frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Review
PUBLISHED 07 October 2022
DOI 10.3389/fbioe.2022.998298

Check for updates

OPEN ACCESS

EDITED BY
Ravindra Soni,
Indira Gandhi Krishi Vishva Vidyalaya,
India

REVIEWED BY
Lata Jain,
National Institute of Biotic Stress
Management, India
Xiao Long,
Peking Union Medical College Hospital
(CAMS), China

*CORRESPONDENCE
Zhichang Zhang,
zczhang@cmu.edu.cn

SPECIALTY SECTION
This article was submitted to Bioprocess Engineering,
a section of the journal
Frontiers in Bioengineering and Biotechnology

RECEIVED 19 July 2022
ACCEPTED 23 September 2022
PUBLISHED 07 October 2022

CITATION
Xu D, Liu B, Wang J and Zhang Z (2022), Bibliometric analysis of artificial intelligence for biotechnology and applied microbiology: Exploring research hotspots and frontiers.
Front. Bioeng. Biotechnol. 10:998298.
doi: 10.3389/fbioe.2022.998298

Bibliometric analysis of artificial intelligence for biotechnology and applied microbiology: Exploring research hotspots and frontiers

Dongyu Xu[1], Bing Liu[2], Jian Wang[3] and Zhichang Zhang[1]*

[1]Department of Computer, School of Intelligent Medicine, China Medical University, Shenyang, Liaoning, China, [2]Department of Bone Oncology, The People's Hospital of Liaoning Province, Shenyang, Liaoning, China, [3]Department of Pathogenic Biology, School of Basic Medicine, China Medical University, Shenyang, Liaoning, China

Background: In the biotechnology and applied microbiology sectors, artificial intelligence (AI) has been extensively used in disease diagnostics, drug research and development, functional genomics, biomarker recognition, and medical imaging diagnostics. In our study, from 2000 to 2021, science publications focusing on AI in biotechnology were reviewed, and quantitative, qualitative, and modeling analyses were performed.

Methods: On 6 May 2022, the Web of Science Core Collection (WoSCC) was screened for AI applications in biotechnology and applied microbiology; 3,529 studies were identified between 2000 and 2022, and analyzed. The following information was collected: publication, country or region, references, knowledgebase, institution, keywords, journal name, and research hotspots, and examined using VOSviewer and CiteSpace V bibliometric platforms.

Results: We showed that 128 countries published articles related to AI in biotechnology and applied microbiology; the United States had the most publications. In addition, 584 global institutions contributed to publications, with the Chinese Academy of Science publishing the most. Reference clusters from studies were categorized into ten headings: deep learning, prediction, support vector machines (SVM), object detection, feature representation, synthetic biology, amyloid, human microRNA precursors, systems biology, and single cell RNA-Sequencing. Research frontier keywords were represented by microRNA (2012–2020) and protein-protein interactions (PPIs) (2012–2020).

Abbreviations: AI, Artificial Intelligence; WoSCC, Web of Science Core Collection; CPI, Compound-Protein Interactions; ACPred-FL, Anti-Cancer Peptide Predictor with Feature Representation Learning; ACPs, Anti-cancer peptides; CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; AD, Alzheimer's Disease; CNN, Convolutional Neural Network; Aβ, amyloid-β; RNN, Recurrent Neural Network; RNA, Ribonucleic Acid; DNN, Deep Neural Networks; DISC, Deep learning Imputation model with semi-supervised learning for Single Cell transcriptomes; PPI, Protein-Protein Interaction; DPPI, Direct physical Protein-Protein Interactions; SVM, Support Vector Machine.

Conclusion: We systematically, objectively, and comprehensively analyzed AI-related biotechnology and applied microbiology literature, and additionally, identified current hot spots and future trends in this area. Our review provides researchers with a comprehensive overview of the dynamic evolution of AI in biotechnology and applied microbiology and identifies future key research areas.

KEYWORDS

artificial intelligence, biotechnology, bibliometric, deep learning, machine learning, applied microbiology

Introduction

Since the beginning of the 21st century, the life sciences and biotechnology and applied microbiology sectors have exemplified mankind's technological and revolutionary evolution. In these sectors, among the top 10 scientific breakthroughs published by science journals in recent decades, more than half of research outputs were revolutionarily innovative and breakthrough in nature. Emerging biological sectors include, biomedicine, bio-based chemicals, bioenergy, and genetically modified crop technology (Celikkanat Ozan and Baran, 2013). These areas are cutting-edge, and next-generation biotechnology industries are anticipated to develop rapidly in the future (Lee, 2016). As the front end of these biological industries and value chains, biotechnology and applied microbiology research has adopted a leading position in these industries. Therefore, exploring rapid developments and hot trends in basic biotechnology and applied microbiology research is pivotal in guiding biotechnology current achievements and developing new, downstream bio-industry markets.

AI represents advanced computer technology, and is a highly complex system integrating mathematics, statistics, probability, logic, ethics, and other disciplines. It primarily includes deep learning, machine learning, convolution and recurrent neural networks (CNN and RNN, respectively), full revolutionary networks (FCNs), and other specific methods. AI is extensively used in different industries, in particular biotechnology and the life sciences. In recent years, several major research developments have been achieved, including the AI-mediated prediction of protein structure, which was breakthrough of the year in 2021 (Baek et al., 2021). By exploiting complex simulation algorithms, AI has revolutionized disease diagnostics, drug research and development, functional genomics, biomarker recognition, and medical imaging diagnostics, and critically, has provided a vital reference point for disease diagnostic, prediction, and treatment strategies (Dlamini et al., 2020).

To facilitate AI research and progress in biotechnology and applied microbiology, bibliometric analyses and reviews are used to equip scientists with in-depth understandings of the application, its ongoing evolution, and future prospects. From a database search spanning 1 January 2000 to 31 December 2021, we used bibliometric methods to analyze scientific papers on AI applications in biotechnology and applied microbiology, including papers published in different jurisdictions and by institutions. We examined journals where AI biotechnological research studies were published, investigated the "top 10 cited studies", and enumerated how many times popular studies were cited. We clustered the reference network of cited studies, and investigated the subject knowledge base. Research hotspots were identified using burst keywords, which provided invaluable indicators for future research. Our research remit was to provide researchers with a macro understanding and micro analysis of the AI biotechnological field. When compared with traditional systematic reviews, we provided an intuitive, timely, and logical framework to track biotechnological developments and explore specific knowledge areas.

Methods

On 6 January 2022, we used the Web of Science Core Collection (WoSCC) to download data (2000–2021), which were independently verified by DX and ZZ. The following search terms were used: ("deep learning" OR "machine learning" OR "convolutional neural network*" OR CNN* OR RNN OR "Recurrent neural network*" OR "Fully Convolutional Network*" OR FCN*). The Web of Science category was "Biotechnology Applied Microbiology", and documents were gathered. From studies, the following basic information was gathered: authors, abstract, title, institution, journal, keywords, country/region, and references. Studies indexed in the database were included, whereas the following were excluded: 1) book chapters, data papers, meeting abstracts and proceedings papers, repeated articles, and editorials, and 2) unpublished studies with limited data for analysis. In total, 79 duplicates were excluded. A study overview (search process and analyses) is provided (Figure 1).

We described publication traits, including country, institute, journals, and keywords. The H-index is an important indicator and was used to reflect the value of scientific research (Eyre-Walker and Stoletzki, 2013). The

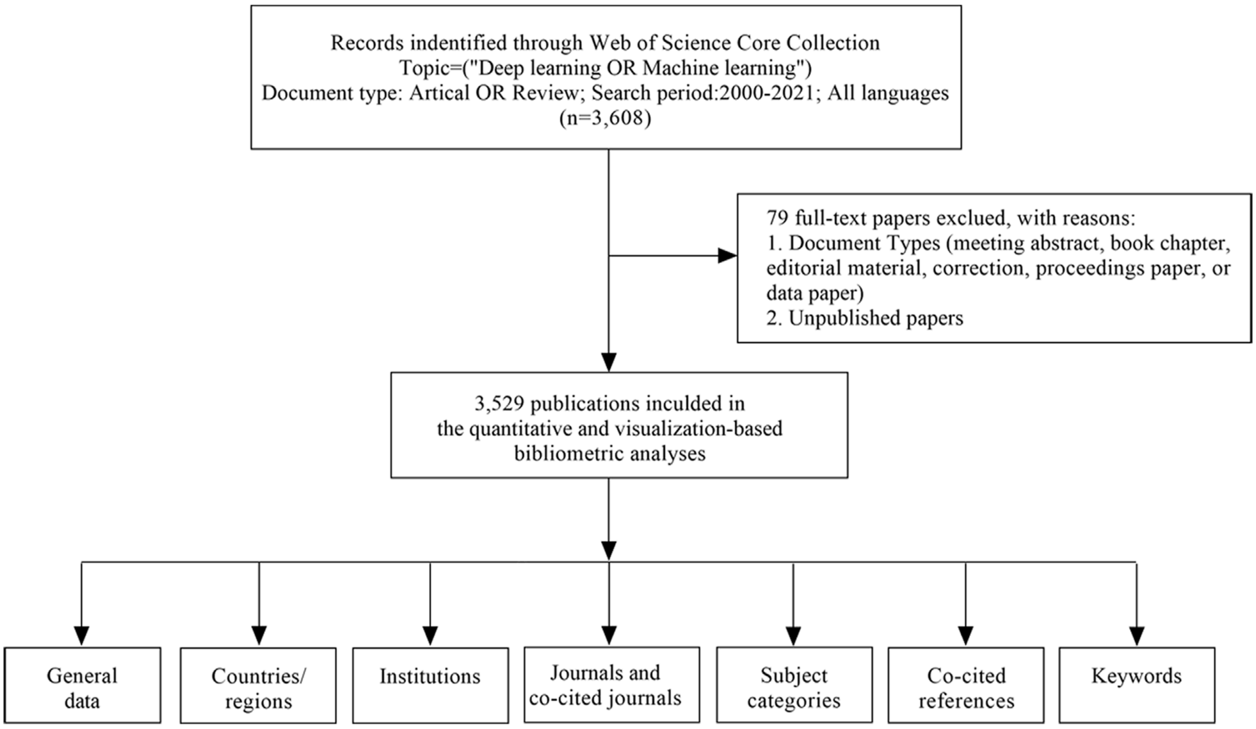

Records indentified through Web of Science Core Collection
Topic=("Deep learning OR Machine learning")
Document type: Artical OR Review; Search period:2000-2021; All languages
(n=3,608)
79 full-text papers exclued, with reasons:
1. Document Types (meeting abstract, book chapter, editorial material, correction, proceedings paper, or data paper)
2. Unpublished papers
3,529 publications inculded in the quantitative and visualization-based bibliometric analyses
General data
Countries/ regions
Institutions
Journals and co-cited journals
Subject categories
Co-cited references
Keywords

FIGURE 1
A frame flow diagram. The diagram showed details selection criteria for ABAM publications from WoSCC database and the steps of bibliometric analysis.

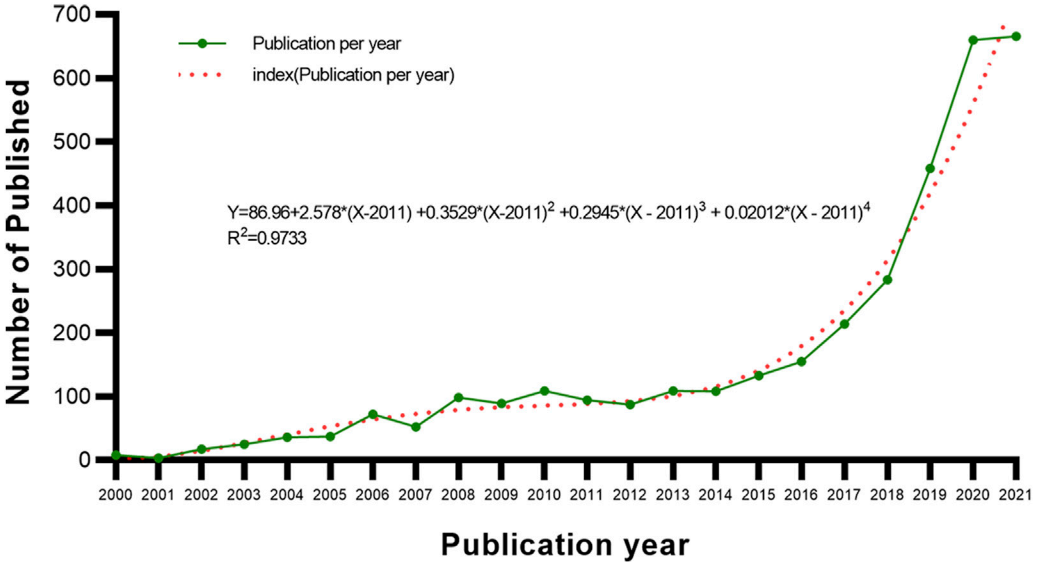


FIGURE 2
Trends in the number of publications on ABAM from 2000 to 2021.

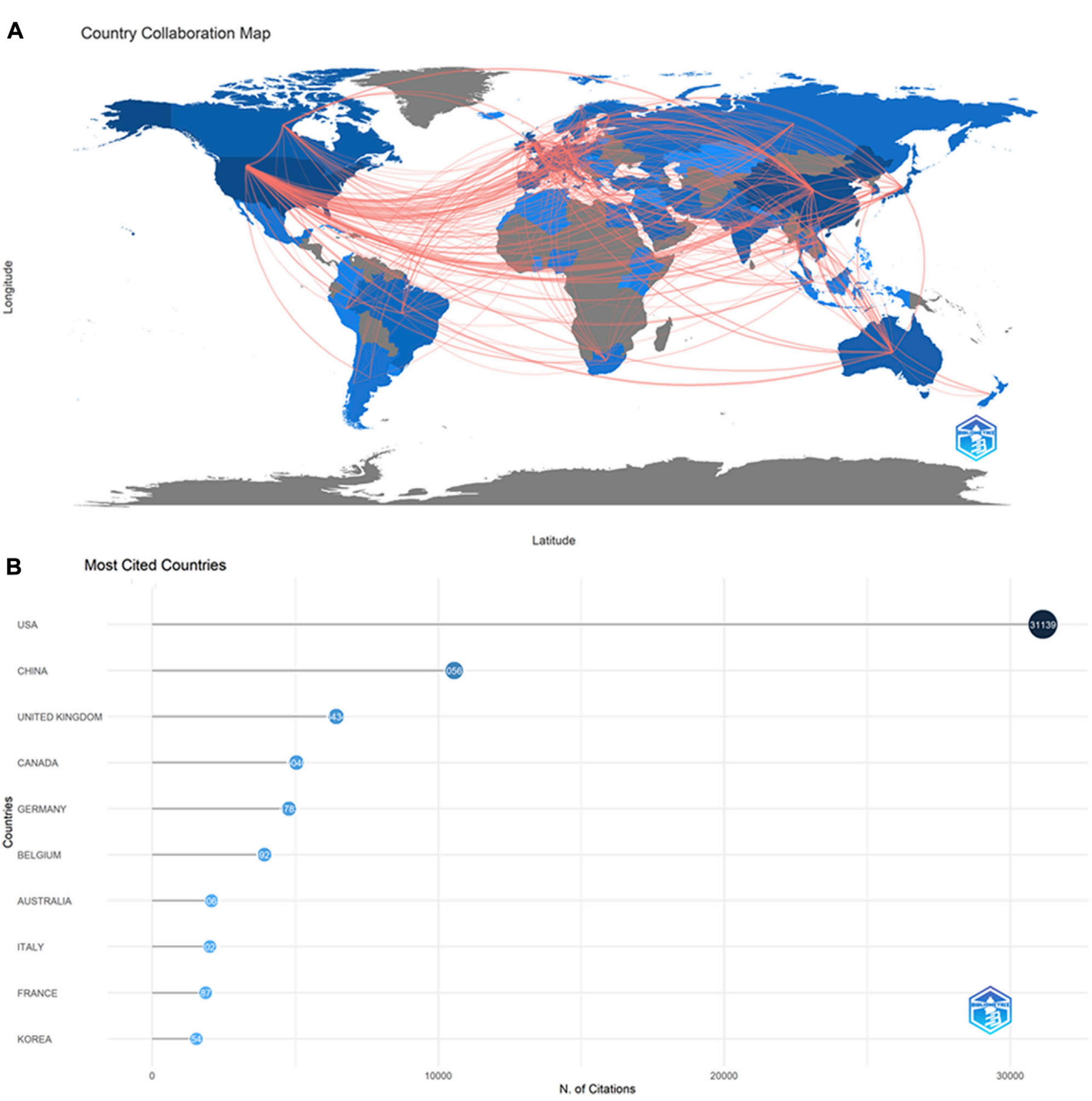


FIGURE 3
The cooperation of countries/regions contributed to publications. **(A)** Country Collaboration map. **(B)**Most Cited Countries.

Literature Metrology websites; http://bibliometric.com/, VOSviewer (Leiden University, Leiden, Netherlands), and CiteSpace V (Drexel University, Philadelphia, PA, United States) were used to visualize collaborative networks in institutes/countries/keywords/journals and co-occurrence analyses. In CiteSpace, we conducted reference co-citation analyses, constructed knowledge maps, and identified burst keywords to generate new recurrent keywords (Chen, 2006).

Results

Article distribution by publication year

The literature retrieval showed that the research on AI in this topic began in 2000. From 2000 to 2021, 3,529 papers were published, and AI with Biotechnology and Applied Microbiology (ABAM) related publication trends identified (Figure 2). Studies

TABLE 1 Top 10 countries/regions and relevant institutions.

Rank	Countries/regions	Count	Total citations	H-index	Institutions	Count	H-index
1	United States	1308	31139	91	CHINESE ACAD SCI	85	26
2	China	826	10561	61	STANFORD UNIV	52	32
3	Germany	258	4783	44	SHANGHAI JIAO TONG UNIV	46	17
4	United Kingdom	223	6434	49	UNIV CAMBRIDGE	37	16
5	Canada	158	5040	37	CARNEGIE MELLON UNIV	36	16
6	Australia	128	2068	27	TSINGHUA UNIV	32	18
7	Italy	121	2022	28	UNIV ELECT SCI & TECHNOL CHINA	31	16
8	Japan	121	1327	30	HARVARD UNIV	31	30
9	South Korea	120	1547	25	TIANJIN UNIV	30	17
10	France	99	1875	29	UNIV WASHINGTON	30	17

TABLE 2 Top 10 cited references on artificial intelligence for biotechnology and applied microbiology.

Rank	Source titles	Title of reference	Count	Interpretation of findings
1	NATURE	Deep learning.	203	This paper studied the back propagation algorithm of deep learning
2	J MACH LEARN RES	Dropout: a simple way to prevent neural networks from overfitting.	176	This study improved neural network performance in supervised learning tasks
3	NAT BIOTECHNOL	Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning.	151	This study used deep learning techniques to identify sequence specificities in DNA and RNA binding proteins
4	INT C LEARNING R	Adam: A method for stochastic optimization.	132	In this paper, a stochastic gradient descent optimization algorithm, based on the first derivative, was proposed for the first time. It was used for large data, sparse data processing, and super parameter easy adjustment.
5	J MACH LEARN RES	Scikit-learn: Machine learning, in python.	126	This article introduced scikit learning, a python module that integrated different contemporary machine learning algorithms
6	NAT METHODS	Predicting effects of noncoding variants with deep learning-based sequence model.	118	Based on deep learning, this study developed an algorithmic framework to identify functional effects from noncoding mutations
7	ACM T INTEL SYST TEC	LIBSVM: A library for support vector machines.	91	This study helped users apply support vector machine (SVM) to their applications.
8	COMMUN ACM	ImageNet classification with deep convolutional neural networks.	90	In this study, a large-scale deep CNN was used to classify 1.2 million high-resolution images
9	NATURE	An integrated encyclopedia of DNA elements in the human genome.	81	This study systematically mapped chromatin structure, transcription, transcription factor association, and histone modification regions.
10	SIGKDD EXPLORATIONS	The WEKA data mining software: an update.	79	The widely used, open source machine learning software Weka was introduced in this paper and allowed researchers access the latest technologies in machine learning.

in this area are increasing year on year, and suggest the establishment of an important research trend.

Institutes, countries, and regions

We observed that 128 countries/regions published ABAM studies: collaborations between countries (Figure 3) and the top 10 countries (Table 1) are outlined. The United States published the most studies (1308), then China (826), Germany (258), and the United Kingdom (223). Some countries, such as United States, China, Germany, and United Kingdom, showed high centrality (marked by dark blue), indicating that these countries likely played an important role in research of this topic and made great contributions.

We identified 584 institutes which contributed to ABAM publications; the top 10 are outlined (Table 1). Institutional collaborations are shown (Figure 4). The Chinese Academy of Sciences recorded the most publications (85), followed by the universities of Stanford (52), Shanghai Jiao Tong (46), and Cambridge (37).

Figure 4 emphasizes the close and complex cooperative relationship between different organizations. The VOSviewer platform can be used to analyze the centrality of organizations. The purple circle represents centrality, and the

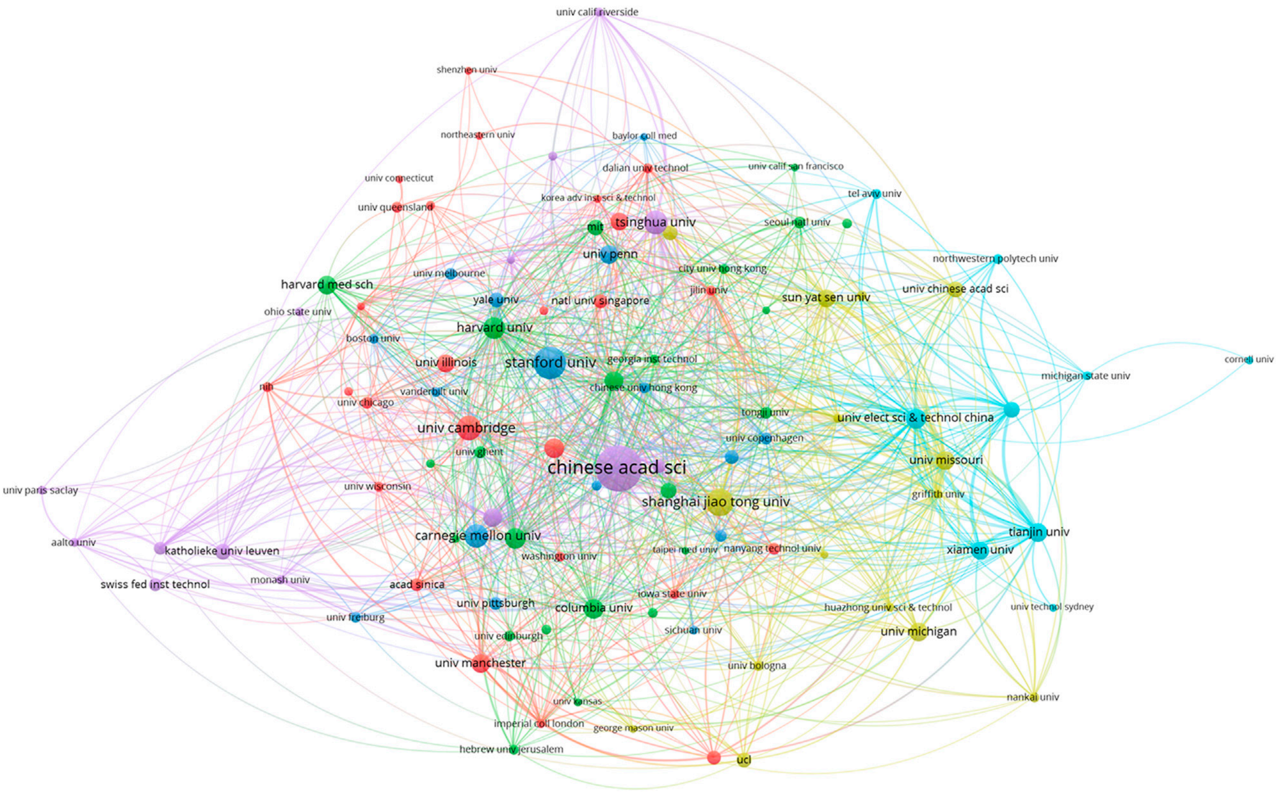


FIGURE 4
The cooperation of institutions contributed to publications.

TABLE 3 Highly link strength of the top 20 occurrence keywords.

Rank	Keyword	Occurrence	Total link strength	Rank	Keyword	Occurrence	Total link strength
1	Machine learning	782	420	11	Prediction	44	52
2	Deep learning	318	183	12	Biomarkers	39	49
3	Classification	82	95	13	Algorithms	36	40
4	Convolutional neural network	72	51	14	Bioinformatics	33	51
5	Artificial intelligence	60	61	15	Cancer	30	44
6	Random forest	57	58	16	Genomics	28	39
7	Support vector machine	56	41	17	Clustering	25	20
8	Feature selection	52	62	18	Data mining	25	29
9	Gene expression	50	61	19	Rna-seq	21	28
10	Neural networks	46	41	20	Natural language processing	19	16

area of the circle is proportional to the centrality. The Chinese Academy of Sciences and Stanford University are the most prominent organizations, showing that they conducted more research in this area.

Journals

In any research field, referential relationships between academic journals often reflect knowledge exchange, where

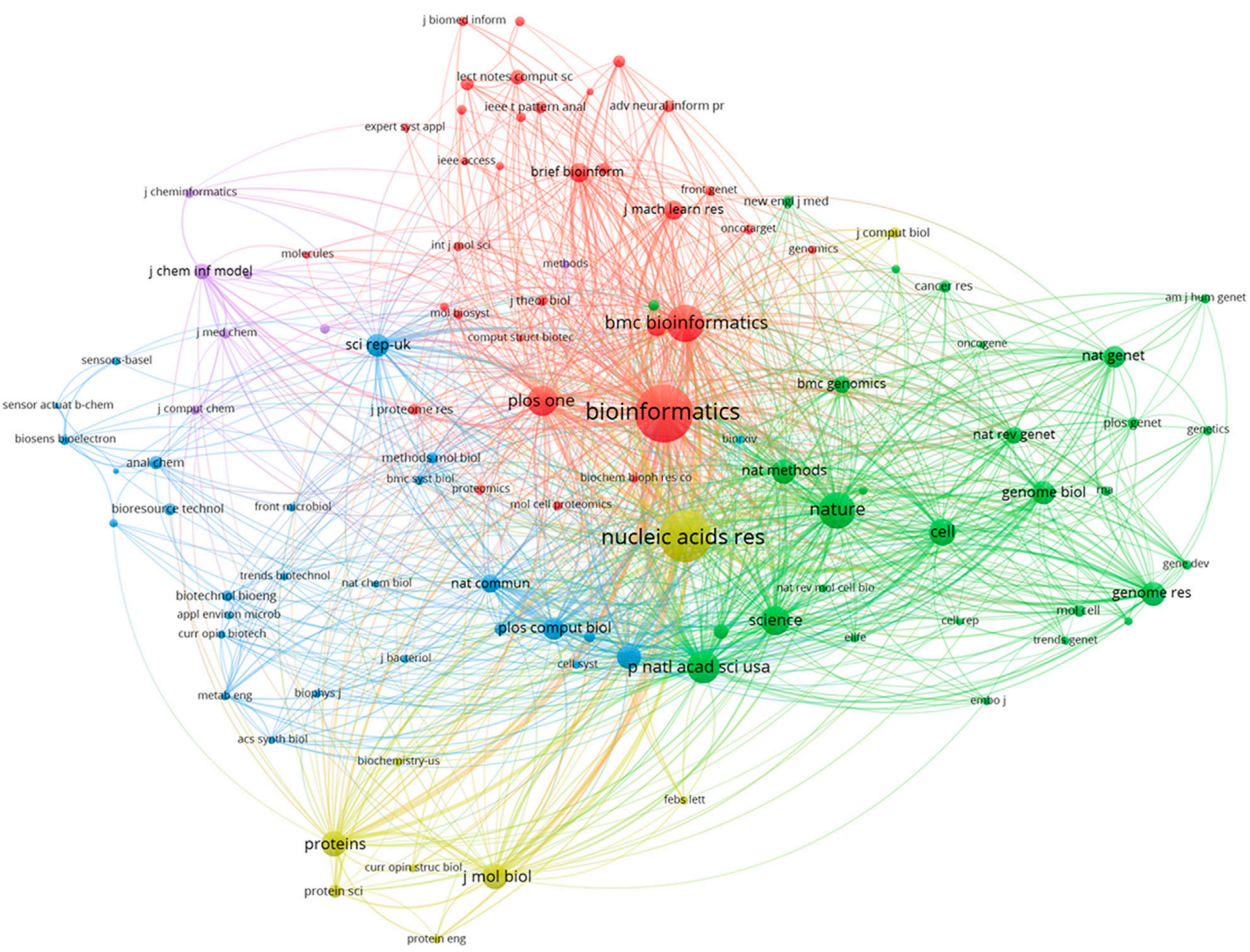


FIGURE 5
The network map of cited journals contributed to publications.

citing studies are knowledge frontiers, and referenced studies the knowledge basis. The top 10 references from studies (2000–2021) (Table 2) and collaborations between related journals (Figure 5) are outlined (Hall et al., 2009; Chang and Lin, 2011; Pedregosa et al., 2011; The ENCODE Project Consortium, 2012; Kingma and Ba, 2014; Srivastava et al., 2014; Alipanahi et al., 2015; LeCun et al., 2015; Zhou and Troyanskaya, 2015; Krizhevsky et al., 2017). Figure 5 shows that such journals as Bioinformatics, BMC Bioinformatics, Nature, Nucleic Acids Research, and PLoS One have higher centrality, and are the most popular journals for publishing research on this topic. The cooperative relationship between these journals is relatively balanced. This suggests that the research on the topic has aroused the interest of mainstream medicine and biology journals.

A dual-map overlay of journals (Figure 6) was used to show citing and cited journals on the left and right, respectively, while citation relationships were reflected by colored paths—these analyses showed that studies published in Genetics/Molecular/Biology journals were typically published in Biology/Molecular/Immunology journals.

References

Reference analysis is a vital bibliometric indicator; frequently cited studies typically and significantly influence their respective research fields. Using this approach, co-cited document-based clustering analyses generated connecting nodes and subfields for ABAM analyses.

We generated a co-citation reference network to measure the scientific relevance of related studies (Figure 7). Cluster setting parameters: top N% = 0.5 and # years per slice = 1. The Modularity Q score = 0.7135, which was > 0.5 and showed the network was reasonably separated into loosely coupled clusters. Weighted mean silhouette score = 0.9229, which was > 0.5, therefore cluster homogeneity was acceptable.

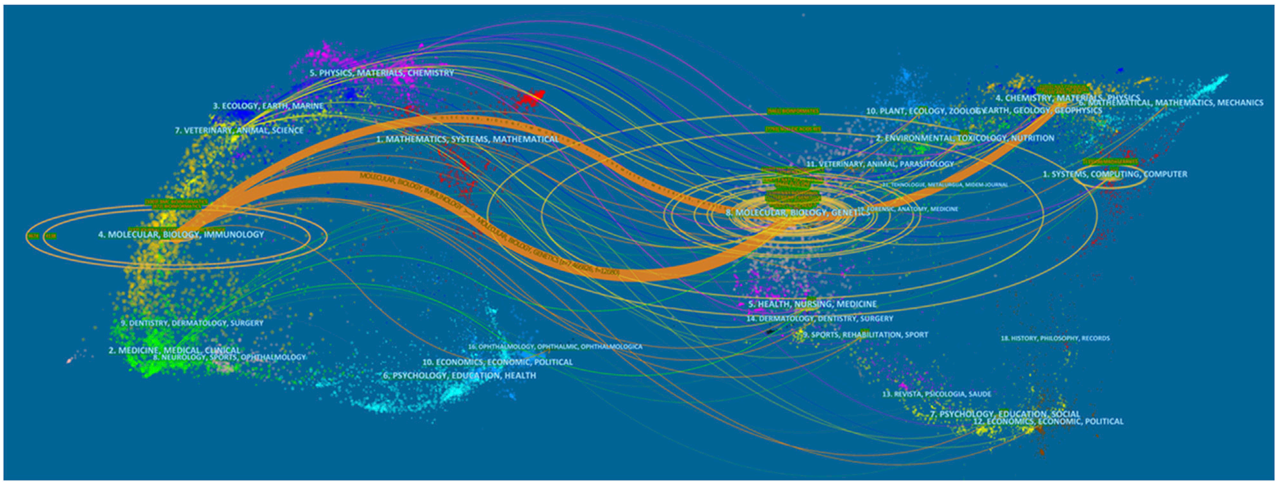

5. PHYSICS, MATERIALS, CHEMISTRY
3. ECOLOGY, EARTH, MARINE
7. VETERINARY, ANIMAL, SCIENCE
1. MATHEMATICS, SYSTEMS, MATHEMATICAL
4. MOLECULAR, BIOLOGY, IMMUNOLOGY
MOLECULAR, BIOLOGY, IMMUNOLOGY → MOLECULAR, BIOLOGY, GENETICS
9. DENTISTRY, DERMATOLOGY, SURGERY
2. MEDICINE, MEDICAL, CLINICAL
8. NEUROLOGY, SPORTS, OPHTHALMOLOGY
16. OPHTHALMOLOGY, OPHTHALMIC, OPHTHALMOLOGICA
10. ECONOMICS, ECONOMIC, POLITICAL
6. PSYCHOLOGY, EDUCATION, HEALTH
4. CHEMISTRY, MATERIALS, PHYSICS
6. MATHEMATICAL, MATHEMATICS, MECHANICS
10. PLANT, ECOLOGY, ZOOLOGY
3. EARTH, GEOLOGY, GEOPHYSICS
2. ENVIRONMENTAL, TOXICOLOGY, NUTRITION
11. VETERINARY, ANIMAL, PARASITOLOGY
1. SYSTEMS, COMPUTING, COMPUTER
15. TEHNOLOGIJE, METALURGIJA, MIDEM-JOURNAL
8. MOLECULAR, BIOLOGY, GENETICS
19. FORENSIC, ANATOMY, MEDICINE
5. HEALTH, NURSING, MEDICINE
14. DERMATOLOGY, DENTISTRY, SURGERY
9. SPORTS, REHABILITATION, SPORT
18. HISTORY, PHILOSOPHY, RECORDS
13. REVISTA, PSICOLOGIA, SAUDE
7. PSYCHOLOGY, EDUCATION, SOCIAL
12. ECONOMICS, ECONOMIC, POLITICAL

FIGURE 6
The dual-map overlay of journals contributed to publications.

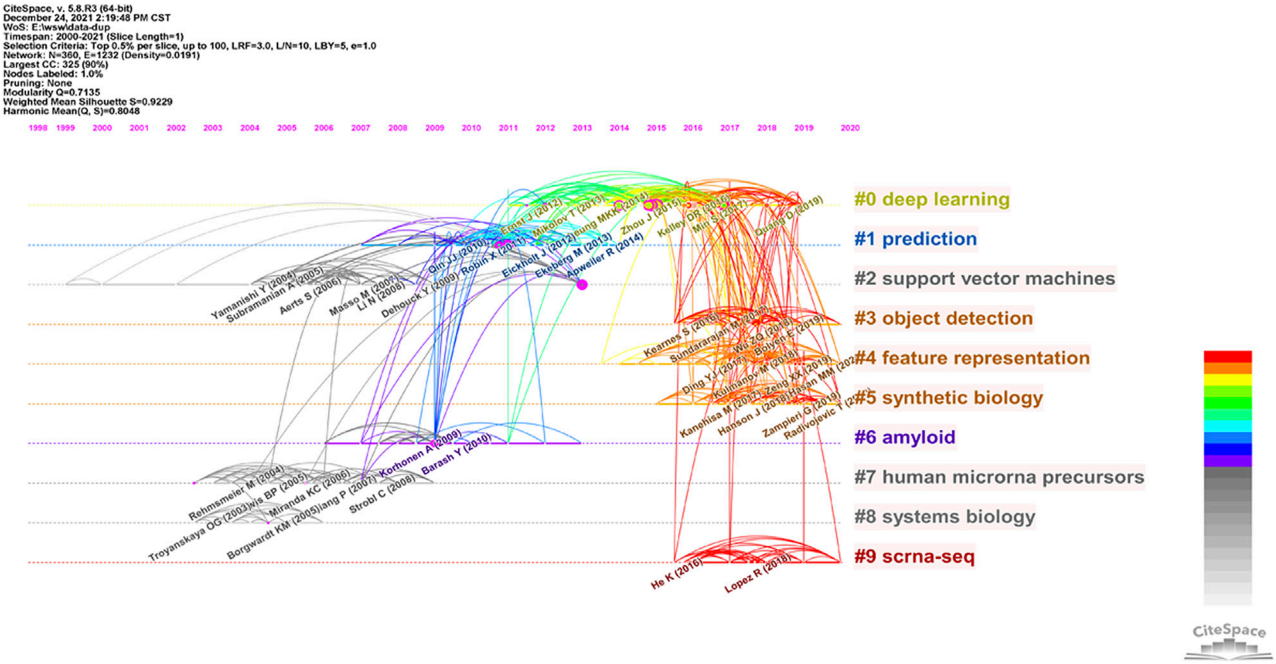
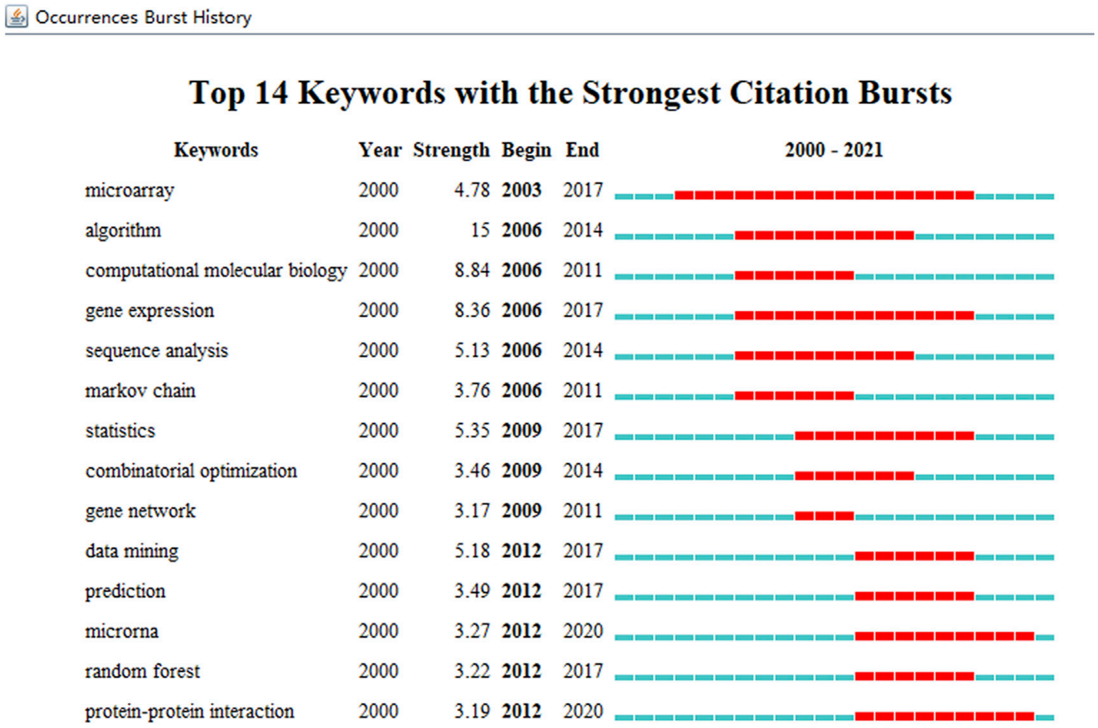


FIGURE 7
Reference co-citation map of publications on ABAM from 2000 to 2021.

Index items, as cluster markers, were extracted from studies. The largest cluster #0 was "deep learning" (Alipanahi et al., 2015), cluster #1 "prediction" (Kourou et al., 2015), cluster #2 "support vector machines (SVM)" (Furey et al., 2000), cluster #3 "object detection" (Hung et al., 2020), cluster #4 "feature representation" (Manavalan et al., 2019), cluster #5 "synthetic biology" (Wu et al., 2016), cluster #6 "amyloid" (Charoenkwan et al., 2021), cluster #7 "human microRNA precursors" (Wang et al., 2011), cluster #8 "systems biology" (Zou et al., 2015b), and cluster #9 "scRNA-Seq (single cell RNA-Sequencing)" (Arisdakessian et al., 2019).



Top 14 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2000 - 2021
microarray	2000	4.78	**2003**	2017	
algorithm	2000	15	**2006**	2014	
computational molecular biology	2000	8.84	**2006**	2011	
gene expression	2000	8.36	**2006**	2017	
sequence analysis	2000	5.13	**2006**	2014	
markov chain	2000	3.76	**2006**	2011	
statistics	2000	5.35	**2009**	2017	
combinatorial optimization	2000	3.46	**2009**	2014	
gene network	2000	3.17	**2009**	2011	
data mining	2000	5.18	**2012**	2017	
prediction	2000	3.49	**2012**	2017	
microrna	2000	3.27	**2012**	2020	
random forest	2000	3.22	**2012**	2017	
protein-protein interaction	2000	3.19	**2012**	2020	

FIGURE 8
The keywords with the strongest citation bursts of publications on ABAM from 2000 to 2021.

Keywords

Analysis of keywords can provide a summary of the topics of each study and explore the hotpots and directions in this research area.

Keywords extracted from ABAM studies were processed, and the top 20 are given in Table 3. Temporal hotspot trend shifts, on the basis of the top 14 keywords with the strongest citation bursts, were analyzed and included the following. The burst keywords in 2006–2011 were computational molecular biology (2006–2011), Markov chain (2006–2011), and gene network (2009–2011). The burst keywords in 2006–2014 were algorithm (2006–2014), sequence analysis (2006–2014), and combinatorial optimization (2009–2014). The burst keywords in 2003–2017 were microarray (2003–2017), gene expression (2006–2017), statistics (2009–2017), data mining (2012–2017), prediction (2012–2017), and random forest (2012–2017). The current research hotspots are microRNA (2012–2020) and protein–protein interaction (2012–2020) (Figure 8).

Discussion

General data

In this study, 3,529 ABAM papers, confirming to search terms and inclusion/exclusion criteria, were published between 2000 and 2021. The United States published most studies (1308, 26.6%), with China second (826, 16.8%). China had five of the top 10 institutions, with four in the United States, and one in the United Kingdom. The journal in which most publications were published was Bioinformatics, which majorly contributed to ABAM research. Additionally, the top 10 cited studies were examined: the top study was cited 203 times and was published by LeCun et al. in NATURE (LeCun et al., 2015). The second rated study was cited 176 times and published by Srivastava et al. in J MACH LEARN RES (Srivastava et al., 2014).

Knowledge base

From previous studies, the application of deep learning related technologies to microbiology and biotechnology has been significant and generated many research achievements. As indicated (Figure 6), when we clustered co-cited references, key clustering nodes identified knowledge bases in this research field: #0 “deep learning”, #1 “prediction”, #2 “SVM”, #3 “object detection”, #4 “feature representation”, #5 “synthetic biology”, #6 “amyloid”, #7 “human microRNA precursors”, #8 “systems biology”, and

#9 "scRNA-Seq". Herein, we describe the knowledge bases according to different clusters.

In #0 "deep learning", a DeepBind software tool, based on deep learning, was developed by Alipanahi et al. (2015), and identified DNA and RNA binding protein sequence specificity. The tool was used to develop regulatory process models in biological systems and identify pathogenic variants. In other work, Beck et al. (2020) generated a deep learning-based, pre-trained, drug-target interaction model, Molecule Transformer-Drug Target Interaction, which identified commercially available drugs targeting SARS-CoV-2 proteins.

In #1 "prediction", Tsubaki et al. (2019) studied end-to-end representation learning of compounds and proteins, and developed a Compound-Protein Interactions (CPI) prediction strategy for virtual screening in drug discovery by combining protein convolution neural networks (CNN) and compound graph neural networks. In other work, Almagro Armenteros et al. (2017) developed a prediction algorithm based on deep neural networks which relied only on sequence information for protein subcellular localization.

In #2 "SVM", Ozer et al. (2020) showed that SVM provided solutions for high-throughput data analyses and contextualization; the approach rapidly determined timelines for invasive cancer diagnostics and treatment, and provided solutions for biomedical, bioengineering, and clinical applications. In other work, a SVM technology model constructed by Zhang et al. (2018) used joint information from multiple bone turnover markers, which improved diagnostic efficiency for osteoporosis, almost in perfect agreement with the dual-energy X-ray absorptiometry.

In #3 "object detection", the approach by Zhang et al. (2020), exploited a deep object detection technique and was used to study contacts between protein secondary structure elements, and predict tertiary structural protein topology. Einhäuser et al. (2017) developed a foveal object detector to detect eye movement, which significantly reduced metabolic costs and computational complexity, and provided insights on visual system evolution with eye movement.

In #4 "feature representation", an effective feature representation learning model ACPred-FL was developed by Wei et al. (2018), and used to rapidly and accurately identify new Anti-cancer peptides (ACPs)in many candidate proteins and peptides. The learning method developed by Peng et al. (2020) was based on feature representation learning and deep neural network (DTI-CNN), and was used to predict drug-target interactions and reduce time and experimental costs. In other research, from deep representation learning features with 107 dimensions, Lv et al. (2020) devised a sub-Golgi protein localization identification method, which exploited one feature type to accurately predict sub-Golgi protein localization.

#5 "synthetic biology" is a logical extension of recombinant technology or genetic engineering fields (Katz et al., 2018). Using integrated synthetic biology, Nguyen et al. (2021) developed a wearable face-mask, with a lyophilized CRISPR sensor, to non-invasively detect SARS-CoV-2 at room temperature within 90 min. Cubillos-Ruiz et al. (2021) proposed that synthetic biology could be used to program living cells with therapeutic functions; their cell-based therapeutic design is currently undergoing rapid development in medicine, and may provide effective treatment solutions for human diseases.

In #6 "amyloid", Charoenkwan et al. (2021) generated the first scorecard-based predictor for the accurate analysis, prediction, characterization, and identification of amyloid, on a large scale, to generate functional information for therapeutic intervention strategies. Cerebral amyloid-β (Aβ) is an Alzheimer's disease (AD) trait. Machine learning methods were used to identify cognitive performance and demographic variables for noninvasive testing of Aβ deposition, which can detect the effect of anti-amyloid drugs in the non-dementia population (Ko et al., 2019).

In #7 "human microRNA precursors", Zheng et al. (2020) used CNN and RNN approaches to automatically extract complex RNA sequence features to efficiently detect and predict human pre microRNAs. Kamenetzky et al. (2016) identified a novel pre-microRNA in the *Echinococcus multilocularis* genome using a machine learning approach, which could help control and prevent the global zoonotic infectious disease alveolar echinococcosis.

In #8 "systems biology", Reel et al. (2021) integrated different machine learning prediction algorithms to analyze different omics data to identify new biomarkers for systems biology. In their research, Weiskittel et al. (2021) outlined how systems biology algorithms layer machine learning and biological components could provide system-level analyses of single-cell omics data to clarify complex biological mechanisms. The powerful combination of systems biology, single cell omics, and machine learning could promote further, beneficial biomedical research.

In #9 "scRNA-Seq", in an unbiased manner in single cells, scRNA-Seq assesses functions in individual cells and cell-to-cell variability (Lin et al., 2020). Based on deep neural networks, Arisdakessian et al. (2019) formulated an interpolation algorithm Deepimpute based on DNN. Dropout layers and loss function were used to learn data patterns and to deal with gaps in scRNA-Seq data. He et al. (2020) developed DISC, a deep learning imputation model with semi-supervised learning for single cell transcriptomes. DISC can deduce gene expression and structures obscured by dropouts, enhanced gene and cell structures, recovered poor gene expression, and improved cell identification. Using machine learning methods (deep learning) combined with scRNA-Seq datasets, issues such as reducing dimensions, missing values, denoizing sc data, and explaining zero expansion, can be solved. Machine learning methods can be exploited to comprehensively process scRNA-Seq data, improve follow-up analyses in stem cells, identify cell subsets, and support regenerative medicine and cell therapy strategies (Yan et al., 2021).

Research frontiers and hotspots

Typically, keywords are used to concentrate on contemporary research concepts, while burst keywords represent research frontiers and emerging trends. CiteSpace was used to capture burst keywords, from which two research frontiers were identified: microRNA (2012–2020) and Protein-Protein Interaction (PPIs) (2012–2020). Importantly, we hypothesize these keywords exemplify future research frontiers.

MicroRNAs are noncoding single stranded RNAs that regulate development and gene transcription. Predicting and identifying connections between miRNAs and disease using AI-related methods is highly significant for unraveling pathogenic, preventative, prognostic, and pathological mechanisms implicated in diseases.

Zou et al. (2015a) predicted correlations between microRNAs and disease using two approaches: KATZ combined social network analysis and machine learning, while CATAPULT was a supervised machine learning method. Both were applied to 242 known associations between microRNAs and disease, and used 3-fold cross validation and leave-one-out cross-validation to evaluate method performance.

Wen et al. (2018) used the deep learning-based approach DeepMir Tar and extracted 750 features from a relatively large data set at different levels to predict human miRNA target sites. DeepMir Tar provided a new way to reveal miRNA biological function, as well as gene therapy and drug discovery for human diseases.

In large-scale RNA sequencing studies, Liu *et al.* developed a computational model called MirTarget which predicted genome-wide miRNA targets. Machine learning methods were used to train miRNA targeting feature data with miRNA binding and target down-regulation features, thus MirTarget showed better performances when compared with other algorithms (Liu and Wang, 2019).

Zheng et al. (2020) used CNN and RNN models to predict human pre-miRNAs; sequences were combined with predicted pre-miRNA secondary structures as input features to avoid feature extraction and selection processes by hand. Models were easily trained for handling training datasets; they demonstrated low generalization errors and were satisfactory for test datasets (Zheng et al., 2020).

Protein–protein interactions are very important in such cell life activities as transcriptional regulation, signal transduction, and drug signal transduction. Study of PPIs has become a research hotspot in bioinformatics. However, it is time-consuming and costly to identify PPIs using experimental methods (Chen et al., 2019).

People are more inclined to use artificial intelligence methods, like machine-learning, to automatically identify PPIs, which helps understanding of the molecular roots of disease on one hand, and provides new ideas for drug research and development on the other hand. Also, this effectively reduces experimental costs (Yu et al., 2021).

Based on a deep learning algorithm, Sun et al. (2017) designed a stacked autoencoder and investigated sequence-based PPIs predictions; the prediction accuracy of different external datasets was 87.99%–99.21%. These high-throughput methods increased our understanding of protein roles, disease etiology, and therapy design.

Hashemifar et al. (2018) developed a Direct Physical Protein-Protein Interactions (DPPI) deep learning framework, which modeled and predicted PPIs from sequence information. By adopting a deep, Siamese-like CNN which used high-quality experimental PPI data, evolutionary information from a predicted protein pair, and combined these data with random projection and data enhancement, PPIs were successfully predicted (Hashemifar et al., 2018).

Zeng et al. (2019) formulated DeepPPISP, a novel end-to-end deep learning framework. To examine local contextual features, authors used a sliding window to acquire neighbor features from target amino acids. To analyze global sequence features, a text CNN extracted features from protein sequences. To predict PPI sites, local contextual and global sequence characteristics were combined (Zeng et al., 2019).

Sequence-based deep learning technologies have been successfully used to predict PPIs. However, Yang et al. (2020) indicted these methods only focus on sequence information and ignore structural information in PPI networks. Such information, including degree, location, and adjacent nodes in graphs, are vital for PPI predictions. Theses authors generated a graph-based deep learning method for predicting PPIs, and demonstrated an accuracy of 99.15%, which improved on existing sequence-based methods (Yang et al., 2020).

In their method based on deep learning, Liu-Wei et al. (2021) developed deepviral, which predicted PPIs between humans and viruses. The method processed protein sequences and phenotypic characteristics to reveal infectious disease mechanisms and elucidate potential treatment methods (Liu-Wei et al., 2021).

Conclusion

We generated an objective, systematic, and comprehensive bibliometric analysis of scientific studies associated with deep learning, machine learning, CNN, RNN, and FCNs in ABAM. Moreover, we identified the research basis, future trends, and current hotspots in this field. Identified knowledge bases were: deep learning, prediction, SVMs, object detection, feature representation, synthetic biology, amyloid, human microRNA precursors, systems biology, and scRNA-Seq. Furthermore, microRNAs and PPIs were identified as future research frontiers and trends.

We identified some study limitations; publications over an extended period (2000–2021) were gathered, therefore, some studies were incomplete and may have introduced publication bias into our research, potentially affecting analysis outcomes.

Author contributions

DX and ZZ conceived and designed the study; ZZ and DX wrote and integrated sections; BL and JW retrieved references. All authors read and approved the final manuscript.

Funding

Grants, supporting this study, were funded to Dongyu Xu (2021-AFCEC-529) from the Association of Fundamental Computing Education in Chinese Universities (AFCEC).

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.

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, 831–838. doi:10.1038/nbt.3300

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., Winther, O., and Hancock, J. (2017). DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 4049–3395. doi:10.1093/bioinformatics/btx548

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L. X. (2019). DeepImpute: An accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol.* 20, 211. doi:10.1186/s13059-019-1837-6

Baek, M., Dimaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754

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 (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* 18, 784–790. doi:10.1016/j.csbj.2020.03.025

Celikkanat Ozan, D., and Baran, Y. (2013). Comparative development of knowledge-based bioeconomy in the European Union and Turkey. *Crit. Rev. Biotechnol.* 34, 269–280. doi:10.3109/07388551.2013.792771

Chang, C.-C., and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intelligent Syst. Technol.* 2, 1–27. doi:10.1145/1961189.1961199

Charoenkwan, P., Kanthawong, S., Nantasenamat, C., Hasan, M. M., and Shoombuatong, W. (2021). iAMY-SCM: Improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. *Genomics* 113, 689–698. doi:10.1016/j.ygeno.2020.09.065

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inf. Sci. Technol.* 57, 359–377. doi:10.1002/asi.20317

Chen, C., Zhang, Q., Ma, Q., and Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom. Intelligent Laboratory Syst.* 191, 54–64. doi:10.1016/j.chemolab.2019.06.003

Cubillos-Ruiz, A., Guo, T., Sokolovska, A., Miller, P. F., Collins, J. J., Lu, T. K., et al. (2021). Engineering living therapeutics with synthetic biology. *Nat. Rev. Drug Discov.* 20, 941–960. doi:10.1038/s41573-021-00285-3

Dlamini, Z., Francies, F. Z., Hull, R., and Marima, R. (2020). Artificial intelligence (AI) and big data in cancer and precision oncology. *Comput. Struct. Biotechnol. J.* 18, 2300–2311. doi:10.1016/j.csbj.2020.08.019

Einhäuser, W., Akbas, E., and Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLoS Comput. Biol.* 13, e1005743. doi:10.1371/journal.pcbi.1005743

Eyre-Walker, A., and Stoletzki, N. (2013). The assessment of science: The relative merits of post-publication review, the impact factor, and the number of citations. *PLoS Biol.* 11, e1001675. doi:10.1371/journal.pbio.1001675

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914. doi:10.1093/bioinformatics/16.10.906

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. *SIGKDD Explor. Newsl.* 11, 10–18. doi:10.1145/1656274.1656278

Hashemifar, S., Neyshabur, B., Khan, A. A., and Xu, J. (2018). Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* 34, i802–i810. doi:10.1093/bioinformatics/bty573

He, Y., Yuan, H., Wu, C., and Xie, Z. (2020). Disc: A highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. *Genome Biol.* 21, 170. doi:10.1186/s13059-020-02083-3

Hung, J., Goodman, A., Ravel, D., Lopes, S. C. P., Rangel, G. W., Nery, O. A., et al. (2020). Keras R-CNN: Library for cell detection in biological images using deep neural networks. *BMC Bioinforma.* 21, 300. doi:10.1186/s12859-020-03635-x

Kamenetzky, L., Stegmayer, G., Maldonado, L., Macchiaroli, N., Yones, C., and Milone, D. H. (2016). MicroRNA discovery in the human parasite Echinococcus multilocularis from genome-wide data. *Genomics* 107, 274–280. doi:10.1016/j.ygeno.2016.04.002

Katz, L., Chen, Y. Y., Gonzalez, R., Peterson, T. C., Zhao, H., and Baltz, R. H. (2018). Synthetic biology advances and applications in the biotechnology industry: A perspective. *J. Industrial Microbiol. Biotechnol.* 45, 449–461. doi:10.1007/s10295-018-2056-y

Kingma, D. P., and Ba, J. (2014). *Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.*

Ko, H., Ihm, J.-J., and Kim, H.-G. (2019). Cognitive profiling related to cerebral amyloid beta burden using machine learning approaches. *Front. Aging Neurosci.* 11, 95. doi:10.3389/fnagi.2019.00095

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi:10.1145/3065386

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539

Lee, D.-H. (2016). Bio-based economies in Asia: Economic analysis of development of bio-based industry in China, India, Japan, Korea, Malaysia and Taiwan. *Int. J. Hydrogen Energy* 41, 4333–4346. doi:10.1016/j.ijhydene.2015.10.048

Lin, E., Mukherjee, S., and Kannan, S. (2020). A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. *BMC Bioinforma.* 21, 64. doi:10.1186/s12859-020-3401-5

Liu, W., and Wang, X. (2019). Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol.* 20, 18. doi:10.1186/s13059-019-1629-z

Liu-Wei, W., Kafkas, Ş., Chen, J., Dimonaco, N. J., Tegnér, J., Hoehndorf, R., et al. (2021). DeepViral: Prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes. *Bioinformatics* 37, 2722–2729. doi:10.1093/bioinformatics/btab147

Lv, Z., Wang, P., Zou, Q., Jiang, Q., and Xu, J. (2020). Identification of sub-Golgi protein localization by use of deep representation learning features. *Bioinformatics* 36, 5600–5609. doi:10.1093/bioinformatics/btaa1074

Manavalan, B., Basith, S., Shin, T. H., Wei, L., Lee, G., and Hancock, J. (2019). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi:10.1093/bioinformatics/bty1047

Nguyen, P. Q., Soenksen, L. R., Donghia, N. M., Angenent-Mari, N. M., De Puig, H., Huang, A., et al. (2021). Wearable materials with embedded synthetic biology sensors for biomolecule detection. *Nat. Biotechnol.* 39, 1366–1374. doi:10.1038/s41587-021-00950-3

Ozer, M. E., Sarica, P. O., and Arga, K. Y. (2020). New machine learning applications to accelerate personalized medicine in breast cancer: Rise of the support vector machines. *OMICS A J. Integr. Biol.* 24, 241–246. doi:10.1089/omi.2020.0001

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.

Peng, J., Li, J., and Shang, X. (2020). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinforma.* 21, 394. doi:10.1186/s12859-020-03677-1

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/j.biotechadv.2021.107739

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.

Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinforma.* 18, 277. doi:10.1186/s12859-017-1700-2

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

Tsubaki, M., Tomii, K., Sese, J., and Wren, J. (2019). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 35, 309–318. doi:10.1093/bioinformatics/bty535

Wang, Y., Chen, X., Jiang, W., Li, L., Li, W., Yang, L., et al. (2011). Predicting human microRNA precursors based on an optimized feature subset generated by GA–SVM. *Genomics* 98, 73–78. doi:10.1016/j.ygeno.2011.04.011

Wei, L., Zhou, C., Chen, H., Song, J., Su, R., and Hancock, J. (2018). ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi:10.1093/bioinformatics/bty451

Weiskittel, T. M., Correia, C., Yu, G. T., Ung, C. Y., Kaufmann, S. H., Billadeau, D. D., et al. (2021). The trifecta of single-cell, systems-biology, and machine-learning approaches. *Genes* 12, 1098. doi:10.3390/genes12071098

Wen, M., Cong, P., Zhang, Z., Lu, H., Li, T., and Kelso, J. (2018). DeepMirTar: A deep-learning approach for predicting human miRNA targets. *Bioinformatics* 34, 3781–3787. doi:10.1093/bioinformatics/bty424

Wu, G., Yan, Q., Jones, J. A., Tang, Y. J., Fong, S. S., and Koffas, M. a. G. (2016). Metabolic burden: Cornerstones in synthetic biology and metabolic engineering applications. *Trends Biotechnol.* 34, 652–664. doi:10.1016/j.tibtech.2016.02.010

Yan, R., Fan, C., Yin, Z., Wang, T., and Chen, X. (2021). Potential applications of deep learning in single-cell RNA sequencing analysis for cell therapy and regenerative medicine. *Stem Cells* 39, 511–521. doi:10.1002/stem.3336

Yang, F., Fan, K., Song, D., and Lin, H. (2020). Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. *BMC Bioinforma.* 21, 323. doi:10.1186/s12859-020-03646-8

Yu, B., Chen, C., Wang, X., Yu, Z., Ma, A., and Liu, B. (2021). Prediction of protein–protein interactions based on elastic net and deep forest. *Expert Syst. Appl.* 176, 114876. doi:10.1016/j.eswa.2021.114876

Zeng, M., Zhang, F., Wu, F.-X., Li, Y., Wang, J., Li, M., et al. (2019). Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 36, 1114–1120. doi:10.1093/bioinformatics/btz699

Zhang, Q., Zhu, J., Ju, F., Kong, L., Sun, S., Zheng, W.-M., et al. (2020). Issec: Inferring contacts among protein secondary structure elements using deep object detection. *BMC Bioinforma.* 21, 503. doi:10.1186/s12859-020-03793-y

Zhang, T., Liu, P., Zhang, Y., Wang, W., Lu, Y., Xi, M., et al. (2018). Combining information from multiple bone turnover markers as diagnostic indices for osteoporosis using support vector machines. *Biomarkers* 24, 120–126. doi:10.1080/1354750x.2018.1539767

Zheng, X., Fu, X., Wang, K., and Wang, M. (2020). Deep neural networks for human microRNA precursor detection. *BMC Bioinforma.* 21, 17. doi:10.1186/s12859-020-3339-7

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12, 931–934. doi:10.1038/nmeth.3547

Zou, Q., Li, J., Hong, Q., Lin, Z., Wu, Y., Shi, H., et al. (2015a). Prediction of MicroRNA-disease associations based on social network analysis methods. *BioMed Res. Int.* 2015, 1–9. doi:10.1155/2015/810514

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2015b). Similarity computation strategies in the microRNA-disease network: A survey. *Briefings Funct. Genomics*.