med. Invest.* 64, 197–201. doi: 10.2152/jmi.64.197

Zeng, X. X., Zhu, S. Y., Liu, X. R., Zhou, Y. D., Nussinov, R., and Cheng, F. X. (2019). DeepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics (Oxf. Engl.)* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zhang, J. Y., Zhong, S. Q., Wang, J., Wang, L., Yang, Y. Q., Wei, B. Y., et al. (2020). "A review on blockchain-based systems and applications," in *Internet of Vehicles. Technologies and Services Toward Smart Cities. IOV 2019. Lecture Notes in Computer Science*, Vol. 11894, eds C. H. Hsu, S. Kallel, K. C. Lan, and Z. Zheng (Cham: Springer), 237–249. doi: 10.1007/978-3-030-38651-1_20

Zheng, Y., and Wu, Z. (2021). A machine learning-based biological drug-target interaction prediction method for a tripartite heterogeneous network. *ACS Omega* 6, 3037–3045. doi: 10.1021/acsomega.0c05377

Zhou, R. Y., Lu, Z. L., Luo, H. M., Xiang, J., Zeng, M., and Li, M. (2020). NEDD: a network embedding based method for predicting drug-disease associations. *BMC Bioinformatics* 21:387. doi: 10.1186/s12859-020-03682-4

Zhou, Z. H., and Feng, J. (2019). Deep forest. *Natl. Sci. Rev.* 6, 74–86. doi: 10.1093/nsr/nwy108

Zhu, D. J., Sun, Y. D., Li, X. F., Du, H. W., Qu, R. N., Yu, P. P., et al. (2020). MINE: a method of multi-interaction heterogeneous information network embedding. *Comput. Mater. Cont.* 63, 1343–1356. doi: 10.32604/cmc.2020.010008

Zhuang, X. H., Ye, R., So, M. T., Lam, W. Y., Karim, A., Yu, M., et al. (2020). A random forest-based framework for genotyping and accuracy assessment of copy number variations. *NAR Genomics Bioinform.* 2:lqaa071. doi: 10.1093/nargab/lqaa071

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 September 2021
doi: 10.3389/fgene.2021.652907

Graph Representation Forecasting of Patient's Medical Conditions: Toward a Digital Twin

Pietro Barbiero[*†], *Ramon Viñas Torné*[†] *and Pietro Lió*[†]

Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom

Objective: Modern medicine needs to shift from a wait and react, curative discipline to a preventative, interdisciplinary science aiming at providing personalized, systemic, and precise treatment plans to patients. To this purpose, we propose a "digital twin" of patients modeling the human body as a whole and providing a panoramic view over individuals' conditions.

Methods: We propose a general framework that composes advanced artificial intelligence (AI) approaches and integrates mathematical modeling in order to provide a panoramic view over current and future pathophysiological conditions. Our modular architecture is based on a graph neural network (GNN) forecasting clinically relevant endpoints (such as blood pressure) and a generative adversarial network (GAN) providing a proof of concept of transcriptomic integrability.

Results: We tested our digital twin model on two simulated clinical case studies combining information at organ, tissue, and cellular level. We provided a panoramic overview over current and future patient's conditions by monitoring and forecasting clinically relevant endpoints representing the evolution of patient's vital parameters using the GNN model. We showed how to use the GAN to generate multi-tissue expression data for blood and lung to find associations between cytokines conditioned on the expression of genes in the renin–angiotensin pathway. Our approach was to detect inflammatory cytokines, which are known to have effects on blood pressure and have previously been associated with SARS-CoV-2 infection (e.g., CXCR6, XCL1, and others).

Significance: The graph representation of a computational patient has potential to solve important technological challenges in integrating multiscale computational modeling with AI. We believe that this work represents a step forward toward next-generation devices for precision and predictive medicine.

Keywords: digital twin, generative adversarial networks, monitoring, graph representation learning, precision medicine

OPEN ACCESS

Edited by:
Wei Lan,
Guangxi University, China

Reviewed by:
Cheng Liang,
Shandong Normal University, China
Ruiqing Zheng,
Central South University, China

****Correspondence:***
Pietro Barbiero
pb737@cam.ac.uk

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *13 January 2021*
Accepted: *24 June 2021*
Published: *16 September 2021*

Citation:
Barbiero P, Viñas Torné R and Lió P (2021) Graph Representation Forecasting of Patient's Medical Conditions: Toward a Digital Twin. Front. Genet. 12:652907. doi: 10.3389/fgene.2021.652907

1. INTRODUCTION

Modern medicine is shifting from a wait and react, curative discipline to a preventative, interdisciplinary science aiming at providing personalized, systemic, and precise treatment plans to patients. Systems and network medicine are rapidly emerging in medical research providing new paradigms to address.

In the next decades, precision and predictive medicine will have a pivotal role in revolutionizing the healthcare system making it more flexible and efficient. Precision and predictive medicine are challenging research fields as they need to deal with the complexity of the human body (Ginsburg and Willard, 2009; Naylor and Chen, 2010). Precision requires integrating large amount of observations at individual and population levels simultaneously. These measures need to be taken at different scales, from genome to clinical and family history and at systemic levels, i.e., considering multiple tissues and organs. In the last years, systems and network medicine have introduced a variety of novel approaches with the aim of integrating and gaining knowledge on the human body. We have no capacity to integrate such disparate information into equation-based models but we can use machine learning and, in particular, deep learning methods to achieve this integration goal.

The primary objective of this work is exploring challenges and opportunities in modeling the human body as a whole, providing a panoramic view over individuals' conditions. To this aim, we propose a proof of concept of a "digital twin," i.e., a virtual prototype of patients mirroring the underlying biological system (Gelernter, 1993; Laubenbacher et al., 2021) combining information at organ, tissue, and cellular level. Existing prominent examples of digital twins in healthcare include "the artificial pancreas" (Brown et al., 2019; Kovatchev, 2019), pediatric cardiac digital twins (Gutierrez et al., 2019; Shang et al., 2019), and diabetes models (Eddy and Schlessinger, 2003). However, all these examples focus on just one single aspect of the human body due to its extreme complexity. As a result, they are not suitable to provide a holistic overview over the whole human body. We believe that recent graph representation approaches could overcome digital twin's limitations scaling across all the variety of body signals at different levels, making possible a revolution in healthcare. This work provides a first proof of concept providing the first elements for a novel class of machine-learning-assisted tools that scale to medical device deployment and run time monitoring and verification. By fusing ideas from systems medicine with scientific computing and machine learning, our software integrates and automates the analysis of vital parameters models under large uncertainty. A high degree of automation could transform how we use models in the scientific and medical discovery cycle and open up for a next-generation of powerful medical devices for probing the inner workings of full body in well-being and disease conditions.

The proposed architecture combines the qualities of generative and (Goodfellow et al., 2014) graph-based models (Scarselli et al., 2008) (see **Figure 1**). On the one hand, the generative model can be used to produce synthetic data under different biological states that might not be observed in reality. By augmenting the set of explorable states of the underlying biological system, the generative model may be employed for the simulation of extremely rare clinical scenarios representing precarious conditions, which might be difficult to analyze otherwise (Yi et al., 2019). In clinical contexts, this means that physicians will be able to set up personalized experiments in a virtual environment representing their patients in a very detailed and realistic way. On the other hand, the graph model represents the actual digital twin, providing a general and flexible framework to run probabilistic simulations. A panoramic view of individuals' conditions is provided by the final network configuration that combines information at organ, tissue, and cellular level. Cross-modal signals are also supported by the most recent graph learning frameworks, thus allowing the combination of different data sources, both structured and unstructured, real or simulated by generative methods. Finally, by relying upon flexible and modular architectures, our "digital twin" model can be conveniently deployed in dedicated hardware modules paving the way for a next-generation of medical devices.

2. DESIGN OF A BIOMEDICAL DIGITAL TWIN

The birth of the term "digital twin" could be the NASA's Apollo program where one spacecraft was launched into the outer space, while a "twin" spacecraft remained on earth to mirror flight conditions. Digital twin has been defined as "an integrated multiphysics, multiscale, probabilistic simulation of a vehicle or system that uses the best available physical models, sensor updates, fleet history, etc., to mirror the life of its flying twin" (Shafto et al., 2010; Grieves, 2015). The digita l twin is a virtual prototype; the analysis of its digital life cycle provides information to understand a product's functionality, manufacturing, behavior, and usage prior to building it. Here, the meaning of digital twin is slightly different: there is no product to be built, instead experimenting therapies on a digital twin will be cost-effective and will provide us with a rigorous testbed to conduct medical interventions. Within this framework, the artificial intelligence model could enable the prediction of disease trajectories before the insurgence of symptoms. The personal medical digital twin could also represent a pragmatic way for the cyber-physical fusion, as a new approach to support biomedical engineering design. In our vision, a composable AI architecture could enable the development of automatic analysis and verification techniques that are key to translational medicine.

Our digital twin consists of a modular AI-aided system that can be used to model the human body as a whole and to forecast the evolution of pathophysiological conditions (see **Figure 2**). The first module is based on a graph neural network (GNN) forecasting clinically relevant endpoints (such as blood pressure), while the second one is represented by a generative adversarial network (GAN) providing a proof of concept of multi-omic integrability.

2.1. The Effectiveness of GNNs and GANs in Biomedical Signal Analysis

The lack of interpretability of deep learning models has been one of the most significant barriers preventing their application in healthcare. Such models exhibit great capacity (Hornik, 1991) but understanding their behavior and following their decision-making process is not trivial (Castelvecchi, 2016). There is a growing body of literature focusing on interpretable artificial intelligence and interpretable deep learning aiming at developing white box models or at explaining black box ones (Das and Rad,

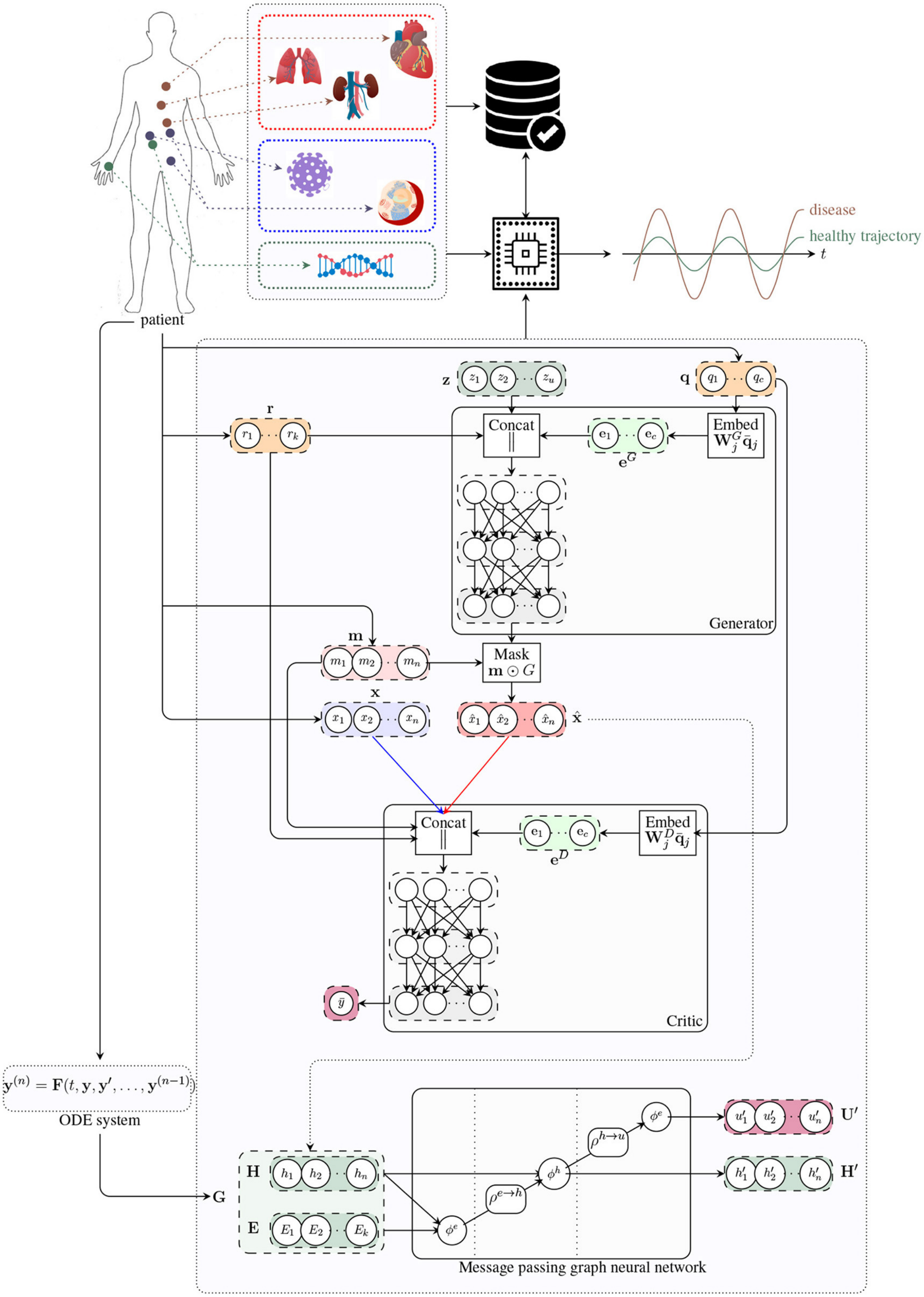

disease
healthy trajectory
t
patient
$\mathbf{z}$ z_1 z_2 z_u
$\mathbf{q}$ q_1 q_c
$\mathbf{r}$ r_1 r_k
Concat ||
e_1 e_c $\mathbf{e}^G$
Embed $\mathbf{W}_j^G\bar{\mathbf{q}}_j$
Generator
$\mathbf{m}$ m_1 m_2 m_n
Mask $\mathbf{m}\odot G$
$\mathbf{x}$ x_1 x_2 x_n
$\hat{x}_1$ $\hat{x}_2$ $\hat{x}_n$ $\hat{\mathbf{x}}$
Concat ||
e_1 e_c $\mathbf{e}^D$
Embed $\mathbf{W}_j^D\bar{\mathbf{q}}_j$
$\bar{y}$
Critic
$\mathbf{y}^{(n)} = \mathbf{F}(t, \mathbf{y}, \mathbf{y}', \ldots, \mathbf{y}^{(n-1)})$
ODE system
$\mathbf{G}$
$\mathbf{H}$ h_1 h_2 h_n
$\mathbf{E}$ E_1 E_2 E_k
ϕ^e $\rho^{e\to h}$ ϕ^h $\rho^{h\to u}$ ϕ^e
u'_1 u'_2 u'_n $\mathbf{U}'$
h'_1 h'_2 h'_n $\mathbf{H}'$
Message passing graph neural network

FIGURE 1 | Architecture of the digital twin model. The generator receives a noise vector z, and categorical (e.g. tissue type; q) and numerical (e.g. age; r) covariates, and outputs a vector of synthetic data (x̂). The critic receives data from two input streams (real, blue; and synthetic, red), a mask m indicating which components of the input vector are missing, and the numerical r and categorical q covariates. The critic produces an unbounded scalar $\bar{y}$ that quantifies the degree of realism of the input samples from the two input streams. The handcrafted ODE system proposed in Barbiero and Lió (2020) is used to determine a graph representation of patient's physiology. The message passing neural network updates latent node features to estimate global attributes describing the evolution of the underlying physiological system.

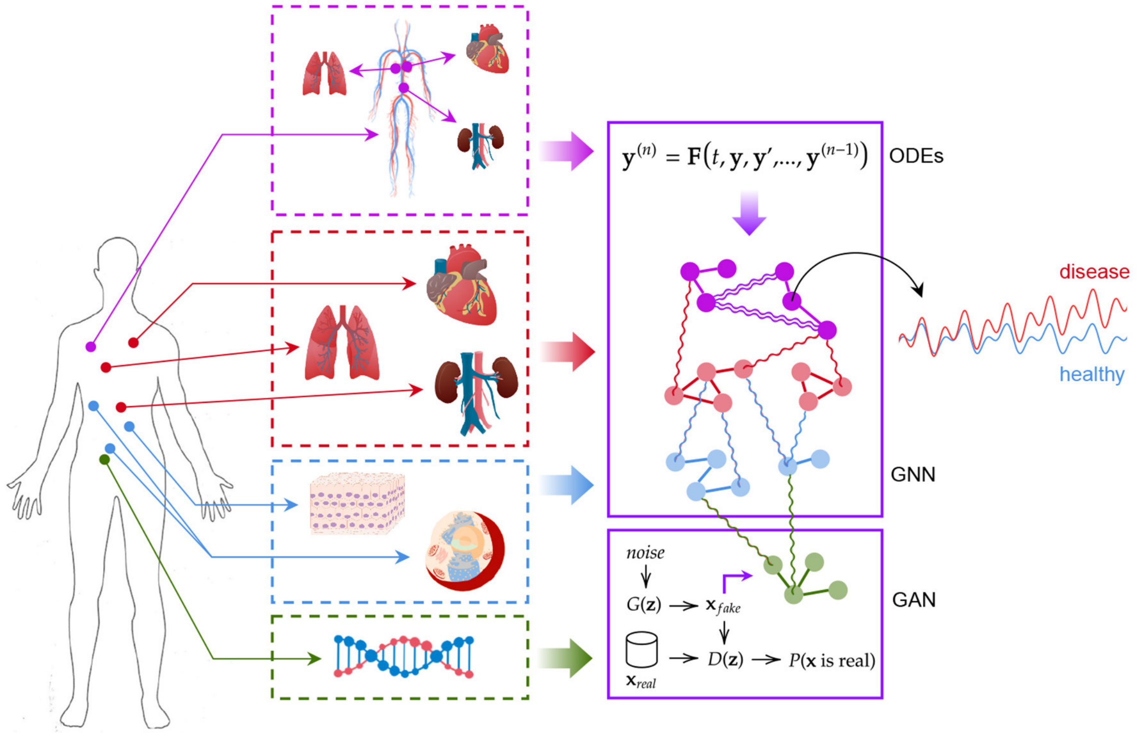

$\mathbf{y}^{(n)} = \mathbf{F}(t, \mathbf{y}, \mathbf{y}', \dots, \mathbf{y}^{(n-1)})$
ODEs
disease
healthy
GNN
noise
$G(\mathbf{z}) \to \mathbf{x}_{fake}$
GAN
$\mathbf{x}_{real} \to D(\mathbf{z}) \to P(\mathbf{x}$ is real$)$

FIGURE 2 | The digital twin model. Ordinary differential equations, graph neural networks, and generative adversarial networks are used synergically to model patient's conditions.

2020). Among such techniques, GNNs have started drawing the attention of both research and industry communities (Bronstein et al., 2017; Zhou et al., 2018). Such models are much more interpretable with respect to other neural approaches thanks to their graph structure, which is quite easy to understand from a human standpoint, and a few studies have already shown how graph networks can be effectively employed in biology and healthcare (Zitnik et al., 2018; Gysi et al., 2020).

Several properties of graph and generative adversarial neural networks make them suitable for medical data analysis. *(1) Non-linearity*: Both GNNs and GANs are able to detect non-linear patterns, which is of key interest as most systems are inherently non-linear in nature. Examples in medicine include heart rate dynamics, pulmonary functions, vascular structure, and gait dynamics. There is often a loss of non-linearity and multiscale fractal in aging and disease conditions (Goldberger et al., 2002). *(2) Interpretability*: Graph-based models are much easier to interpret with respect to other neural approaches thanks to their structure. The possibility of interpreting the behavior of models and the reason for their predictions is pivotal if not critical in many fields including clinical practice. *(3) Non-Euclidean geometry*: As a unique non-Euclidean data structure for machine learning, graphs can be used to model a variety of biological systems at different scales. Tissue and organ distributions could be modeled as graph models where each node or the graph contain time-dependent signals, similarly for pressure and electric sensors positioned at various parts of the body. Lymphatic vessels can also be modeled as a network where lymph nodes are vertices. At lower scale, cell arrangements in tissues form particular manifolds; proteins and genes are organized in regulatory networks; other examples are cytoskeleton and organelles (mitochondria networks). Additionally, diseases could be seen as nodes in a graph where edges represent comorbidity or underlying polygenic causes. *(4) Modularity*: A key property of GNNs is modularity, which allows to learn independent mechanisms that can be reused in several parts of the graph. Modularity facilitates scalability and allows to model dynamic properties of graphs. *(5) Cross-modality*: Both GNNs and GANs can learn how to combine structured and unstructured data sources, spanning different levels of biological complexity. This is particularly relevant when integrating signals at different levels of biological scale such as DNA methylation and functional magnetic resonance imaging (fMRI) data. *(6) Generative*: Both GNNs and GANs can learn how to generate new data preserving the statistical properties of the training set. This could be used to compare statistics at individual level with those at specific groups identified with stratification analysis or at general population levels. *(7) Multiscale*: The graph representation has the capability of integrating granular information organized as networks at different layers of biological complexity. This allows to recognize patterns in higher-order structures such as motifs, pathways, tissues (as compositions of cells), organs (as composition of tissues), processes and apparatus (as composition of organs), and stratification (as composition of individuals). *(8) Spectral* *density*: Together with spatial properties, GNN are amenable to frequency domain analysis. This allows to investigate network motifs, substructures, and periodical patterns at network levels.

2.2. Graph Neural Model

Graphs are mathematical structures that are used to model a set of objects (nodes) and their mutual relationships (edges) (Bollobás, 2013). Graphs are employed in a variety of research areas as they provide a general and flexible data structure for modeling real-world systems (Lieberman et al., 2005; Zhou et al., 2018; Rakocevic et al., 2019; Bica et al., 2020). GNNs are deep learning-based models working on the graph domain (Scarselli et al., 2008; Battaglia et al., 2018; Wu et al., 2020). Their properties have been recently drawn the attention of the artificial intelligence research community given their high interpretability (Lecue, 2019; Huang et al., 2020). The combination of graph theory and neural network elements have made GNNs one of the most promising tools to analyze complex systems in the graph domain. From neural networks, GNNs inherit a data-driven approach associated with a multi-layer architecture, which is the key to extract hierarchical patterns from data. However, unlike other deep-learning models, GNNs exploit additional features from graph theory and other mathematical disciplines. The main advantage with respect to other machine learning models relies in their extremely flexible and interpretable architecture. Once defined, the main endpoints of a system together with their mutual relationships directly induce a corresponding graph representation, which can be easily interpreted from a human standpoint. The abstract graph representation can be handcrafted, when the complexity of the underlying system allows it, or even automatically induced from data using generative approaches (Li et al., 2018). Hybrid techniques may also be explored taking advantage of generative algorithms for handling system complexity and human modeling to customize the most relevant endpoints. The design of GNNs is based on two basic principles, flexibility, and composability. GNNs support different graph structures as well as flexible representations of global, node, and edge attributes, customizable according to specific demands of tasks.

2.2.1. Stratification of Human Body Layers in a GNN

GNNs natively allow the design of complex systems using a modular approach. First, the complexity of the human body is broken up by developing independent subsystems representing genomic alterations, biological pathways, and organ physiology. Each subsystem can be represented as a different node or a network of nodes in a GNN, while inter-process signals can be reframed as message passing operations supporting multiscale ripple effects. Homogeneous subsystems can be aggregated into layers according to their characteristics. Our digital patient model is composed of four biological layers: the transcriptomic layer, the cellular layer, the organ layer, and the exposomic layer. Other layers can be easily implemented.

2.2.1.1. Transcriptomic Layer

The transcriptomic layer operates on the set of RNA transcripts produced by the genome at a particular time. Currently, RNA sequencing (RNA-seq) can measure RNA abundance across the entire genome with high resolution. The resulting high-throughput gene expression data can be used to uncover disease mechanisms (Emilsson et al., 2008; Cookson et al., 2009; Gamazon et al., 2018), propose novel drug targets (Evans and Relling, 2004; Sirota et al., 2011), provide a basis for comparative genomics (Colbran et al., 2019), and address a wide range of fundamental biological problems.

In this work, we study the crosstalk between tissues in the organ layer (see **Figure 1**) through the communicome, e.g., communication factors in blood (Ray et al., 2007). Specifically, we analyze to what extent the expression of genes involved in the renin–angiotensin system (RAS) can be explained by genes from signaling and receptor pathways, including the chemokine, TNF, and TGF-β pathways. We further develop a transcriptomics generative model based on a generative adversarial network (Goodfellow et al., 2014) and simulate the effects of SARS-CoV-2 infection by conditioning on high expression of ACE2 in the lung, kidney, and pancreas.

2.2.1.2. Cellular Layer

The cellular layer involves biological processes affecting individual cells from metabolism and protein synthesis to replication and motility. In this study, we focus on modeling the RAS, one of the main biological pathways regulating blood pressure and closely related to SARS-CoV-2 infectivity. Hence, it represents a suitable case study to demonstrate the flexibility and expressiveness of our GNN-based approach. The RAS is a hormone system regulating vasoconstriction and inflammatory response (Fountain and Lappin, 2019). The key hormone of the system is the peptide angiotensin II (ANG-II) generated from the decapeptide angiotensin I by the angiotensin-converting enzyme (ACE). ANG II promotes vasoconstriction, hypertension, inflammation, and fibrosis by activating the ANG-II type 1 receptor (AT1R) (Kuba et al., 2010; Gironacci et al., 2011). Glucose concentration, ACE inhibitor treatments, and viral infections binding to ACE2, such as SARS-CoV-2, can all have a significant impact on the RAS. A high glucose concentration may determine chronic hypertensive conditions. Reducing ANG II production with ACE inhibitors increases vasodilation and vasoprotection effects stimulated by the overproduction of AT2R and ANG-(1-7) (Zaman et al., 2002). Viral infections such as SARS-CoV-2 may also have an impact on RAS, as the virus binds to ACE2 in order to gain entry into the host cell. This results in an altered ACE2 activity and concentration, possibly leading to hypertension and inflammatory response (South et al., 2020).

2.2.1.3. Organ Layer

The organ layer comprises group of tissues with similar functions (organs) and complex networks of cooperating organs. Given the nature of the multi-factorial disease under study, we limited the organ layer to the circulatory system and a physiological representation of a few organs (Barbiero and Lió, 2020): heart, lungs, and kidneys. The heart model includes four compartments known as chambers (Neal and Bassingthwaighte, 2007). Deoxygenated blood collected from the superior and inferior venae cavae flows into the right atrium. When the right atrium contracts, the blood is pumped through the tricuspid valve into the right ventricle. From the right ventricle, the blood is pumped into the pulmonary trunk through the pulmonary valve flowing toward the lungs where carbon dioxide is exchanged for oxygen. The pulmonary circulation is composed of five vascular segments: proximal and distal pulmonary artery, small arteries, capillaries, and veins. Oxygenated blood collects into the left atrium via the pulmonary veins. From there, it flows into the left ventricle through the mitral valve and it is pumped into the aorta through the aortic valve for systemic circulation, providing oxygen and nutrients to body cells for metabolism in exchange for carbon dioxide and waste products. The mean arterial blood pressure is controlled by baroreceptors, special sensory neurons excited by a stretch in the carotid sinus and aortic arch vessels. They relay sensory information regarding blood pressure changes to the central nervous system where it is processed and utilized primarily in autonomic reflexes, regulating short-term blood pressure.

2.2.1.4. Exposomic Layer

The exposome refers to the totality of exposure individuals experience from conception until death and its impact on chronic and acute diseases (Wild, 2005). Toxicants, dietary regimens, treatments, physical exercise, posture, and lifestyle habits are possible exposures taking part to individual's well-being or disease condition. All such environmental factors are deeply coupled among themselves but also with individuals influencing the effects of new or present exposures. The exposome is intrinsically co-dependent on a person's genetics, epigenetics, health status, and physiology. For instance, regular exposure to pollution may lead to the outbreak of a lung carcinoma, which in turn may call for clinical intervention. In this work, we consider four types of exposures: dietary habits, physical activity, therapeutic treatments, and viral infections.

2.2.2. Inter-process Signals and Clinical Endpoints

One of the main advantages of using GNN-based models relies in that inter-process and multiscale communications can be natively implemented using message passing. In a GNN, each biological entity can be represented as a node, while the relationship between two entities can be modeled using directional edges. Signals exchanged between nodes are implemented using message functions ϕ^h (see Equation 1), which are used to update the hidden states of nodes. Such state transition will then have an impact on messages exchanged at the following time steps. Another strength of GNN models consists of the possibility of supervising the evolution of the underlying system by using the readout functions ϕ^u. Hence, the endpoints of multi-factorial diseases can be directly controlled by checking the output of readout functions in critical nodes. The resulting GNN model will combine a simple and modular design with a versatile structure accommodating for complex multiscale systems where clinical endpoints can be easily monitored and forecast in real time.

2.3. Generative Adversarial Model

One way of studying probability distributions is by means of generative models, which describe the random phenomenon in terms of the joint probability distribution of observed and target variables (Jebara, 2012). Generative adversarial networks (GANs) are a framework for estimating generative models via an adversarial process (Goodfellow et al., 2014). They are often described as a two-player game in which both players are encouraged to improve. One player, the *generator*, creates samples that are intended to be indistinguishable from the ones coming from a given data distribution. The other player, the *discriminator*, learns to determine whether samples come from the *fake* distribution (*fake* samples) or the *real* data distribution (*real* samples). **Figure 3** shows the basic idea of generative adversarial networks. With respect to other generative models, they provide a general and flexible framework for the analysis of joint probability distributions. The architecture itself allows a fine control of the data generation process and a high level of customization, making them suitable for a variety of experimental scenarios.

2.3.1. Crosstalk Between Tissue Types

The activity of biological systems is determined by internal factors, determined by intrinsic and functional properties, and by external factors shaping the interconnections between different systems. Chemical and molecular events, like oxygenation or protein phosphorylation, are often the vehicles of biological signals' transduction. A chain of biochemical events forms a signaling pathway whose activation may give rise to a biochemical cascade of events affecting the organism at different levels. In complex organisms, several signal transduction pathways communicate and react reciprocally generating biological crosstalks. Crosstalks have been widely characterized and observed in a variety of biological processes from micro- to macroscale from genomics (Poyton and McEwen, 1996; Du et al., 2015), to internal and external cell activity (Geiger et al., 2001; Li et al., 2016), and even between tissues (Lengyel et al., 2018). In particular, receptors and signaling factors from the chemokine, TNF, and TGF$-\beta$ pathways are known to take an active role in tissue communication as well as inflammatory-associated diseases (e.g., cardiovascular diseases affecting that the heart and the stiffness of blood vessels). Here, we develop a generative model based on a generative adversarial network to produce synthetic transcriptomics data describing the ripple effects of a viral infection on crosstalks between different tissues. The aim is to demonstrate how generative approaches can be used both to reproduce and enhance the set of observable states of a patient allowing for a deeper understanding of complex biological processes.

3. RESULTS

3.1. Clinical Case Studies

In Barbiero and Lió (2020), the authors proposed a computational tool for running simulations integrating a variety of mechanistic and phenomenological models describing

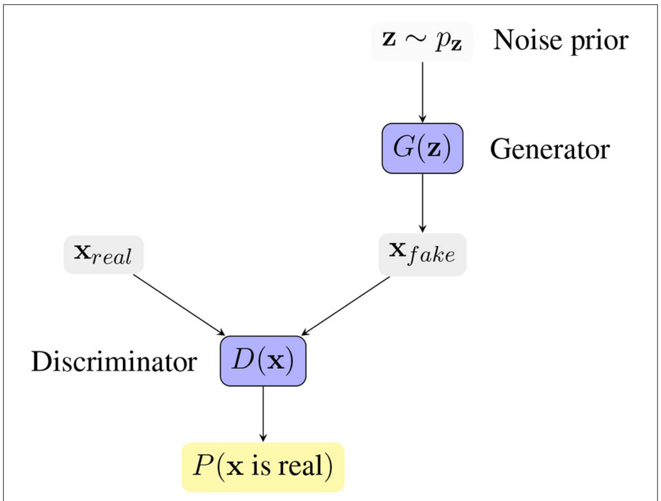


FIGURE 3 | Generative Adversarial Network framework. The generator *G*(z) receives a vector z sampled from a noise prior distribution p_z, and generates a synthetic sample $x_f ake$. The discriminator *D*(x) tries to distinguish *real* samples from *fake* samples, producing the probability of x coming from the *real* data distribution. The competition between the two players drives the game and makes both players increasingly better.

the human body with ordinary differential equations (ODEs). This computational framework is hereby used to generate two clinical case studies. The main difference of the proposed approach with respect to the computational tool proposed in Barbiero and Lió (2020) consists of a different modeling approach based on state-of-the-art AI models instead of ODEs.

The first scenario consists of an elderly patient experiencing hypertension and type 2 diabetes with diabetic nephropathy. Her lifestyle is mainly sedentary and her diet is rich in carbohydrates. The patient needs a therapeutic plan for the treatment of her hypertension. The task for the clinician is to personalize the therapy assigning a proper daily dosage of benazepril. This case study is used to show how the digital patient model can be employed to simulate the evolution over time of clinical endpoints under a set of possible therapeutic plans and to choose the best option.

In the second scenario, the same patient is seeking medical help for a mild flu caused by a SARS-CoV infection. For this case study, the model can be used to constantly monitor and forecast clinical endpoints to prevent complications threatening patient's life. The decreased oxygenation caused by flu may have detrimental effects on both heart and brain activities indeed. Studies have reported that SARS-CoV infections can activate the blood clotting pathway by impairing left heart pumping performance, which results in a blood back up in the lungs and in a increased blood pressure. High blood pressure can reduce blood vessel's compliance decreasing blood and oxygen flows and leading to a higher risk of developing systemic conditions. For this reason, heparin-based therapies have been recommended to prevent clot formation or tissue plasminogen activator (tPA) (Sardu et al., 2020; Tang et al., 2020). Although some variation in blood pressure throughout the day is normal, a high blood pressure variability is associated with a higher risk of cardiovascular disease (O'Rourke and Nichols, 2005; Mitchell et al., 2010; Wen et al., 2015; Clark et al., 2019; Bangalore et al., 2020) and all-cause mortality (Tao et al., 2017; Kim et al., 2018). Clogged arteries, fibrosis, and strokes caused by blood pressure spikes are among the main complications threatening patient's life and calling for the foremost necessity for treatment. Hence, blood pressure is one of the most relevant clinical endpoints that need to be constantly monitored in real time and accurately forecast.

3.2. Forecasting Clinical Scenarios

3.2.1. Dataset

Our digital twin model is hereby used to actively monitor and forecast the endpoints highlighted in the two clinical case studies. First, the computational system described in Barbiero and Lió (2020) based on ordinary differential equations (ODEs) is used to generate a time series of clinical endpoints for each differential equation with a window size of $\tau = 500$ time steps (Barbiero and Lio, 2020). Time series are collected, randomly shuffled, and stacked in a dataset. Each item of the collection is randomly assigned either to a training ($n_{train} = 3,200$), validation ($n_{val} = 800$), or test set ($n_{test} = 1,000$).

3.2.2. Training

The graph model is derived from the structure of the ODE system, thus leveraging human knowledge (an example is shown in **Figure 4**). Nodes correspond to variables represented by the differential equations in Barbiero and Lió (2020) while edges follow the underlying relationships. In a GNN-based model, each node learns a latent representation of the state using the messages received from its neighborhood. Hence, the rigid mathematical structure of the ODE system is relaxed in our model as such structure can be learned directly from data. The learning process lasts for $\eta = 50$ epochs with a learning rate of $\epsilon = 0.01$. Once trained and validated, the model is used to generate a bundle of possible trajectories for the elements of the test set. As a result, the model estimates a 95% confidence interval of the evolution of each variable over time.

3.2.3. Results

Providing a complete overview of the clinical state of a patient is not trivial. Focusing just on one endpoint might be misleading. On the contrary, a global vision comprising pathophysiological conditions is required in order to provide a clear and effective overview where organs and physiological systems can be monitored as a whole. One of the most effective approaches consists of applying a dimensionality reduction technique (Van Der Maaten et al., 2009) condensing the information of each organ and projecting forecasts in a lower-dimensional space.

Figure 5 shows an overview of the clinical state of the heart in a two-dimensional projected phase space. For each clinical case study, a GNN-based model is used to simulate a therapeutic intervention and its impact on blood pressure in heart chambers (right and left atrium and ventricle). In order to provide an overview of heart conditions, we

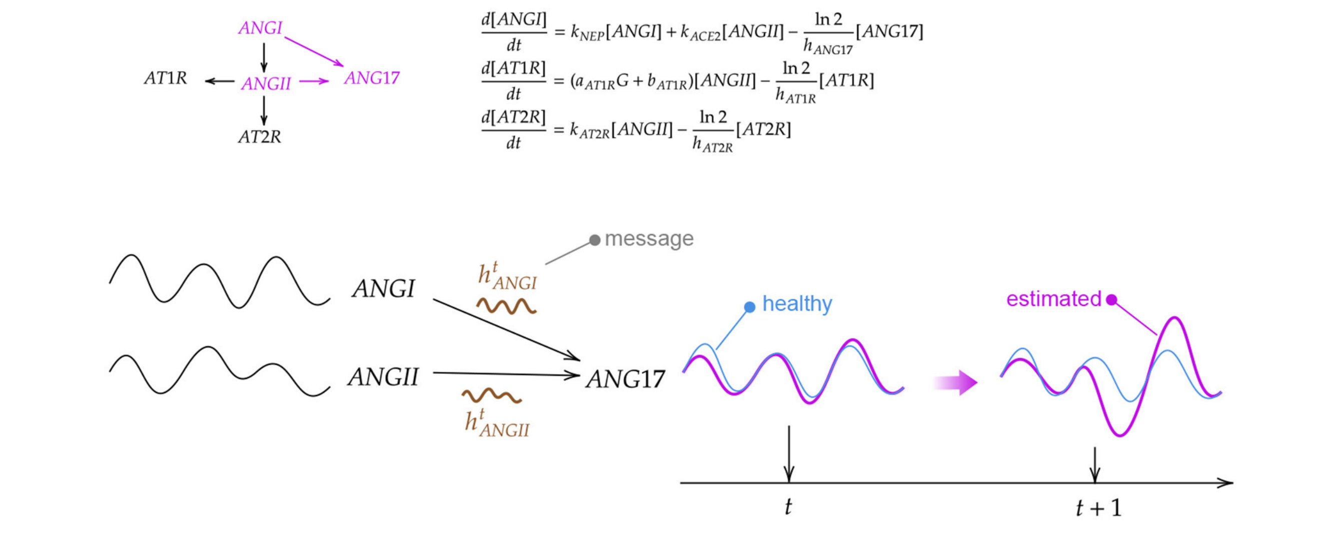

ANGI
AT1R
ANGII
ANG17
AT2R
$\frac{d[ANGI]}{dt} = k_{NEP}[ANGI] + k_{ACE2}[ANGII] - \frac{\ln 2}{h_{ANG17}}[ANG17]$
$\frac{d[AT1R]}{dt} = (a_{AT1R}G + b_{AT1R})[ANGII] - \frac{\ln 2}{h_{AT1R}}[AT1R]$
$\frac{d[AT2R]}{dt} = k_{AT2R}[ANGII] - \frac{\ln 2}{h_{AT2R}}[AT2R]$
message
h^t_{ANGI}
ANGI
ANGII
ANG17
h^t_{ANGII}
healthy
estimated
t
$t+1$

FIGURE 4 | Example of how a biological system can be modeled in a graph neural network through differential equations. First, an ordinary differential equation (ODE) system is derived from a biochemical reaction network. Then, the ODE system is solved for different initial conditions generating a set of trajectories for each variable. Finally, a graph neural network aggregates the information coming from neighbor nodes to update the current state of the variable.

projected the predicted trajectories using principle component analysis (PCA) (Pearson, 1901). The interpretation of both pictures is straightforward. The first one shows the effect of a therapeutic intervention comprising an increased physical exercise, a reduced amount of calorie intake, and the subscription of a daily dosage of benazepril (5 mg). The predicted result of the prescription (green density reporting the 95% CI of the trajectories) reveals an overall reduction of blood pressure mean and variability in heart chambers. This results in a reduced risk of developing severe cardiovascular conditions with detrimental ripple effects for the whole system.

The second figure reports the simulation corresponding to the second case study. The same patient is seeking medical help to treat the first symptoms of a SARS-CoV-2 infection. The first simulation (red density) shows the long-term impact on heart blood pressure of an untreated viral infection. In this case, blood pressure spikes may cause irreparable damages to blood vessel walls, reducing their compliance, and impairing their capacity for adaptation to different environmental conditions. A synergic therapy including both benazepril (5 mg/day) and intravenous injection of heparin (5,000 U/ml) may have a beneficial effect on blood pressure mean and variability (orange density). On the one hand, benazapril lowers blood pressure by inhibiting ACE activity in cleaving ANG-I and producing ANG-II, which is the key RAS regulator of blood pressure. On the other hand, heparin is used to prevent and dissolve blood clots (Sardu et al., 2020; Tang et al., 2020). The treatment has an indirect impact on blood pressure by making blood less dense, reducing clotting formation, and lowering inflammation.

A lower-dimensional representation of an organ or system as a whole could be interesting to get a rapid and clear overview of the long-term impact of a disease or a therapeutic intervention. Nonetheless, bundle of predicted trajectories can be visualized and monitored individually in real time when needed in order to investigate patterns in the time domain. **Figure 5** shows an example where blood pressure trajectories in heart chambers are predicted in real time starting from a healthy state condition (green density). In some cases, this representation in the time domain might be closer to common clinical approaches, thus providing a more conventional visualization tool for monitoring clinical endpoints in real time.

3.3. Transcriptomics Analysis of the Crosstalk Between Tissue Types

We hypothesize that the communication factors in blood might be playing an important role in the development of the SARS-CoV-2 infection by facilitating the spread of the virus in the human body. Here, we study whether the expression of genes involved in the RAS can be explained by genes that take part of the communicome in blood. This analysis might shed light on whether it is sensible to model the crosstalk between tissue types with a GNN where tissue nodes communicate with each other through whole blood.

3.3.1. Dataset

We leverage data from the Genotype-Tissue Expression (GTEx) project (v8), a resource that has generated a comprehensive collection of human transcriptome data in a diverse set of tissues (Aguet et al., 2019). The dataset contains 15,201 RNA-Seq samples collected from 49 tissues of 838 unique donors. We select genes based on expression thresholds of $\geq$ 0.1 TPM in $\geq$ 20% of samples and $\geq$ 6 reads in $\geq$ 20% of samples. We normalize the read counts between samples using the trimmed mean of M-values (TMM) normalization method (Robinson and Oshlack, 2010) and we inverse normal transform the expression values for each gene. From all the donors, we select those that have gene expression measurements for whole blood, yielding

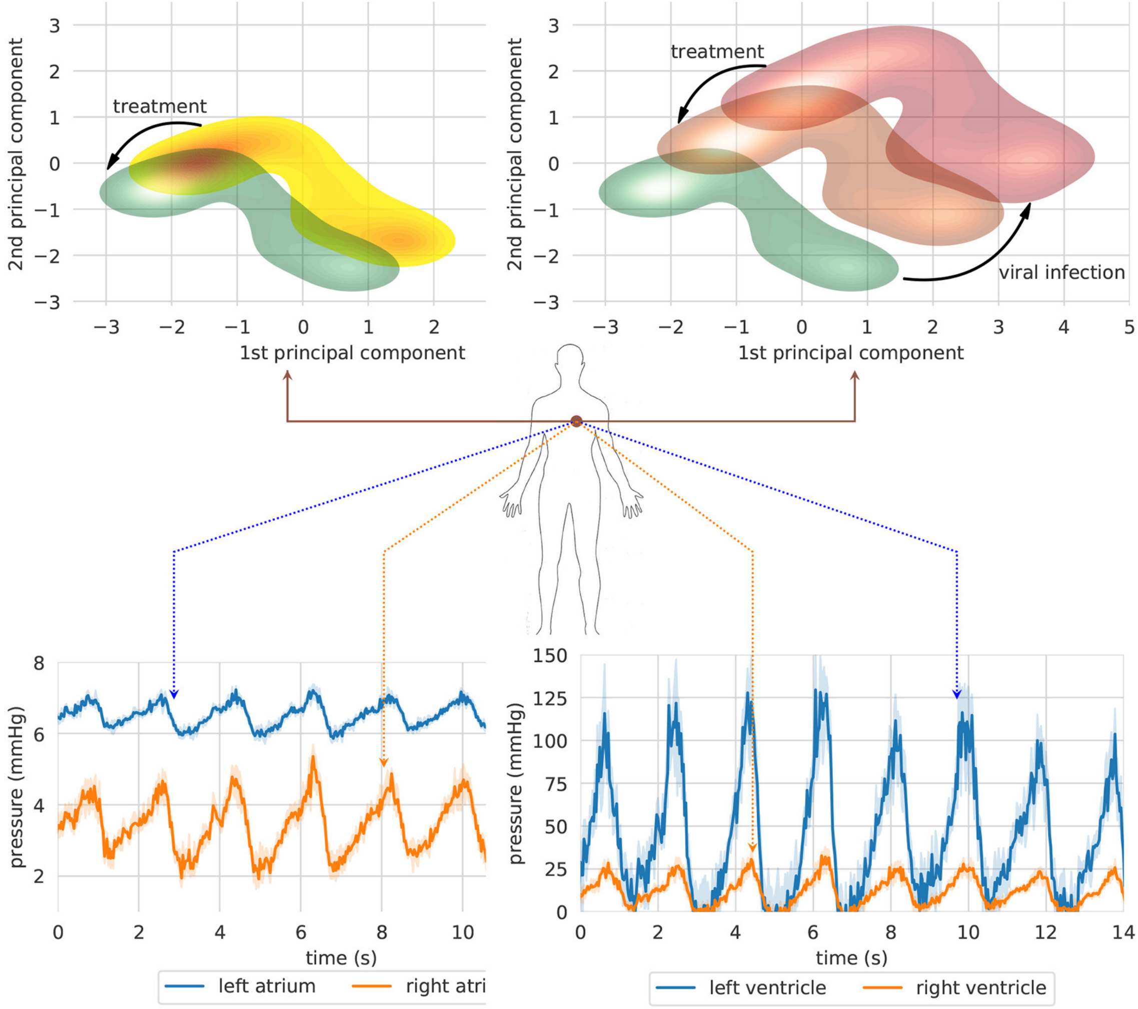


FIGURE 5 | Two clinical case studies represented in a projected heart phase-space. The first case study (left) shows the effect of a therapeutic intervention comprising an increased physical exercise, a reduced amount of calorie intake, and the subscription of a daily dosage of Benazepril (5 mg). The second simulation (right) shows the long-term impact on blood pressure of an untreated SARS-CoV-2 infection (red density) and the effects of a therapy including both Benazepril (5 mg/day) and intra venous injection of heparin (5000 µ/ml) (orange density). (Top) Bundle of predicted trajectories can be visualized and monitored in real time in order to investigate patterns in the time domain. The simulation shows blood pressure in heart chambers starting from healthy state conditions. Error bands represent 95% CI (Bottom).

670 unique individuals. We then match the patients' whole blood samples with the corresponding measurements in lung (418), cortex of kidney (62), pancreas (257), and left ventricle of heart (324). Finally, we use the KEGG pathway database (Kanehisa et al., 2010) to select genes from the RAS (*hsa04614*), chemokine (*hsa04062*), TNF (*hsa04668*), and TGF-β (*hsa04350*) pathways.

3.3.2. Results

Figure 6 shows the bootstrapped R^2 scores for each gene in the RAS pathway in different tissue types. To compute the bootstrapped scores, we sampled donors with replacement (sample size: 75% of the total observations), trained the ridge regression model (Equation 5.3) on the sampled data, and evaluated the performance on the remaining out of bag (OOB) observations. Appendix 3 in **Supplementary Material** shows the held-out performances for different regularization strengths. We repeated this process 1,000 times to obtain a distribution of R^2 scores for each gene. Our results show that the expression of some genes in the ACE2 pathway can be partially explained by signaling genes from whole blood. Notably, the associations for the kidney (cortex) are weaker or non-existent, potentially because the data are limited for this tissue (62 samples) or because the biological associations are indeed small. Overall, these results suggest that signaling pathways such as TNF, TGF-β, and chemokine might be playing an important role in the development of the SARS-CoV-2 infection.

We next model the expression of cytokines and receptors from whole blood (TNF, TGF-β, and chemokine pathways) as a function of cytokines from other tissue types (lung, kidney, heart, and pancreas) (Lijnen et al., 2003; Elmarakby et al., 2007; Rudemiller and Crowley, 2017). **Figure 7** shows the bootstrapped R^2 scores for the top 20 cytokines (chemokine pathway) for each

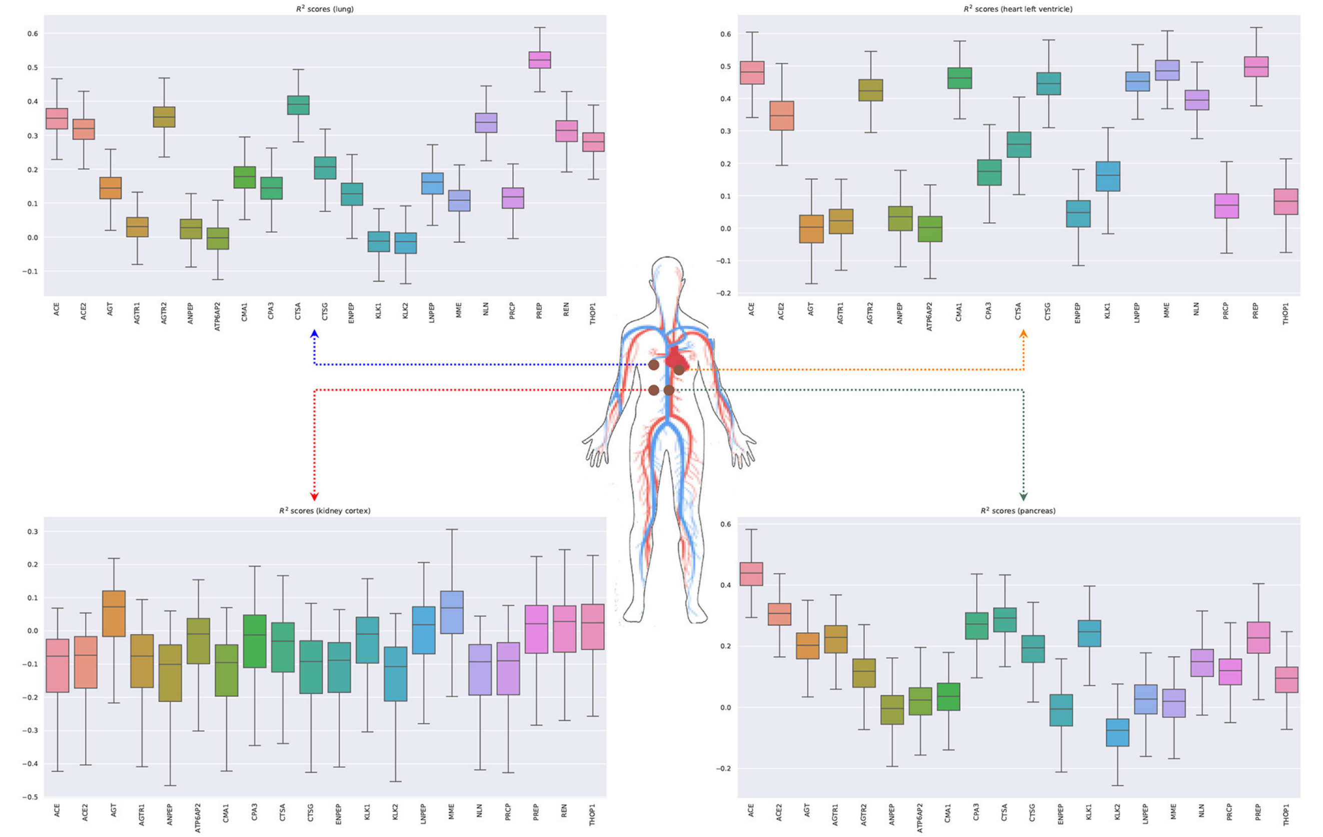


FIGURE 6 | Bootstrapped R^2 scores for genes involved in the renin-angiotensin system for lung, heart (left ventricle), kidney (cortex), and pancreas. The input variables are the expressions of genes in whole blood belonging to the chemokine, TNF, and TGF-β pathways.

target tissue type. These results illustrate the associations between cytokines in blood and other tissue types, which facilitate tissue communication and crosstalks.

3.4. Generative Model for Transcriptomics Data

The generative model is here used to produce synthetic transcriptomics data. By conditioning on high expression of ACE2 in the lung, kidney, and pancreas, we aim to simulate the effects of SARS-CoV-2 infection in the expression of genes involved in communicome and signaling pathways such as TNF, TGF-β, and chemokines. These pathways are implicated in many physiological and pathological processes including the regulation of blood pressure and inflammatory processes, and have been hypothesized to play a central role in SARS-CoV-2 infection (Garvin et al., 2020). For this analysis, we use data from the GTEx project previously described. In Appendix 2 in **Supplementary Material**, we analyze the held-out performance for different architectures of the generator and critic and describe all the training details.

Real datasets often lack transcriptomic measurements that account for multiple tissue types jointly. For example, out of 838 GTEx donors, only 257 of them present joint observations for pancreas and whole blood (**Figure 8** shows the distribution of missing tissues per patient). Importantly, our model allows to sample gene expression data for synthetic patients in every modeled tissue type and without any missing values, facilitating the cross-tissue analysis of gene expression.

3.4.1. Results

Figure 9 shows that the pairwise correlations between genes in the ACE2 pathway (lung) are well-preserved in the synthetic data. We observe that some genes in the RAS pathway (CTSA, AGTR2, NLN, and PREP) that can be relatively well-explained as a function of blood signaling factors (see **Figure 6**) are simultaneously correlated with ACE2. This suggests that these genes could be playing an important role in the spread of SARS-CoV-2 in our body through blood. Next, we use the GAN to generate multi-tissue expression data for blood and lung, and fit a linear model to predict the expression of 170 chemokines in blood as a function of the expression of 21 genes in the renin–angiotensin pathway from lung. **Figure 10** shows the R^2 scores for the top 20 chemokines. We find that some of the top predicted chemokines (e.g., CXCR6 and XCL1) have previously been associated with SARS-CoV-2 infection (Kusnadi et al., 2020; Liao et al., 2020). Additionally, our GAN captures associations between inflammatory cytokines, which are

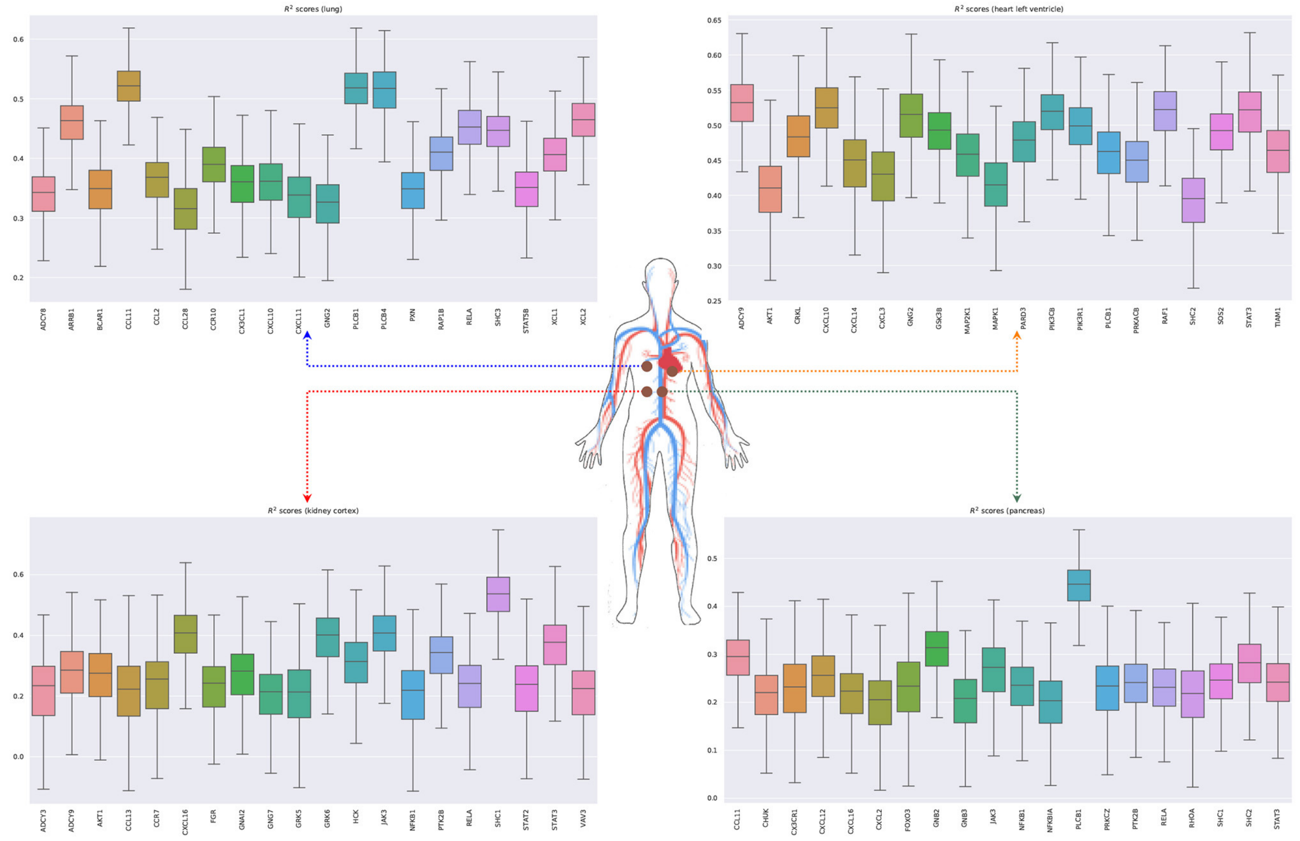


FIGURE 7 | Bootstrapped R^2 scores for several cytokines and receptors for lung, heart (left ventricle), kidney (cortex), and pancreas. For each tissue type, we show the top 20 predicted cytokines. The input variables are the expressions of genes in whole blood belonging to the chemokine, TNF, and TGF-β pathways.

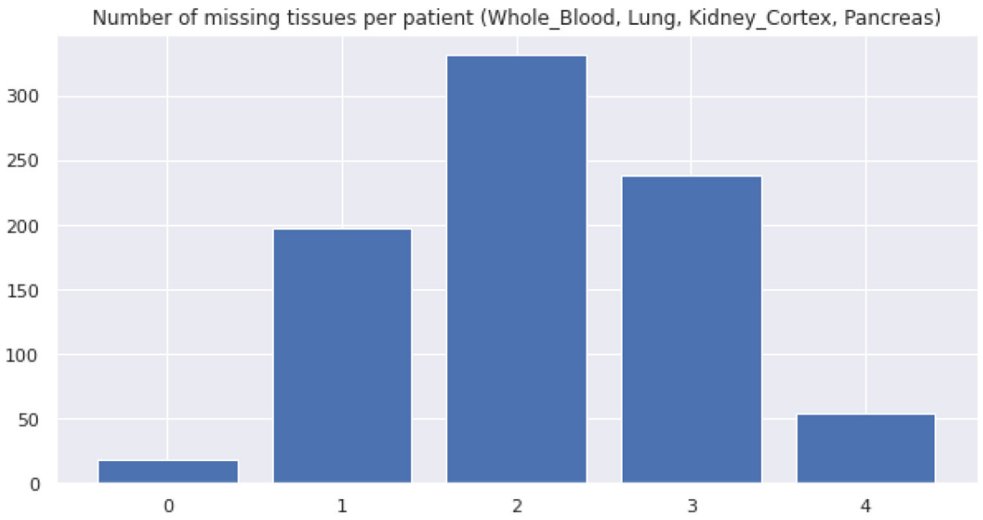


FIGURE 8 | Distribution of missing tissues per GTEx patient. This plot only considers 4 tissue types (whole blood, lung, kidney (cortex), and pancreas).

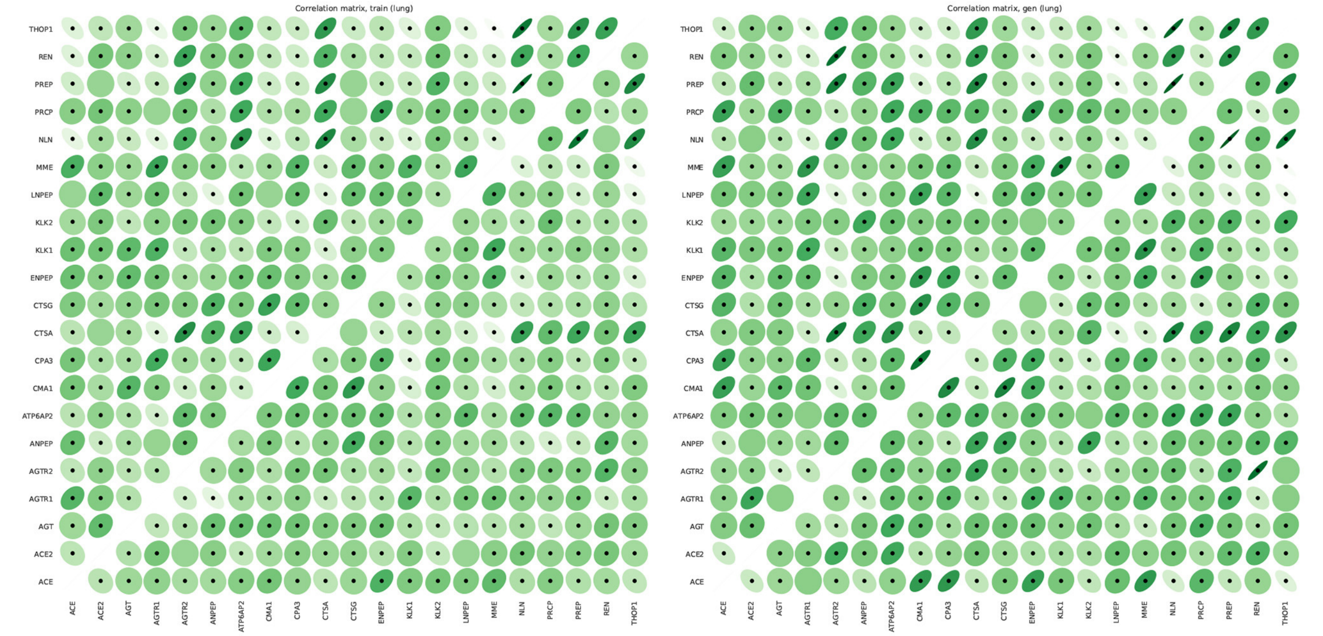


FIGURE 9 | Pairwise Pearson correlations between genes in the renin-angiotensin system pathway in lung for real **(left)** and synthetic **(right)** data. The correlations in the lower and upper matrices are computed from samples with low (61 samples) and high (60 samples) ACE2 expression, respectively. We use dots to label statistically significant correlations (two-sided p-value < 0.05).

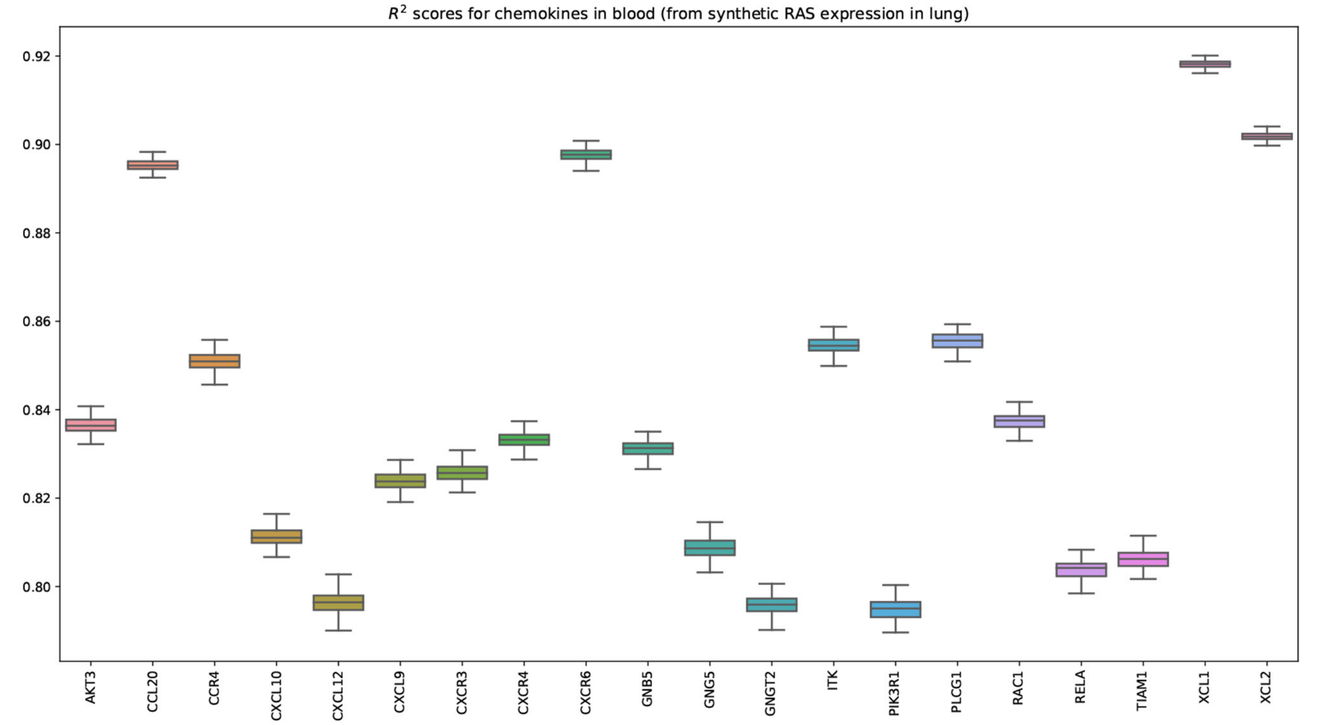


FIGURE 10 | Bootstrapped R^2 scores for chemokines in blood. The input variables are the expressions of 21 genes belonging to the renin-angiotensin system pathway in lung. This plot shows the top 20 predicted chemokines (out of 170). The transcriptomics data was generated by our GAN. Importantly, some of the top predicted chemokines (e.g., CXCR6) have been previously associated with SARS-CoV-2 (Liao et al., 2020).

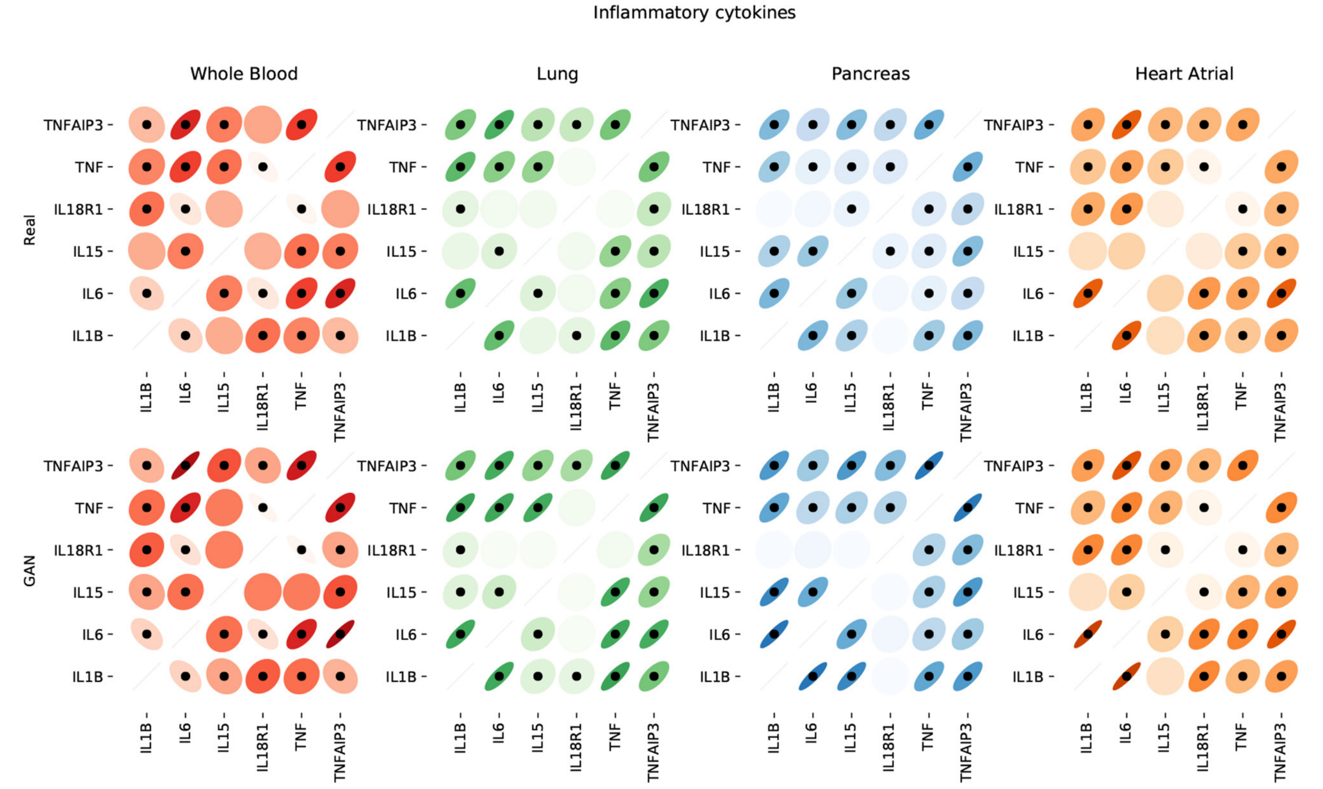


FIGURE 11 | Pairwise correlations between inflammatory cytokines in the 4 modeled tissue types. We use dots to label statistically significant correlations (two-sided *p*-value < 0.05).

known to have effects on blood pressure (Groth et al., 2014). **Figure 11** illustrates the real and synthetic pairwise correlations for 6 inflammatory cytokines in the 4 modeled tissue types. Finally, **Figure 12** shows that it is also possible to sample data for synthetic patients conditioned on different levels of ACE2 expression in lung.

4. DISCUSSION

In this work, we presented an interpretable digital twin model providing an holistic view over patients' conditions. We tested our proof of concept on two clinical case studies combining information at organ, tissue, and cellular level showing the potential of our framework in clinical practice. We demonstrate the feasibility of representing and integrating physiological models and molecular information using GNNs and generative adversarial networks. This composite approach provides modularity and scalability across layers of biomedical data, it is amenable of a battery of modeling approaches, and generates integrated predictions that translate into patients trajectories. We have assimilated our product to a digital twin of the patient.

4.1. Technological Perspectives

4.1.1. Digital Twin Deployment

Mechanistic computational modeling and machine learning should be considered together when building innovative healthcare solutions. Building a puzzle is often an example of participatory activity. Clinicians, mechanistic computational modeling and machine learning researchers, data policy makers, and public and private sectors could build a puzzle (i.e., the healthcare) together and they should first develop a shared vision about what is the puzzle. Our vision is to consider a co-simulation (say doctor checkup visits vs. computational experiments) of the two twins to allow co-verification. From a theoretical computer science perspective, this could open the direction of an interplay between AI and verification/synthesis and the use of reachability analysis to identify constraints over the well-being and disease system state space. Although different architectures seem suitable (e.g., only GNNs, only GANs, VAEs, etc.), our design has important advantages: the GNN could provide a physical mapping of the human body (in the same way a tube map or bus route is a map of a city); GANs could be specialized on processing molecular information or they could operate cross-modal operations such as omic–omic, omic–clinical, and clinical–clinical.

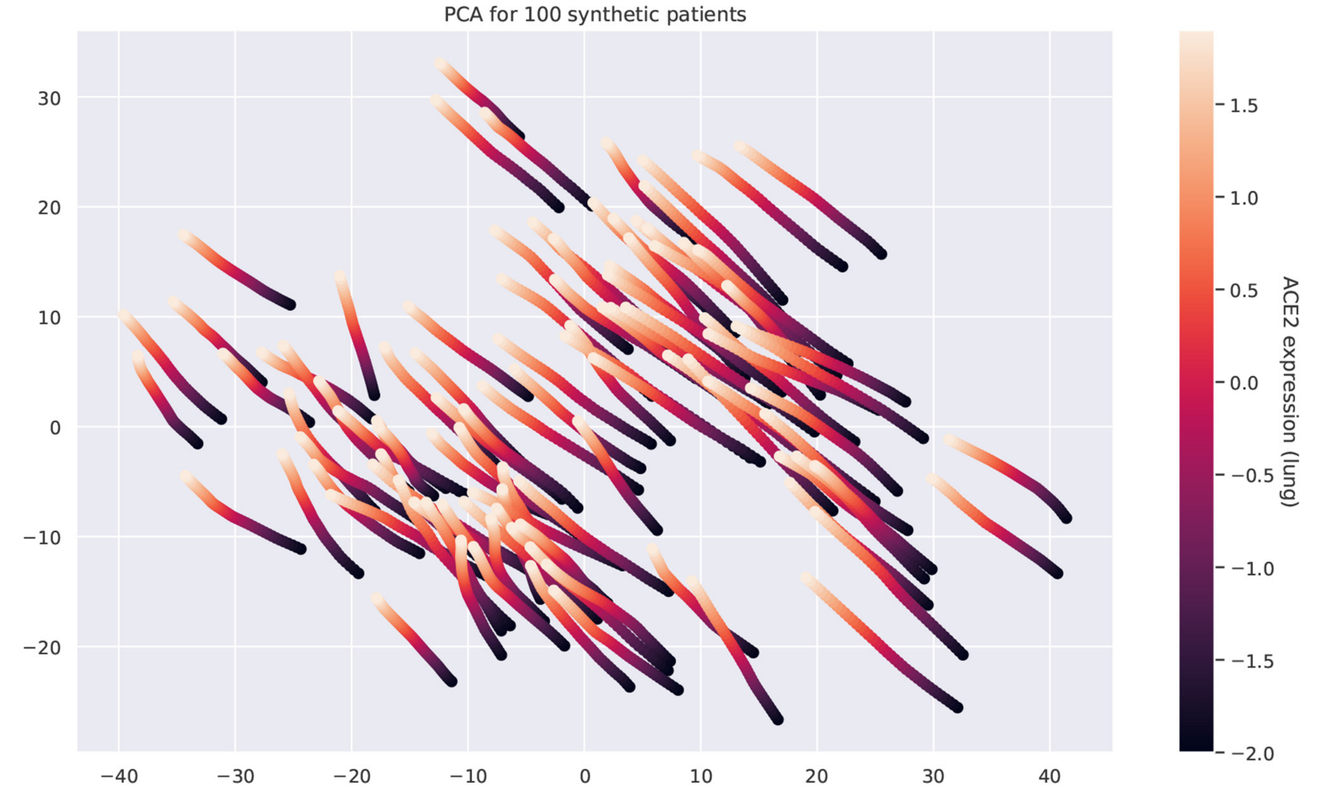


FIGURE 12 | Principal component analysis of the multi-tissue expression of 100 synthetic patients for different levels of ACE2 expression. Each line corresponds to a unique patient. For each patient, we fix all the latent covariates and modify the levels of ACE2 in lung. Overexpressing ACE2 leads to changes in the expression of other genes and these changes follow a well-defined trajectory.

4.1.2. A Modular Approach

The models presented in this work (GAN and GNN) are independent of each other. On the one hand, the main goal of the GNN model is to forecast various patient's conditions based on real or synthetic data, integrating information that spans multiple layers of the human body. On the other hand, the GAN model is able to generate data under different states, effectively enriching the space of pathophysiological conditions and endowing the digital twin with the ability to simulate the effects of counterfactual events. The independence of these two models enables a modular framework wherein each module can be trained separately on a distinct data modality. Importantly, these modules can be composed and reused through transfer learning. In this work, we have shown how computational models can be used to generate synthetic training data representing physiological conditions. Following the same principles, each module of a complex architecture could be pre-trained on synthetic simulations, refined using data obtained from horizontal population studies, and finally personalized according to clinical health records.

4.1.3. Next-Generation Datasets

The GAN and the GNN models can be interconnected in a synergistic way. In order to train the GNN effectively, it is necessary to have access to heterogeneous, paired data modalities (from different layers: genomic, transcriptomic, cellular, organ, exposomic, etc.) collected from a comprehensive collection of patients and encompassing a wide variety of conditions. However, to the best of our knowledge, to this date no such dataset exists. This is mainly because collecting paired, multilayer data from patients is expensive and entails important ethical and privacy concerns (Jobin et al., 2019; Mittelstadt, 2019). To address this issue, our GAN framework can synthesize data at multiple layers conditioned on the patient's conditions (e.g., diabetic, hypertension, etc.) and clinical information (e.g., heart rate, blood pressure, age, sex, ethnicity, exercise, nutrition, etc.). This synthetic data can be used to train the GNN and impute missing data modalities of real patients. Yet, the lack of real data from patients remains the key limitation for the introduction of our framework in clinical practice.

4.1.4. Explainable AI

The lack of interpretability of deep learning models has been one of the most significant barriers preventing their application in healthcare. Such models exhibit great capacity (Hornik, 1991) but understanding their behavior and following their decision-making process is not trivial (Castelvecchi, 2016). There is a growing body of literature focusing on interpretable artificial intelligence and interpretable deep learning aiming at developing white box models or at explaining black box ones (Das and Rad, 2020). Among such techniques, GNNs have started drawing the attention of both research and industry communities (Bronstein et al., 2017; Zhou et al., 2018). Such models are much more interpretable with respect to other neural approaches thanks to their graph structure, which is quite easy to understand from a human standpoint and a few studies have already shown how graph networks can be effectively employed in biology and healthcare (Zitnik et al., 2018; Gysi et al., 2020).

4.2. Advantages, Limitations, and Visions

4.2.1. Toward Precision and Predictive Medicine

The future of medicine is already bound to AI (Topol, 2019). Technological innovations are completely changing medicine perspectives expanding its horizons and moving toward an holistic view of human beings. The destiny of the whole healthcare system depends on this radical paradigm shift. Embracing AI innovations is just a technological prerequisite, and the first step toward a total transformation of how medicine currently works is delivered and perceived by patients. Thinking that AI will just and mainly improve clinical decision making is wrong. AI may actually open the doors to completely new ways of investigating the human body as a whole. The core and ultimate purpose of health will be developing preventative and personalized pathways to well-being rather than delivering treatments. The future foreseen is that AI will assist medicine in improving diagnosis and devising novel therapeutic strategies to deliver more effective solutions. The current healthcare revolution will not take back all the past technological advances, but it will show them under a new light.

4.2.2. Patient's Benefit

A meaningful quote about twins is the following: being a twin is like being born with a best friend. The data integration will make a better portrait of patient's condition trajectories but will require data inter-operability and data security. Technology is often not neutral, but transformed to be biased in one way or another (Ellul et al., 1954). Individuals can have different unforeseen readings and usage of new technologies. It may increase both user vulnerability and user empowerment. The vulnerability is the combination of exposure to the variety of personal medical data and the coping capabilities of users that could be different between young and mature people, as young are usually quicker in incorporating a new technology into everyday life. The user is empowered if he/she acquires awareness and control of his/her condition and context. A common example are online (website and blogs) initiatives such as patientslikeme that allow the user to search and make up his/her mind about a disease (Wicks et al., 2010). Instead, the user disempowerment depends on the lack of technical knowledge of how mechanisms work; this is even enhanced in black box techniques such as deep learning.

4.2.3. Training Clinicians

We believe that improving both data integration and predictability will provide physicians with improved medical decisions support systems and a decrease in both costs, through the evaluation of best therapies, and errors. A limitations is the poor interpretability and explainability in deep learning architectures. This limitation will also greatly affect the training of the new clinicians on AI technologies. There are growing efforts to make neural networks more interpretable in order to keep the human (doctors and patients) in the loop. The interpretability could be improved by using parallel mechanistic computational modeling and simulations (Milanesi et al., 2009; Bartocci and Lió, 2016), model extraction libraries (see, for instance, Kazhdan et al., 2020), and visual inference tools (Bodnar et al., 2020). This tool could also be complemented by clinical decision support systems such as Müller and Lio (2020). The complexly structured and multilevel comorbidity and frailty patterns of most diseases describe a highly dynamical system and are, therefore, challenging current medical therapies.

4.2.4. AI for Evidence-Based Medicine

From a clinical standpoint, AI will support a plethora of different tasks from medical check up to personalized intervention strategies to contrast ripple effects or to promote healthy habits. In non-acute states, predictive inference will propose prevention plans for comorbidity management, particularly in presence of multiple therapies (Rivera, 2020). Increasingly large amount of personal data will be collected to feed modular machine learning (ML) models organized to address specific and personalized medical issues. Clinical endpoints will be constantly monitored, shared, and compared in order to answer relevant research questions and to deliver the best possible service. A deeper understanding and practice of modeling in medicine will produce better investigation of complex biological processes, and even new ideas and better feedback into medicine. Modeling-based approaches combined with data-driven ML techniques will progressively provide models with higher degree of interpretability and generalization ability (Barbiero et al., 2020a), which will make evidence-based medicine even more accessible intensifying the involvement of patients in the decision-making process. AI simulations forecasting the evolution of clinical endpoints over time will also reshape clinical guidelines (Rivera, 2020), which will no longer be based just on *horizontal* population studies. Cross-modality data will be collected for each patient and machine learning models will be used to predict a bundle of possible trajectories representing the future states of the patient allowing for personalized prescriptions, surgical planning, and medical interventions.

4.2.5. Social Impact

Ethical repercussions will also be huge (Jobin et al., 2019; Mittelstadt, 2019). The transition will call for deeper trans-disciplinary research and a substantial technological innovation in a variety of research and social areas. Here, education will play a key role in changing lifestyle habits and the way health is perceived, communicated, and delivered (Yu et al., 2017). For each individual, both healthcare systems and private companies will collect, save, and eventually exploit an enormous amount of personal data. Providing an effective, stable, and unified juridical overview is critical on this matter (Panch et al., 2019).

4.2.6. Next-Generation Medical Devices

AI will change the leading *vehicle* of medicine. The demand for AI-powered and internet of things (IoT) devices is increasing worldwide. The future equipment for precision medicine will likely required to be cheap and extremely modular, but more importantly it needs to be deployable in dedicated hardware to be distributed in larger markets. Our digital twin model aims at providing the first example of a novel class of AI-assisted tools for precision and predictive medicine. Our framework is designed to scale to medical device deployment and run time monitoring and verification combining ideas from systems medicine with scientific computing and machine learning. The integration of interpretable AI models in clinical devices may lead to a deep transformation in healthcare paving the way for a next generation of tools for precision medicine probing the inner workings of full body in well-being and disease conditions.

5. METHODS

5.1. Graph Neural Network

5.1.1. Graph Network Blocks

The GNN framework proposed by Battaglia et al. (2018) is based on modules called graph network blocks (GN blocks) representing the core computation units of a GNN. Multiple GN blocks can be composed of or even combined with other neural networks to generate complex architectures. A GNN can be defined as a 3-tuple $G = (\mathbf{u}, H, E)$. $H = \{\mathbf{h}_i\}_{i=1:N^v}$ is the node set where the feature of each node is denoted by $\mathbf{h}_i$. $E = \{(\mathbf{e}_k, r_k, s_k)\}$ is the edge set where each node is represented by its features $\mathbf{e}_k$, the receiver node r_k, and the sender node s_k. $\mathbf{u}$ denotes a set of global attributes representing the state of the underlying system. Each GN block consists of three update functions, ϕ, and three aggregation functions, ρ:

$$\begin{aligned} \mathbf{e}'_k &= \phi^e(\mathbf{e}_k, \mathbf{h}_{r_k}, \mathbf{h}_{s_k}, \mathbf{u}) & \bar{\mathbf{e}}'_i &= \rho^{e\to h}(E'_i) \\ \mathbf{h}'_i &= \phi^h(\bar{\mathbf{e}}_k, \mathbf{h}_i, \mathbf{u}) & \bar{\mathbf{e}}' &= \rho^{e\to u}(E') \\ \mathbf{u}' &= \phi^u(\mathbf{e}', \mathbf{h}', \mathbf{u}) & \bar{\mathbf{h}}' &= \rho^{h\to u}(H') \end{aligned} \quad (1)$$

where $E'_i = \{(\mathbf{e}'_k, r_k, s_k)\}$, $H' = \{(\mathbf{h}'_i)\}_{i=1:N^v}$, and $E' = \bigcup_i E'_i = \{(\mathbf{e}'_k, r_k, s_k)\}_{k=1:N^e}$. In order to train a GN block in full, six computation steps are required, alternating the update and aggregation functions. For each edge, E'_i is computed through the update function ϕ^e. The result is then aggregated by means of the function $\rho^{e\to v}$. The output $\bar{\mathbf{e}}'_i$ corresponds to an edge update and it is employed to update node representations $\mathbf{h}'_i$ by means of ϕ^h. $\rho^{e\to u}$ and $\rho^{h\to u}$ perform aggregation steps generating $\bar{\mathbf{e}}'$ and $\bar{\mathbf{h}}'$ from edge and node updates, respectively. Global attributes represented by $\mathbf{u}'$ are computed leveraging the information from $\bar{\mathbf{e}}'$, $\bar{\mathbf{h}}'$, and $\mathbf{u}$ via the function ϕ^u. The learning process of each GN block may be independent or co-dependent with other blocks. Constraints may apply on edges, information flows, or global attributes, depending on the application. In this work, we are just interested in the evaluation of global attributes to monitor clinical endpoints and we did not apply any learning constraint, even if in clinical practice may still be of great interest. Given a set of labels for global attributes $\mathbf{t} = \{t_i\}_{i=1:N^v}$ and the corresponding predictions provided by the GN block $\widehat{\mathbf{u}}' = \{\widehat{u}'_i\}_{i=1:N^v}$ representing the evolution of the underlying biological system, we aim at minimizing the following objective function:

$$\min_{\theta} \frac{1}{N^v} \sum_{i=1}^{N^v} \left(t_i - \widehat{u}'_i\right)^2 \quad (2)$$

where θ is the set of model's parameters.

5.1.2. Assessing Prediction Uncertainty

The aim of developing a digital patient model is to provide an accurate estimation of the trajectory of a patient by forecasting clinically relevant endpoints. In such a context, quantifying model uncertainty is critical. One of the most established techniques relies upon the use of dropout (Srivastava et al., 2014) at test time, as a Bayesian approximation, without sacrificing either computational complexity or test performance (Gal and Ghahramani, 2016b). In this framework, the first two moments of the predictive distribution q performing T stochastic forward passes for a sample $\mathbf{x}^*$ with label $\mathbf{y}^*$ can be estimated as (Gal and Ghahramani, 2016a):

$$\mathbb{E}_{q(\mathbf{y}^*, \mathbf{x}^*)}(\mathbf{y}^*) \approx \frac{1}{T} \sum_{i=1}^{T} \widehat{\mathbf{y}}^*(\mathbf{x}^*, W_1^t, \ldots, W_L^t) \quad (3)$$

$$\begin{aligned} \mathrm{Var}_{q(\mathbf{y}^*, \mathbf{x}^*)}(\mathbf{y}^*) \approx\ & \tau^{-1} I_D \\ & + \frac{1}{T} \sum_{i=1}^{T} \widehat{\mathbf{y}}^*(\mathbf{x}^*, W_1^t, \ldots, W_L^t)^T \widehat{\mathbf{y}}^*(\mathbf{x}^*, W_1^t, \ldots, W_L^t) \\ & - \mathbb{E}_{q(\mathbf{y}^*, \mathbf{x}^*)}(\mathbf{y}^*)^T \mathbb{E}_{q(\mathbf{y}^*, \mathbf{x}^*)}(\mathbf{y}^*) \end{aligned} \quad (4)$$

where $\widehat{\mathbf{y}}^*$ is the predicted label, $\{W_i\}_{i=1}^{L}$ is a set of random variables representing the weights of a neural network with L layers, I_D is an identity matrix, D is the number of output units of the neural network, and τ is a precision hyper-parameter. The method has also been generalized to convolutional (Gal and Ghahramani, 2015) and recurrent networks (Gal and Ghahramani, 2016c).

Here, we show how such technique can be used to quantify the uncertainty of a GNN by generating a predictive distribution of the trajectories representing the future states of the patient. Let $x_1^*, \ldots, x_k^*$ be a sequence of real values representing a clinical endpoint measured at $1, \ldots, k$ time steps. Let f^t be a stochastic model that takes a sequence $x_1^*, \ldots, x_k^*$ as input and it outputs a prediction $\widehat{y}^* \in \mathbb{R}$. We are interested in estimating a predictive distribution of the trajectories of the variable x over the next $k+1, \ldots, k+h$ time steps. To this aim, we can use an iterative algorithm by generating one trajectory at a time. The first prediction $\widehat{y}_{k+1}^*$ can be generated as:

$$\widehat{y}_{k+1}^{*,t} = f^t(x_1^*, \ldots, x_k^*) \quad (5)$$

By using the obtained prediction and sliding the time window one time step further, we can generate the first prediction for the second time step $k+2$:

$$\widehat{y}^{*,t}_{k+2} = f^t(x^*_2, \ldots, x^*_k, \widehat{y}^{*,t}_{k+1}) \tag{6}$$

The procedure can be repeated for $k + h$ time steps to generate a single trajectory. Model uncertainty can be assessed building multiple trajectories by performing T stochastic forward passes. The resulting algorithm is equivalent to a Monte Carlo sampling as proven by Gal and Ghahramani (2016b). In our GNN model, the approach we just described can be easily applied for each node in order to assess the uncertainty of clinical endpoints.

5.2. Generative Adversarial Network

Consider a dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$ of samples from an unknown distribution $\mathbb{P}_{\mathbf{x},\mathbf{m},\mathbf{r},\mathbf{q}}$, where $\mathbf{x} \in \mathbb{R}^{t \times n}$ represents a matrix of n gene expression values in t tissues; $\mathbf{m} \in \{0, 1\}^t$ is a mask vector indicating whether the expression of each tissue has been measured for the given patient; and $\mathbf{r} \in \mathbb{R}^k$ and $\mathbf{q} \in \mathbb{N}^c$ are vectors of k quantitative covariates (e.g., age) and c categorical (e.g., gender), respectively. Our goal is to produce realistic gene expression samples by modeling the conditional probability distribution $\mathbb{P}(\mathbf{X} = \mathbf{x}|\mathbf{M} = \mathbf{m}, \mathbf{R} = \mathbf{r}, \mathbf{Q} = \mathbf{q})$, where $\mathbf{r}$ includes the expression of ACE2 in different tissues (e.g., lung, kidney, and pancreas). By modeling this distribution, we can sample data for different conditions and quantify the uncertainty of the generated expression values.

To address this problem, we extend the model proposed in Viñas et al. (2021) to simultaneously account for t tissue types from the same donor. In particular, our method builds on a Wasserstein GAN with gradient penalty (WGAN-GP) (Arjovsky et al., 2017; Gulrajani et al., 2017). Similar to Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), WGAN-GPs estimate a generative model via an adversarial process driven by the competition between two players, the *generator* and the *critic*.

The generator aims at producing samples from the conditional $\mathbb{P}(\mathbf{X}|\mathbf{M}, \mathbf{R}, \mathbf{Q})$. Formally, we define the generator as a function $G_\theta : \mathbb{R}^u \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^{t \times n}$ parameterized by θ that generates gene expression values $\hat{\mathbf{x}}$ as follows:

$$\hat{\mathbf{x}} = \mathbf{m} \odot G_\theta(\mathbf{z}, \mathbf{r}, \mathbf{q}) \tag{7}$$

where $\mathbf{z} \in \mathbb{R}^u$ is a vector sampled from a fixed noise distribution $\mathbb{P}_\mathbf{z}$ and u is a user-definable hyperparameter. We apply the mask $\mathbf{m}$ element-wise to match the distribution of missing tissues of the training dataset.

The critic takes gene expression samples $\mathbf{x}$ from two input streams (the generator and the data distribution) and attempts to distinguish the true input source. Formally, the critic is a function $D_\omega : \mathbb{R}^{t \times n} \times \{0, 1\}^t \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}$ parameterized by ω that we define as follows:

$$\bar{y} = D_\omega(\bar{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q})$$

where the output $\bar{y}$ is an unbounded scalar that quantifies the degree of realism of an input sample $\bar{\mathbf{x}}$ given the covariates $\mathbf{r}$ and $\mathbf{q}$ (e.g., high values correspond to real samples and low values correspond to fake samples). When the expression of a certain tissue t is unavailable for a given patient, we set the unobserved values of tissue t in $\bar{\mathbf{x}}$ to 0 and the t-th component of the mask $\mathbf{m}$ to 0.

We optimize the generator and the critic adversarially. Following (Arjovsky et al., 2017), we train the generator G_θ and the critic D_ω to solve the following minimax game based on the Wasserstein distance:

$$\begin{aligned} &\min_\theta \max_\omega \underset{\mathbf{x},\mathbf{m},\mathbf{r},\mathbf{q}\sim\mathbb{P}_{\mathbf{x},\mathbf{m},\mathbf{r},\mathbf{q}}}{\mathbb{E}} \left[D_\omega(\mathbf{x}, \mathbf{m}, \mathbf{r}, \mathbf{q}) - \underset{\mathbf{z}\sim\mathbb{P}_\mathbf{z}}{\mathbb{E}} [D_\omega(\hat{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q})] \right] \\ &\text{subject to} \quad ||D_\omega(\mathbf{x}_i, \mathbf{m}, \mathbf{r}, \mathbf{q}) - D_\omega(\mathbf{x}_j, \mathbf{m}, \mathbf{r}, \mathbf{q})|| \le ||\mathbf{x}_i - \mathbf{x}_j|| \\ &\qquad\qquad \forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{t \times n}, \mathbf{m} \in \{0, 1\}^t, \mathbf{r} \in \mathbb{R}^k, \mathbf{q} \in \mathbb{N}^c \end{aligned} \tag{8}$$

where $\hat{\mathbf{x}}$ is defined as in Equation (7) and the constraint enforces a soft version of the 1-Lipschitz constraint (e.g., the norm of the critic's gradient with respect to $\mathbf{x}$ must be at most 1 everywhere).

Let $\{(\mathbf{x}_i, \mathbf{m}_i, \mathbf{r}_i, \mathbf{q}_i)\}_{i=1}^b$ be a mini-batch of b independent samples from the training dataset $\mathcal{D}$. Let $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_k\}$ be a set of k vectors sampled independently from the noise distribution $\mathbb{P}_\mathbf{z}$ and let us define the synthetic samples corresponding to the mini-batch as $\hat{\mathbf{x}}_i = \mathbf{m}_i \odot G_\theta(\mathbf{z}_i, \mathbf{r}_i, \mathbf{q}_i)$ for each i in $[1, 2, ..., k]$. We solve the minimax problem described in Equation (8) by interleaving mini-batch gradient updates for the generator and the critic, optimizing the following problems:

$$\begin{aligned} &\text{Generator:} \quad \min_\theta \quad -\frac{1}{k}\sum_{i=1}^{k} D_\omega\big(\hat{\mathbf{x}}_i, \mathbf{m}_i, \mathbf{r}_i, \mathbf{q}_i\big) \\ &\text{Critic:} \quad \min_\omega \quad \frac{1}{k}\sum_{i=1}^{k} D_\omega\big(\hat{\mathbf{x}}_i, \mathbf{m}_i, \mathbf{r}_i, \mathbf{q}_i\big) - D_\omega(\mathbf{x}_i, \mathbf{m}_i, \mathbf{r}_i, \mathbf{q}_i) \\ &\qquad\qquad + \frac{\lambda}{k}\sum_{i=1}^{k} \big(||\nabla_{\tilde{\mathbf{x}}_i} D_\omega(\tilde{\mathbf{x}}_i, \mathbf{m}_i, \mathbf{r}_i, \mathbf{q}_i)||_2 - 1\big)^2 \end{aligned} \tag{9}$$

where λ is a user-definable hyperparameter and each $\tilde{\mathbf{x}}_i$ is a random point along the straight line that connects $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$, that is, $\tilde{\mathbf{x}}_i = \alpha_i \mathbf{x}_i + (1 - \alpha_i)\hat{\mathbf{x}}_i$ with $\alpha_i \sim \mathcal{U}(0, 1)$. Intuitively, since enforcing the 1-Lipschitz constraint everywhere is intractable (see Equation 8), the second term of the critic problem is a relaxed version of the constraint that penalizes the gradient norm along points in the straight lines that connect real and synthetic samples (Gulrajani et al., 2017).

5.2.1. Architecture

Figure 1 shows the architecture of both players. The generator G receives a noise vector $\mathbf{z}$ as input (green box) as well as sample covariates $\mathbf{r}$ and $\mathbf{q}$ (orange boxes) and produces a vector $\hat{\mathbf{x}}$ of synthetic expression values (red box). The critic D takes either a real gene expression sample $\mathbf{x}$ (blue box) or a synthetic sample $\hat{\mathbf{x}}$ (red box), in addition to sample covariates $\mathbf{r}$ and $\mathbf{q}$, and attempts to distinguish whether the input sample is real or fake. For both players, we use word embeddings (Mikolov et al., 2013) to model the sample covariates (light green boxes), a distinctive feature that allows to learn distributed, dense representations for the different tissue types and, more generally, for all the categorical covariates $\mathbf{q} \in \mathbb{N}^c$.

Formally, let q_j be a categorical covariate (e.g., tissue type) with vocabulary size v_j, that is, $q_j \in \{1, 2, ..., v_j\}$, where each value in the vocabulary $\{1, 2, ..., v_j\}$ represents a different category (e.g., whole blood or kidney). Let $\bar{\mathbf{q}}_j \in \{0, 1\}^{v_j}$ be a one-hot vector such that $\bar{q}_{jk} = 1$ if $q_j = k$ and $\bar{q}_{jk} = 0$ otherwise. Let d_j be the dimensionality of the embeddings for covariate j. We obtain a vector of embeddings $\mathbf{e}_j \in \mathbb{R}^{d_j}$ as follows:

$$\mathbf{e}_j = \mathbf{W}_j \bar{\mathbf{q}}_j$$

where each $\mathbf{W}_j \in \mathbb{R}^{d_j \times v_j}$ is a matrix of learnable weights. Essentially, this operation describes a lookup search in a dictionary with v_j entries, where each entry contains a learnable d_j-dimensional vector of embeddings that characterizes each of the possible values that q_j can take. To obtain a global collection of embeddings $\mathbf{e}$, we concatenate all the vectors $\mathbf{e}_j$ for each categorical covariate j:

$$\mathbf{e} = \Big\|_{j=1}^{c} \mathbf{e}_j$$

where c is the number of categorical covariates and $\|$ represents the concatenation operator. We then use the learnable embeddings $\mathbf{e}$ in downstream tasks.

In terms of the player's architecture, we model both the generator G_θ and critic D_ω as neural networks that leverage independent instances $\mathbf{e}^G$ and $\mathbf{e}^D$ of the categorical embeddings for their corresponding downstream tasks. Specifically, we model the two players as follows:

$$G_\theta(\mathbf{z}, \mathbf{r}, \mathbf{q}) = \mathrm{MLP}(\mathbf{z}\|\mathbf{r}\|\mathbf{e}^G) \qquad D_\omega(\bar{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q}) = \mathrm{MLP}(\bar{\mathbf{x}}\|\mathbf{m}\|\mathbf{r}\|\mathbf{e}^D)$$

where MLP denotes a multilayer perceptron.

5.3. Ridge Regression

We model the expression of genes from the renin-angiotensin system in lung, kidney, pancreas, and heart as a function of genes in the chemokine, TNF, and TGF-β pathways in blood. Let $\mathbf{Y} = (Y_1, ..., Y_n)^\top$ and $\mathbf{X} = (X_1, ..., X_m)^\top$ be multivariate random variables representing the expression of the n genes in the renin-angiotensin system and the m genes in the signaling pathways, respectively. Our model is based on ridge regression (Hoerl and Kennard, 1970):

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \boldsymbol{\epsilon}$$

where $\mathbf{W} \in \mathbb{R}^{m \times n}$ is a matrix of learnable weights and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ are the residuals. We optimize the following objective:

$$\min_{\mathbf{W}} ||\mathbf{Y} - \mathbf{X}\mathbf{W}||_2^2 + \alpha||\mathbf{W}||_2^2$$

where α is a hyperparameter that controls the regularization strength. Alternative non-linear models such as support vector machines, Gaussian processes, and random forests did not improve our cross-validation scores.

DATA AVAILABILITY STATEMENT

We leverage data from the Genotype-Tissue Expression (GTEx) project (v8), a resource that has generated a comprehensive collection of human transcriptome data in a diverse set of tissues (Aguet et al., 2019). We use the KEGG pathway database (Kanehisa et al., 2010) to select genes from the renin-angiotensin system (*hsa04614*), chemokine (*hsa04062*), TNF (*hsa04668*), and TGF-β (*hsa04350*) pathways.

The computational ODE-based system described in Barbiero and Lió (2020) is used to generate a time series for each differential equation with a window size of $\tau = 500$ time steps (Barbiero and Lio, 2020). Time series are collected, randomly shuffled, and stacked in a dataset. Each item of the collection is randomly assigned either to a training ($n_{train} = 3200$), validation ($n_{val} = 800$), or test set ($n_{test} = 1000$).

CODE AVAILABILITY

All the code for the experiments has been implemented in Python 3, relying upon open-source libraries (Pedregosa et al., 2011; Abadi et al., 2016; Wang et al., 2019). All the experiments have been run on the same machine: Intel® Core™ i7-8750H 6-Core Processor at 2.20 GHz equipped with 8 GB RAM. To enable code reuse, the Python code for the mathematical models including parameter values and documentation is freely available under GNU Public License from a GitHub repository[1] (Barbiero et al., 2020b). Unless required by applicable law or agreed to in writing, software is distributed on an "as is" basis, without warranties or conditions of any kind, either express or implied.

AUTHOR CONTRIBUTIONS

PB and RV performed the experiments. All authors were involved in conceiving the approach, co-wrote the paper, designing the experiments, reviewing and discussing the data, and commented on the manuscript.

FUNDING

This project has received funding from "la Caixa" Foundation (ID 100010434), under agreement 667 LCF/BQ/EU19/11710059, and from the Engineering and Physical Sciences 668 Research Council under grant agreement No EP/R513180/1 (R.V. EPSRC DTG 2018/19).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.652907/full#supplementary-material

[1] https://github.com/pietrobarbiero/digital-patient

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (Savannah, GA), 265–283.

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., et al. (2019). The gtex consortium atlas of genetic regulatory effects across human tissues. *bioRxiv [Preprint]*. doi: 10.1101/787903

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv [Preprint]. arXiv:1701.07875*.

Bangalore, S., Maron, D. J., O'Brien, S. M., Fleg, J. L., Kretov, E. I., Briguori, C., et al. (2020). Management of coronary disease in patients with advanced kidney disease. *N. Engl. J. Med.* 382, 1608–1618. doi: 10.1056/nejmoa1915925

Barbiero, P., and Lió, P. (2020). The computational patient has diabetes and a covid. *arXiv [Preprint]. arXiv:2006.06435*.

Barbiero, P., and Lio, P. (2020). Pietrobarbiero/computational-patient: absolutno. doi: 10.5281/zenodo.4030228

Barbiero, P., Squillero, G., and Tonda, A. (2020a). Modeling generalization in machine learning: A methodological and computational study. *arXiv [Preprint]. arXiv:2006.15680*.

Barbiero, P., Torne, R. V., and Lio, P. (2020b). pietrobarbiero/digital-patient: absolutno. doi: 10.5281/zenodo.4030220

Bartocci, E., and Lió, P. (2016). Computational modeling, formal analysis, and tools for systems biology. *PL S Computat. Biol.* 12:e1004591. doi: 10.1371/journal.pcbi.1004591

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv [Preprint]. arXiv:1806.01261*.

Bica, I., Andrés-Terré, H., Cvejic, A., and Liò, P. (2020). Unsupervised generative and graph representation learning for modelling cell differentiation. *Sci. Rep.* 10, 1–13. doi: 10.1038/s41598-020-66166-8

Bodnar, C., Cangea, C., and Liò, P. (2020). Deep graph mapper: seeing graphs through the neural lens. *arXiv [Preprint]. arXiv:2002.03864*.

Bollobás, B. (2013). *Modern graph Theory*, Vol. 184. New York, NY: Springer Science & Business Media.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Process. Mag.* 34, 18–42. doi: 10.1109/msp.2017.2693418

Brown, S. A., Kovatchev, B. P., Raghinaru, D., Lum, J. W., Buckingham, B. A., Kudva, Y. C., et al. (2019). Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. *N. Engl. J. Med.* 381, 1707–1717. doi: 10.1056/NEJMoa1907863

Castelvecchi, D. (2016). Can we open the black box of AI? *Nat. News* 538:20. doi: 10.1038/538020a

Clark, D., Nicholls, S. J., John, J. S., Elshazly, M. B., Ahmed, H. M., Khraishah, H., et al. (2019). Visit-to-visit blood pressure variability, coronary atheroma progression, and clinical outcomes. *JAMA Cardiol.* 4:437. doi: 10.1001/jamacardio.2019.0751

Colbran, L. L., Gamazon, E. R., Zhou, D., Evans, P., Cox, N. J., and Capra, J. A. (2019). Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat. Ecol. Evol.* 3, 1598–1606. doi: 10.1038/s41559-019-0996-x

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–194. doi: 10.1038/nrg2537

Das, A., and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): a survey. *arXiv [Preprint]. arXiv:2006.11371*.

Du, J., Johnson, L. M., Jacobsen, S. E., and Patel, D. J. (2015). Dna methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* 16, 519–532. doi: 10.1038/nrm4043

Eddy, D. M., and Schlessinger, L. (2003). Archimedes: a trial-validated model of diabetes. *Diabetes Care* 26, 3093–3101. doi: 10.2337/diacare.26.11.3093

Ellul, J., Ellul, J., Ellul, J., Juriste, P., and Ellul, J. (1954). *La technique ou l'enjeu du siècle*. A. Colin Paris.

Elmarakby, A. A., Quigley, J. E., Olearczyk, J. J., Sridhar, A., Cook, A. K., Inscho, E. W., et al. (2007). Chemokine receptor 2b inhibition provides renal protection in angiotensin ii–salt hypertension. *Hypertension* 50, 1069–1076. doi: 10.1161/HYPERTENSIONAHA.107.098806

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., et al. (2008). Genetics of gene expression and its effect on disease. *Nature* 452, 423–428. doi: 10.1038/nature06758

Evans, W. E., and Relling, M. V. (2004). Moving towards individualized medicine with pharmacogenomics. *Nature* 429, 464–468. doi: 10.1038/nature02626

Fountain, J. H., and Lappin, S. L. (2019). "Physiology, renin angiotensin system," in *StatPearls*, (StatPearls Publishing).

Gal, Y., and Ghahramani, Z. (2015). Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv [Preprint]. arXiv:1506.02158*.

Gal, Y., and Ghahramani, Z. (2016a). Dropout as a bayesian approximation: appendix 20. *arXiv [Preprint]. arxiv:1506.02157*.

Gal, Y., and Ghahramani, Z. (2016b). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 1050–1059.

Gal, Y., and Ghahramani, Z. (2016c). "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in Neural Information Processing Systems*, 1019–1027.

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., et al. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nat. Genet.* 50, 956–967. doi: 10.1038/s41588-018-0154-4

Garvin, M. R., Alvarez, C., Miller, J. I., Prates, E. T., Walker, A. M., Amos, B. K., et al. (2020). A mechanistic model and therapeutic interventions for covid-19 involving a ras-mediated bradykinin storm. *eLife* 9:e59177. doi: 10.7554/eLife.59177.sa2

Geiger, B., Bershadsky, A., Pankov, R., and Yamada, K. M. (2001). Transmembrane crosstalk between the extracellular matrix and the cytoskeleton. *Nat. Rev. Mol. Cell Biol.* 2, 793–805. doi: 10.1038/35099066

Gelernter, D. (1993). *Mirror Worlds: Or: The Day Software Puts the Universe in a Shoebox. How It Will Happen and What It Will Mean.* Oxford University Press.

Ginsburg, G. S., and Willard, H. F. (2009). Genomic and personalized medicine: foundations and applications. *Transl. Res.* 154, 277–287. doi: 10.1016/j.trsl.2009.09.005

Gironacci, M. M., Adamo, H. P., Corradi, G., Santos, R. A., Ortiz, P., and Carretero, O. A. (2011). Angiotensin (1-7) induces mas receptor internalization. *Hypertension* 58, 176–181. doi: 10.1161/HYPERTENSIONAHA.111.173344

Goldberger, A. L., Amaral, L. A., Hausdorff, J. M., Ivanov, P. C., Peng, C.-K., and Stanley, H. E. (2002). Fractal dynamics in physiology: alterations with disease and aging. *Proc. Natl. Acad. Sci. U.S.A.* 99, 2466–2472. doi: 10.1073/pnas.012579499

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2672–2680.

Grieves, M. (2015). "Digital twin: manufacturing excellence through virtual factory replication," *Whitepaper*.

Groth, A., Vrugt, B., Brock, M., Speich, R., Ulrich, S., and Huber, L. C. (2014). Inflammatory cytokines in pulmonary hypertension. *Respir. Res.* 15, 1–10. doi: 10.1186/1465-9921-15-47

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of Wasserstein GANs. *arXiv [Preprint]*. arXiv: 1704.00028.

Gutierrez, N. G., Mathew, M., McCrindle, B. W., Tran, J. S., Kahn, A. M., Burns, J. C., et al. (2019). Hemodynamic variables in aneurysms are associated with thrombotic risk in children with Kawasaki disease. *Int. J. Cardiol.* 281, 15–21. doi: 10.1016/j.ijcard.2019.01.092

Gysi, D. M., Valle, Í. D., Zitnik, M., Ameli, A., Gan, X., Varol, O., et al. (2020). Network medicine framework for identifying drug repurposing opportunities for covid-19. *arXiv [Preprint]. arXiv:2004.07229*.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.* 4, 251–257. doi: 10.1016/0893-6080(91)90009-T

Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D., and Chang, Y. (2020). Graphlime: local interpretable model explanations for graph neural networks. *arXiv [Preprint]. arXiv:2001.06216*.

Jebara, T. (2012). *Machine Learning: Discriminative and Generative*, Vol. 755. Springer Science & Business Media.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360. doi: 10.1093/nar/gkp896

Kazhdan, D., Shams, Z., and Liò, P. (2020). Marleme: a multi-agent reinforcement learning model extraction library. *arXiv [Preprint]. arXiv:2004.07928*.

Kim, B. J., Kwon, S. U., Wajsbrot, D., Koo, J., Park, J. M., and Jeffers, B. W. (2018). Relationship of inter-individual blood pressure variability and the risk for recurrent stroke. *J. Am. Heart Assoc.* 7:e009480. doi: 10.1161/jaha.118.009480

Kovatchev, B. (2019). A century of diabetes technology: signals, models, and artificial pancreas control. *Trends Endocrinol. Metab.* 30, 432–444. doi: 10.1016/j.tem.2019.04.008

Kuba, K., Imai, Y., Ohto-Nakanishi, T., and Penninger, J. M. (2010). Trilogy of ace2: a peptidase in the renin–angiotensin system, a sars receptor, and a partner for amino acid transporters. *Pharmacol. Therapeut.* 128, 119–128. doi: 10.1016/j.pharmthera.2010.06.003

Kusnadi, A., Ramírez-Suástegui, C., Fajardo, V., Chee, S. J., Meckiff, B. J., Simon, H., et al. (2020). Severely ill covid-19 patients display augmented functional properties in sars-cov-2-reactive cd8+ t cells. *bioRxiv [Preprint]*.

Laubenbacher, R., Sluka, J. P., and Glazier, J. A. (2021). Using digital twins in viral infection. *Science* 371, 1105–1106. doi: 10.1126/science.ab f3370

Lecue, F. (2019). On the role of knowledge graphs in explainable AI. *Semantic Web* 1, 1–11. doi: 10.3233/SW-190374

Lengyel, E., Makowski, L., DiGiovanni, J., and Kolonin, M. G. (2018). Cancer as a matter of fat: the crosstalk between adipose tissue and tumors. *Trends Cancer* 4, 374–384. doi: 10.1016/j.trecan.2018.03.004

Li, M., Gao, P., and Zhang, J. (2016). Crosstalk between autophagy and apoptosis: potential and emerging therapeutic targets for cardiac diseases. *Int. J. Mol. Sci.* 17:332. doi: 10.3390/ijms17030332

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. (2018). Learning deep generative models of graphs. *arXiv [Preprint]. arXiv:1803.03324*.

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nat. Med.* 26, 842–844. doi: 10.1038/s41591-020-0901-9

Lieberman, E., Hauert, C., and Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature* 433, 312–316. doi: 10.1038/nature03204

Lijnen, P. J., Petrov, V. V., and Fagard, R. H. (2003). Association between transforming growth factor-β and hypertension.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv [Preprint]*. arXiv: 1310.4546.

Milanesi, L., Romano, P., Castellani, G., Remondini, D., and Liò, P. (2009). Trends in modeling biomedical complex systems. *BMC Bioinformatics* 10:I1. doi: 10.1186/1471-2105-10-s12-i1

Mitchell, G. F., Hwang, S.-J., Vasan, R. S., Larson, M. G., Pencina, M. J., Hamburg, N. M., et al. (2010). Arterial stiffness and cardiovascular events. *Circulation* 121, 505–511. doi: 10.1161/circulationaha.109.886655

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* 1, 501–507. doi: 10.1038/s42256-019-0114-4

Müller, T. T., and Lio, P. (2020). PECLIDES neuro: a personalisable clinical decision support system for neurological diseases. *Front. Artif. Intell.* 3:23. doi: 10.3389/frai.2020.00023

Naylor, S., and Chen, J. Y. (2010). Unraveling human complexity and disease with systems biology and personalized medicine. *Pers. Med.* 7, 275–289. doi: 10.2217/pme.10.16

Neal, M. L., and Bassingthwaighte, J. B. (2007). Subject-specific model estimation of cardiac output and blood volume during hemorrhage. *Cardiovasc. Eng.* 7, 97–120. doi: 10.1007/s10558-007-9035-7

O'Rourke, M. F., and Nichols, W. W. (2005). Aortic diameter, aortic stiffness, and wave reflection increase with age and isolated systolic hypertension. *Hypertension* 45, 652–658. doi: 10.1161/01.hyp.0000153793.84859.b8

Panch, T., Mattie, H., and Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *NPJ Digit. Med.* 2, 1–3. doi: 10.1038/s41746-019-0155-4

Pearson, K. (1901). On lines of closes fit to system of points in space, London, E Dinb. *Dublin Philos. Mag. J. Sci* 2, 559–572. doi: 10.1080/14786440109462720

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Poyton, R. O., and McEwen, J. E. (1996). Crosstalk between nuclear and mitochondrial genomes. *Annu. Rev. Biochem.* 65, 563–607. doi: 10.1146/annurev.bi.65.070196.003023

Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., et al. (2019). Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* 51, 354–362. doi: 10.1038/s41588-018-0316-4

Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., Boxer, A., Blennow, K., et al. (2007). Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. *Nat. Med.* 13, 1359–1362. doi: 10.1038/nm1653

Rivera, S. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-AI extension. *BMJ* 26, 1351–1363. doi: 10.1038/s41591-020-1037-7

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biol.* 11, 1–9. doi: 10.1186/gb-2010-11-3-r25

Rudemiller, N. P., and Crowley, S. D. (2017). The role of chemokines in hypertension and consequent target organ damage. *Pharmacol. Res.* 119, 404–411. doi: 10.1016/j.phrs.2017.02.026

Sardu, C., Gambardella, J., Morelli, M. B., Wang, X., Marfella, R., and Santulli, G. (2020). Hypertension, thrombosis, kidney failure, and diabetes: Is COVID-19 an endothelial disease? a comprehensive evaluation of clinical and basic evidence. *J. Clin. Med.* 9:1417. doi: 10.3390/jcm9051417

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. Neural Netw.* 20, 61–80. doi: 10.1109/TNN.2008.2005605

Shafto, M., Conroy, M., Doyle, R., Glaessgen, E., Kemp, C., LeMoigne, J., et al. (2010). *Modeling, Simulation, Information Technology and Processing Roadmap*. National Aeronautics and Space Administration.

Shang, J. K., Esmaily, M., Verma, A., Reinhartz, O., Figliola, R. S., Hsia, T.-Y., et al. (2019). Patient-specific multiscale modeling of the assisted bidirectional glenn. *Ann. Thorac. Surg.* 107, 1232–1239. doi: 10.1016/j.athoracsur.2018.10.024

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3:96ra77. doi: 10.1126/scitranslmed.3001318

South, A. M., Tomlinson, L., Edmonston, D., Hiremath, S., and Sparks, M. A. (2020). Controversies of renin–angiotensin system inhibition during the covid-19 pandemic. *Nat. Rev. Nephrol.* 16, 305–307. doi: 10.1038/s41581-020-0279-4

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Tang, N., Bai, H., Chen, X., Gong, J., Li, D., and Sun, Z. (2020). Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. *J. Thromb. Haemost.* 18, 1094–1099. doi: 10.1111/jth.14817

Tao, Y., Xu, J., Song, B., Xie, X., Gu, H., Liu, Q., et al. (2017). Short-term blood pressure variability and long-term blood pressure variability: which one is a reliable predictor for recurrent stroke. *Journal of Human Hypertension* 31, 568–573. doi: 10.1038/jhh.2017.32

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hachette.

Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J. Mach. Learn. Res.* 10:13.

Viñas, R., Andres-Terre, H., Lio, P., and Bryson, K. (2021). Adversarial generation of gene expression data. *Bioinformatics*. doi: 10.1093/bioinformatics/btab035. [Epub ahead of print].

Wang, M., Yu, L., Zheng, D., Gan, Q., Gai, Y., Ye, Z., et al. (2019). Deep graph library: towards efficient and scalable deep learning on graphs. *arXiv [Preprint]. arXiv:1909.01315*.

Wen, W., Luo, R., Tang, X., Tang, L., Huang, H. X., Wen, X., et al. (2015). Age-related progression of arterial stiffness and its elevated positive association with blood pressure in healthy people. *Atherosclerosis* 238, 147–152. doi: 10.1016/j.atherosclerosis.2014.10.089

Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., et al. (2010). Sharing health data for better outcomes on patientslikeme. *J. Med. Internet Res.* 12:e19. doi: 10.2196/jmir.1549

Wild, C. P. (2005). Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol. *Biomarkers Prev.* 14, 1847–1850. doi: 10.1158/1055-9965.EPI-05-0456

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 58:101552. doi: 10.1016/j.media.2019.101552

Yu, H., Miao, C., Leung, C., and White, T. J. (2017). Towards AI-powered personalization in MOOC learning. *NPJ Science of Learning* 2, 1–5. doi: 10.1038/s41539-017-0016-3

Zaman, M. A., Oparil, S., and Calhoun, D. A. (2002). Drugs targeting the renin–angiotensin–aldosterone system. *Nat. Rev. Drug Discov.* 1, 621–636. doi: 10.1038/nrd873

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., et al. (2018). Graph neural networks: a review of methods and applications. *arXiv [Preprint]. arXiv:1812.08434.*

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34, i457–i466. doi: 10.1093/bioinformatics/bty294

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership