Deep learning methods for drug response prediction in cancer: Predominant and emerging trends

Alexander Partin[1]*, Thomas S. Brettin[1], Yitan Zhu[1],
Oleksandr Narykov[1], Austin Clyde[1], Jamie Overbeek[1] and
Rick L. Stevens[1,2]

[1]Division of Data Science and Learning, Argonne National Laboratory, Lemont, IL, United States, [2]Department
of Computer Science, The University of Chicago, Chicago, IL, United States

Cancer claims millions of lives yearly worldwide. While many therapies have been made available in recent years, by in large cancer remains unsolved. Exploiting computational predictive models to study and treat cancer holds great promise in improving drug development and personalized design of treatment plans, ultimately suppressing tumors, alleviating suffering, and prolonging lives of patients. A wave of recent papers demonstrates promising results in predicting cancer response to drug treatments while utilizing deep learning methods. These papers investigate diverse data representations, neural network architectures, learning methodologies, and evaluations schemes. However, deciphering promising predominant and emerging trends is difficult due to the variety of explored methods and lack of standardized framework for comparing drug response prediction models. To obtain a comprehensive landscape of deep learning methods, we conducted an extensive search and analysis of deep learning models that predict the response to single drug treatments. A total of 61 deep learning-based models have been curated, and summary plots were generated. Based on the analysis, observable patterns and prevalence of methods have been revealed. This review allows to better understand the current state of the field and identify major challenges and promising solution paths.

KEYWORDS

deep learning, drug sensitivity, multiomics, neural networks, precision medicine, precision oncology, personalized medicine, drug response prediction

1. Introduction

Cancer treatment response prediction is a problem of great importance for both clinical and pharmacological research communities. Many believe it will pave the way to devising more efficient treatment protocols for individual patients and provide insights into designing novel drugs that efficiently suppress disease. Currently, however, cancer treatment remains extremely challenging, often resulting in inconsistent outcomes. For example, tumor heterogeneity contributes to differential treatment responses in patients with the same tumor type (1, 2). Nevertheless, conventional tumor type-dependent anticancer treatments such as chemotherapy often lead to suboptimal results and substantial side effects, and therefore, are notoriously regarded as one-size-fits-all therapies (3, 4). Alternatively, targeted therapies and certain immunotherapies are prescribed based on known biomarkers, observable within individual patients (5). Cancer biomarkers refer to abnormalities in omics data (genomic, transcriptomic, etc.) which can be predictive of treatment response (3). Biomarker-driven treatment plans, either standalone or in combination with chemotherapies, are the mainstream of nowadays personalized (or precision) oncology. Discovery of biomarkers and their subsequent utilization

in clinical settings are attributed to advances in tumor profiling technologies and high-throughput drug screenings (3, 4).

An alternative direction to leverage large-scale screenings and high-dimensional omics data in the cancer research community is to build analytical models designed to predict the response of tumors to drug treatments. Typically, such models use tumor and drug information without explicitly specifying biomarkers (6). These models, often referred to as drug response prediction (DRP) models, can be used to prioritize treatments, explore drug repurposing, and reaffirm existing biomarkers. Artificial intelligence (AI) is the core methodology in designing DRP models, demonstrating encouraging results in retrospective evaluation analyzes with pre-clinical and clinical datasets. Many DRP models use classical machine learning (ML) and deep learning (DL), i.e., multi-layer neural networks (NNs). While DL is generally considered a subset of ML, we differentiate between the two terms, where ML is referred here to learning algorithms that do not involve the use of NNs.

Papers in this field are being published constantly, exploiting learning algorithms for DRP. To cope with increasing rate of publications, a recent special issue in Briefings in Bioinformatics was dedicated to DRP in cancer models (7). Collectively, at least 18 review-like papers have been published since 2020 in an attempt to summarize progress and challenges in the field, as well as provide discussions on promising research directions. Each paper aims to review the field from a unique perspective but certain topics substantially overlap as shown in Table 1 which lists some of the major topics and associated references. Common topics include data resources for constructing training and test datasets, prevalent data representations, ML and DL approaches for modeling drug response, and methods for evaluating the predictive performance of models.

Following the revival of artificial neural networks (NNs) more than a decade ago (29), DL methods have become a promising research direction across a variety of scientific and engineering disciplines (30–33). This trend is also observed in cancer research, including the prediction of tumor response to treatments, as shown in Figure 1. In 2013, Menden et al. demonstrated that a single hidden layer NN predicts drug response with comparable performance to random forest (RF) (34). The authors used genomic and response data from the Genomics of Drug Sensitivity in Cancer (GDSC) project which was published in 2012 (35). Despite the accelerated popularity of DL and the availability of substantial screening and omics data, it was not until the emergence of open-source frameworks dedicated for building and training NNs (36–38) that DL has become an integral part of DRP research (note the gap between 2013 and 2018 in Figure 1). Owing to the abundance of omics and screening data, the availability of DL frameworks and the unmet need for precision oncology, we have been witnessing a growth of scientific publications exploring NN architectures for DRP.

Dozens of DL-based models have been published, exploring diverse feature representations, NN architectures, learning schemes, and evaluation methods. However, only three out of the many existing reviews explicitly focus on modeling DRP with DL (10, 11, 18). These reviews present a categorized summary of methods, citing and discussing a limited number of examples from each category. As a result, only a selective overview of methods is provided and limited aspects of published DRP models are considered, exhibiting a primary limitation of these reviews. In addition, the scope of existing reviews does not require a comprehensive search for models which

TABLE 1 Categorization of topics covered in existing review papers on drug response prediction (DRP).

Category	Topic	References
Data resources[a]	Characterization of cancer tumors	(8–18)
	Characterization of drugs	(15, 17, 19–21)
	Monotherapy screenings	(9, 10, 12–16, 18, 19)
	Drug combination screenings	(6, 9–11, 19, 22)
	Complimentary resources	(6, 8, 10, 11, 19, 20, 22, 23)
Data representation methods[b]	Representation of cancer tumors	(8, 9, 12, 18, 24)
	Representation of drugs	(17, 21, 24)
	Representation of treatment response	(8, 9, 14, 18, 24)
	Feature selection and extraction	(6, 8, 13, 16, 25)
Prediction models[c]	DL models for monotherapy	(6, 10, 14, 16–19, 21, 23, 26)
	DL models for drug combination therapies	(6, 10, 11, 16, 19, 21, 22)
	ML models for monotherapy	(6, 8, 12–16, 18, 21, 24, 27)
	ML models for drug combination therapies	(6, 16, 21, 22, 28)
Assessment and evaluation[d]	Experimental validation of prediction models	(14, 24–28)
	Evaluation methods of model performance	(8, 10, 12, 14, 18, 22, 26)

[a]Characterization of cancer tumors: resources of cancer omics data; Characterization of drugs: tools for generating drug feature representations; Monotherapy screenings: resources of single-drug response data; Drug combination screenings: resources of drug-combination response data; Complimentary resources relevant to DRP: additional data resources and computational methods relevant to the DRP problem such as drug-target interactions, pathway information, etc.
[b]Representation of cancer tumors: gene expressions, copy-number variations, and other representations of tumors; Representation of drugs: drug molecular descriptors, fingerprints, molecular graph structures, etc.; Representation of treatment response: drug efficacy metrics such as IC50, AUC, sensitive/resistant, and others; Feature selection and extraction: methods for feature selection and dimensionality reduction.
[c]Models for monotherapy: discussion of models predicting single-drug response; Models for drug-combination therapies: discussion of models predicting drug-combination response.
[d]Experimental validation of prediction models: assessment studies that conduct experiments to assess performance of DRP models; Evaluation methods of model performance: discussion of methods for evaluating performance of DRP models.

could reveal predominant and emerging trends. Considering the current rate of publications, the diversity of approaches, and the limitations of existing reviews, a comprehensive review is necessary and timely.

This review is the first one to conduct a comprehensive search to include all relevant papers that utilize NNs for DRP. As of the time of completing this paper, a total of 61 peer-reviewed publications have been identified. Only models predicting response to single-drug treatments are included in the current review, excluding models that make predictions to combination therapies. We identified three major components involved in developing
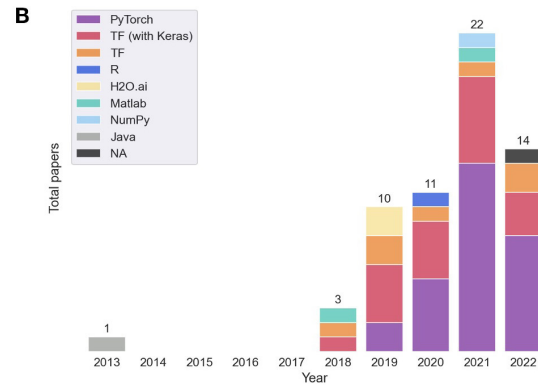
FIGURE 1
(A) A snapshot of Supplementary Table 1 that lists peer-reviewed papers proposing monotherapy drug response prediction (DRP) models (the full list can be found in the Supplementary material). The papers have been curated to identify various properties (shown in table columns), such as deep learning (DL) methods, feature types, evaluation methods, etc., as discussed in detail in this paper. Multiple plots in this paper have been generated using the data from Supplementary Table 1 [**(B)**, Figures 4–7C, D]. **(B)** Distribution of papers by year that use DL methods for DRP. The neural network models are designed using popular computational frameworks which include both proprietary and open-source software. The bar plots are color-coded by the different computational frameworks. TensorFlow/Keras and PyTorch are the most popular frameworks based on this plot. Data were collected until August 2022, considering only peer-reviewed publications.

DRP models, including data preparation, model development, and performance analysis. These components have been used to guide the curation of papers with special focus on representation methods of drugs, cancers, and measures of treatment response, DL related methods including NN modules and learning schemes, and methods for evaluating the prediction performance. This information is summarized in Supplementary Table 1. Summary plots have been generated, revealing the prevalence of methods used in these papers. Observing the prevalence of methods will assist in revealing approaches that have been investigated in multiple studies as well as emerging methods which are rather underexplored for DRP. We believe this review would serve as a valuable reference for new and experienced researchers in this field.

Section 2 formulates the DRP as a DL problem. Sections 3–5 provide a detailed review of the drug, cancer, and response representations used in papers listed in Supplementary Table 1. Existing design choices are summarized in Sections 6.1, 6.2 in terms of fundamental NN building blocks and learning schemes, respectively. Section 7 compiles existing approaches for evaluating model performance. In Section 8, we discuss the current state of DRP field, determine primary challenges, and propose further directions.

2. Deep learning-based drug response prediction workflow

A DRP model can be represented by $r = f(d, c)$, where f is the analytical model designed to predict the response r of cancer c to the treatment by drug d. The function f is implemented with a NN architecture in which the weights are learned through backpropagation. This formulation is for pan-cancer and multi-drug[1] prediction model where both cancer and drug representations are needed to predict response. A special case is drug-specific models designed to make predictions for a drug or drug family [e.g., drugs with the same mechanism of action (MoA)] (39). These models learn

from cancer features only and can be formulated as $r = f_D(c)$. Another type of models is multi-task learning models which take only cancer representations as inputs and generate multiple outputs where each output produces predictions for a specific drug. As compared to drug-specific models, the multi-task formulation enables learning from larger amounts of drug response data while exploiting common characteristics among drugs, thereby allowing to further improve generalization of the entire model (further discussed in Section 6.2.3).

The general workflow for developing DRP models is not much different from developing supervised models for other applications (Figure 2). The challenges come in the specific details typical to predicting treatment response. The process can be divided into three components: (1) data preparation, (2) model development, and (3) performance analysis (10, 12). The choice of methods associated with each one of these components can have a significant impact on the overall workflow complexity and the potential application of the final model.

2.1. Data preparation

Data preparation is generally the initial step in designing a prediction model, requiring expertise in bioinformatics and statistical methods. During this step, heterogeneous data types are aggregated from multiple data sources, preprocessed, split into training and test sets, and structured to conform to an API of a DL framework. The generated drug response dataset with N samples, denoted by $S = \{d, c, r\}_{i=1}^{N}$, includes representations for drug (d), cancer (c), and response (r). Prediction generalization is expected to improve with a larger number of training samples as demonstrated with multiple cell line datasets (40–42), normally creating preference for larger datasets when developing DL models. Recent research focusing on data-centric approaches suggests that efficient data representations and proper choice of a training set are at least as important as the dataset size for improving predictions,
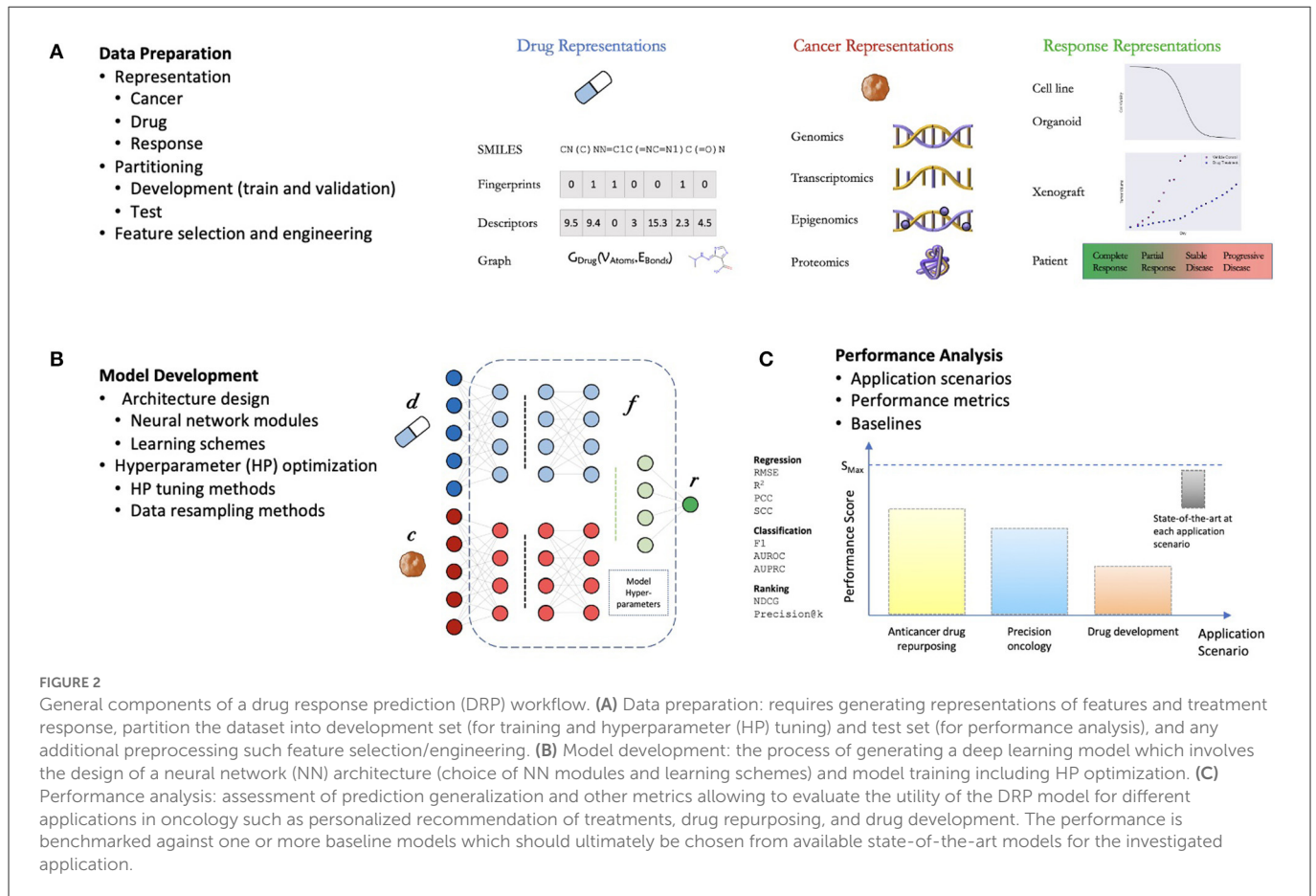
---

1 Sometimes referred as pan-drug.

FIGURE 2

General components of a drug response prediction (DRP) workflow. **(A)** Data preparation: requires generating representations of features and treatment response, partition the dataset into development set (for training and hyperparameter (HP) tuning) and test set (for performance analysis), and any additional preprocessing such feature selection/engineering. **(B)** Model development: the process of generating a deep learning model which involves the design of a neural network (NN) architecture (choice of NN modules and learning schemes) and model training including HP optimization. **(C)** Performance analysis: assessment of prediction generalization and other metrics allowing to evaluate the utility of the DRP model for different applications in oncology such as personalized recommendation of treatments, drug repurposing, and drug development. The performance is benchmarked against one or more baseline models which should ultimately be chosen from available state-of-the-art models for the investigated application.

and further emphasize the importance of the data preprocessing step (43).

2.2. Model development

Model development refers to NN architecture design and optimization of model hyperparameters (HPs). To design NNs, developers often resort to common heuristics which rely on intuition, experimentation, and adoption of architectures from related fields. This process involves choosing the basic NN modules, the architecture, and learning schemes. Diversity of data representations for cancers (9) and drugs (21) and potential utilization of DRP models in several pre-clinical and clinical settings have led researchers to explore a wide range of DL methods.

2.3. Performance analysis

A desirable outcome of a model development workflow is a robust model that produces accurate predictions across cancers and drugs as evaluated by appropriate performance metrics. DRP models can be used in various scenarios such as personalized recommendation of treatments, exploration of drug repurposing, and assisting in development of new drugs. Therefore, both performance metrics and appropriate evaluation schemes (e.g., design of training and test sets) are critical for proper evaluation of prediction performance.

The abundance of DRP papers in recent years (Figure 1) and lack of benchmark datasets (10), strongly suggest that a rigorous assessment of model performance is required where state-of-the-art baseline models serve as a point of reference.

2.4. Source and target domains

A vital characteristic that affects the entire workflow is the source domain data used to develop the DRP model, and the target domain data representing the biological domain on which the model is expected to operate. Data suitable for modeling DRP (Sections 3–5) come from multiple biological domains, such as cell lines, organoids, xenografts, and patients (9). Most models utilize data from a single domain (usually cell lines, Section 3.1, due to the abundance of response data). Certain models exploit data from a mix of domains with the goal to improve predictions in a target domain which suffers from insufficient data. Since a primary goal is to make DRP models useful in improving patient care, the biological domains can be further categorized into clinical (human patient data) and pre-clinical (cell lines, etc.). Considering this distinction, models can exploit data in different ways: train and test on preclinical (44–46), train and test on clinical, train on preclinical and test on clinical (47–49), and models that leverage both preclinical and clinical data during the training process and then test on clinical (50–54). Models that use data from mixed domains are generally more challenging to develop because they require extra steps in data preparation,

advanced modeling techniques, and robust performance evaluation analysis (Section 6.2.2).

3. Representations of treatment response

Drug screening platforms enable testing the sensitivity of cancer samples in controlled lab environments, ultimately producing data for predictive modeling (Figure 3). A major objective is a discovery of potential anticancer treatments through the screening of compound libraries against diverse cancer panels. Systematic drug screening platforms have been established for *in vitro* cancer models such as cell lines (55–57) and organoids (58), and *in vivo* models such as patient-derived xenografts (PDXs) (59). An overview of preclinical cancer models and the corresponding methods for sensitivity profiling is available in a recent review article (9). Drug sensitivity data can be transformed into continuous or categorical variables, representing the treatment response. Accordingly, prediction models have been designed to solve regression, classification, and ranking problems. This section reviews methods for representing treatment response and the corresponding prediction tasks.

3.1. Cell lines

Cell line studies constitute the most abundant resource of response data. High-throughput drug screenings with cell lines is performed with a compounds library, where each cell-drug combination is screened at multiple drug concentrations. The response of *in vitro* cells at each concentration is assessed *via* a cell viability assay which quantifies the surviving (viable) cells after treatment vs. the untreated control cells. Performing the experiments over a range of concentrations results in a vector of non-negative continuous dose-response values for each cell-drug pair. Those data points are summarized *via* a dose-response curve obtained by fitting a four parameters logistic Hill equation.

3.1.1. Continuous measures of response

The cell line dose-dependent responses lack a direct translation into the space of *in vivo* cancer models. A common approach is to extract from the dose-response curve a single-value summary statistic which represents the response for a cell-drug pair. Several methods exist allowing to calculate continuous response values which serve as the prediction target with supervised regression models. The two most common metrics are IC50 and AUC, with IC50 being substantially more prevalent (Figure 4).

Observing such predominance of IC50, an immediate question is whether IC50 exhibits substantial benefits over other measures. The half-maximal inhibitory concentration, i.e., IC50, is the concentration at which the drug reaches half of its maximal inhibition power on the fitted dose-response curve. Alternatively, measures such as AUC (area under the dose-response curve), AAC (area above the dose-response curve or activity area), and DSS (drug sensitivity score), are obtained by aggregating the cell viability values across a range of concentrations of the dose-response curve, providing what is considered a more global measure of response. Arguments in favor of these global measures are available, discussing the benefits of AUC

(41, 60) and DSS (61) as opposed to IC50. In addition, empirical analyzes suggest better prediction generalization in the case of using AAC values as opposed to the alternative of using IC50 (14).

3.1.2. Categorical measures of response

Despite the prevalence of IC50 and AUC, there are arguments suggesting that continuous measures of response (blue in Figure 4) lack a straightforward interpretation in the context of decision-making purposes. Alternatively, a categorical output representing a discrete level of response such as *sensitive* vs. *resistant* is more comprehensible for humans, and thus, better supports actionable outcomes. Two primary approaches were used to produce categorical responses with DRP models.

In the more common approach, continuous responses were first categorized usually using one of the available methods such as waterfall algorithm (55, 62), LOBICO (56, 63), or a histogram-based method (64, 65). DL classifiers were trained on the transformed values to predict class probability, which is subsequently translated into a discrete response label (orange in Figure 4). The predicted probability can also be utilized as a quantitative measure of prediction uncertainty, an essential aspect in decision-making. Most models were trained to predict a binary response, representing that cancer is either sensitive or resistant to treatment. A multi-class classifier with three classes corresponding to low, intermediate, and high responsiveness, was also explored (66). The second approach is to train regression models to predict one of the continuous responses and then categorize it into two or more classes (green in Figure 4). Only a few papers explored this approach, including binary (67–70) and multi-class labels (71). This approach naturally allows assessing model performance using both regression and classification metrics, possibly offering a more robust generalization analysis.

3.1.3. Ranking

In addition to producing continuous or categorical predictions, models can be trained to learn a ranking function with drug response data. In the context of clinical precision oncology, the clinician might be interested in obtaining a ranked list of the top-k drugs that are likely to be most beneficial for the patient. In our search, we obtained three ranking models, all of which targeting personalized treatment recommendation. By framing the problem as a ranking task, the authors proposed models that learn to produce a ranked list of drugs per cell-line that are most likely to inhibit cell viability. Prasse et al. (72) transformed IC50 into drug relevance scores and derived a differentiable optimization criteria to solve the ranking problem which can be combined with different NN architectures. By combining their ranking learning method with the PaccMann architecture (73) and a fully-connected NN (FC-NN), they significantly improved the ranking performance as compared to baseline models. SRDFM used a deep factorization machine (DeepFM) with pairwise ranking approach which generates relative rankings of drugs rather than exact relevance scores (74). PPORank generates drug rankings using deep reinforcement learning which enables to sequentially improve the model as more data becomes available. SRDFM and PPORank rankings outperform non-DL models.

FIGURE 3
Drug screening experiments are performed with various cancer models such as cell lines, organoids, and xenografts, where cancer samples are screened against a library of drug compounds. The screening data is transformed into a drug response dataset that can be used for developing drug response prediction models, including regression, classification, and ranking.

FIGURE 4
Cell-line drug response data is usually represented with continuous or categorical values. Drug response prediction (DRP) models use the different drug response representations to train regression, classification and ranking models. The histogram illustrates the prevalence of the difference representations and learning tasks. Certain papers exploit several representations of response and learning tasks, and therefore, these papers contribute more than one item to the histogram. The label *categorical* means that the continuous response was first categorized and then a DRP classifier was trained, while *continuous to categorical* means that a DRP regressor was trained and then the predicted response was categorized (in both cases, classification metrics were used for performance analysis).

3.2. Patient derived xenografts and patient tumors

Whereas, cell line data serve as the primary resource for training DL models, several papers proposed methods for predicting response in PDX and patient tumors. PDX is a contemporary cancer model that was developed to better emulate human cancer in medium-scale drug screenings, providing a controlled environment for studying the disease and systematically testing treatments in pre-clinical settings. While systematic drug screenings with patients is practically impossible (6), public data containing treatment response in individual patients are available.

Tumor response in humans is obtained *via* a standardized evaluation framework called Response Evaluation Criteria in Solid Tumors (RECIST) (75). RECIST involves non-invasive imaging followed by evaluation of change in tumor size. Four categories used for grading tumor change include Complete Response (CR), Partial Response (PR), Progressive Disease (PD) and Stable Disease (SD). Tumor change in PDXs is evaluated by monitoring tumor volume over time using more traditional methods (e.g., calipers) as opposed to using RECIST, primarily due to cost. The main data resource for modeling drug response with PDXs, provides continuous response values calculated based on tumor volume (59). To create a counterpart to RECIST, they also set cutoff criteria allowing to transform the continuous response values into the four categories.

Because discrete response labels are available in patient and PDX datasets, DRP models are primarily designed as classifiers, where the four-label RECIST categories are transformed into sensitive (CR and PR) and resistant (PD and SD) labels. Small sample size, however, is a major challenge in developing DL models with these datasets. This issue is generally addressed by incorporating abundant cell line data and transfer learning schemes to improve predictions in PDX and patients (discussed in Section 6.2.2).

4. Cancer representations

As personalized oncology treatments rely on omics biomarkers, progress in high-throughput tissue profiling plays an important role in existing and future cancer therapies. Pharmaco-omic studies conduct multiomic profiling of cancer models and drug screening experiments (9). Advances in sequencing technologies allowed to substantially scale the experimental studies by increasing the throughput rate and reducing the cost of profiling. The algorithms for processing the raw data evolve as well, thereby providing more reliable representations of the underlying cancer biology and allowing projects to update their repositories with refined versions of existing data while utilizing improved processing techniques (e.g., DepMap portal[2]).

Dedicated bioinformatics pipelines for transforming raw data depend on the omic type and the profiling technology, where common steps include alignment, quantification, normalization, and quality control. Profiling at each omic level produces high-dimensional representation of cancer that can be used as features in the downstream modeling of drug response. Existing prediction models utilize features that were obtained primarily at four omic levels including genomic [mutation, copy-number variation (CNV)], transcriptomic (gene expression microarrays, RNA-Seq), epigenomic (methylation), and proteomic [Reverse Phase Protein Arrays (RPPA)] (9). Figure 5A shows the prevalence of these representations in DRP models. Many models use these omic features directly as inputs to NNs while others optimize representations *via* additional preprocessing with the goal of increasing predictive power (76, 77).

In 2014, the NCI-DREAM challenge was a community effort that assessed performance of various classical ML models in predicting drug response in breast cancer cell-lines (27). It was reported that for most models participated in the challenge, gene expression microarrays provided greater predictive power than other data types.

Results of this challenge seem to set the tone for future research in this field as gene expression was used in approximately 90% of the models either alone or in combination with other feature types, including mutation, CNV, methylation, and RPPA. Mutation and CNV are also common but rarely used without gene expression (with only 12% of such models).

Learning from multiple omic types have shown to improve generalization, and as consequence, integration of multiomics have been a recent trend (Figure 5C). A straightforward approach is to concatenate the multiomic profiles to form a feature vector and pass it as an input to a NN model, a method commonly referred to as *early integration*. It has been shown, however, that *late integration* significantly improves predictions where the different omic profiles are passed through separate subnetworks before the integration (39, 40, 78). Yet, caveats related to data availability hinder the pertinence of multiomics as opposed to using single-omics. Not having all omics available for all samples is a common occurrence in pharmaco-omic projects. A common approach to address this issue is filtering the dataset to retain a subset of cancer samples that contain all the required omic types as well as drug response data. The multiomics data is closely related to multimodal learning which refers to prediction models that learn jointly from multiple data modalities (i.e., representations). It was demonstrated that leveraging multimodal data generally improves predictions.

5. Representations of drug compounds

Contemporary cancer treatment often involves administering therapeutic drugs to patients to inhibit or stop growth of cancer cells, destroy tumors, or boost cancer-related immune system. Drug molecules are 3-D chemical structures consisting of atoms and bonds with complex atomic interactions. Developing numerical representation of molecules is an active research area in several related disciplines, including *in silico* design and discovery of anticancer drugs.

Ultimately, drug representations should be able to encode essential physical and chemical properties of molecules in compact formats. Alternatively, when using these representations as drug features for DRP, the model should be able to properly ingest these features and extract information predictive of treatment response. Therefore, the inherent information encoded with numerical drug representations and the model capability to learn from that specific representation are closely related and mutually important for producing efficient DRP models.

Neural networks provide substantially more flexibility than classical ML in learning from unstructured data such as strings, images, and graphs. This flexibility facilitates a noticeable trend of exploring various drug representations as features in DL-based DRP models (21). Primary types of drug representations include SMILES, fingerprints, descriptors, and graph-based structures (Figure 5B).

5.1. SMILES

Perhaps the most common format for querying, handling and storing molecules when working with modern chemoinformatics software tools and databases is a linear notation format called SMILES (simplified molecular-input line-entry system). With this

---

FIGURE 5
Feature representations in drug response prediction (DRP) models. Various representations can be used to represent cancers and drugs in DRP models, as described in Sections 4, 5, respectively. Each DRP model usually exploits one or multiple feature types (Supplementary Table 1 lists the feature types that each model have used). **(A)** The prevalence of omics (cancer) feature representations in DRP models. **(B)** The prevalence of drug feature representations in DRP models. **(C)** The distribution of papers that utilized single-omics and multiomics features in DRP models. We used Supplementary Table 1 to generate these figures, where we considered only peer-reviewed publications, collected until August 2022. CNV, copy number variations; RPPA, reverse phase protein arrays.

data structure, each molecule is represented with a string of symbolic characters generated by a graph traversal algorithm (21, 79). Although the use of SMILES in DRP is quite limited as compared to other applications (e.g., molecule property prediction, QSAR), this format provides several benefits.

Following common text preprocessing steps (e.g., tokenization, one-hot encoding), SMILES can be naturally used with common sequence-aware modules such as RNN (72, 73, 80) and 1-D CNN (81). Moreover, each molecule can be represented with different strings depending on the initial conditions when applying the SMILES generating algorithm (i.e., starting node of graph traversal). The resulting strings are referred to as randomized or enumerated SMILES and have been reported to improve generalization when utilized for drug augmentation (82, 83). SMILES are less prevalent in DRP models but remain a popular notation for describing molecules because it often serves as an intermediate step for generating other representations such as fingerprints, descriptors, and graph structures.

5.2. Descriptors and fingerprints

Fingerprints (FPs) and descriptors are the two most common feature types for representing drugs in DRP papers (Figure 5B). Unlike with SMILES, where the string length varies for different molecules, we can specify the same number of features for all

drugs in a dataset with either FPs or descriptors. The consistent feature dimensionality across drugs makes descriptors and FPs easy to use with NNs and classical ML. With FPs, a drug is a binary vector where each value encodes presence or absence of a molecular substructure with a common vector size of 512, 1,024, or 2,048. Multiple algorithms for generating FPs are available, implemented by several chemoinformatics packages. For example, Morgan FPs refers to Extended Connectivity Fingerprints (ECFPs) generated *via* the Morgan algorithm. ECFP is a class of circular FPs where atom neighborhoods are numerically encoded with binary values. The open-source package RDKit provides an API for generating these FPs which are often used for DRP.[3] Descriptors are a vector of continuous and discrete values representing various physical and chemical properties, usually containing hundreds or a few thousands of variables. Both open-source and proprietary software tools are available for generating descriptors, where PaDEL (84), Mordred (85), and Dragon (86) being particularly used for DRP. A systematic assessment of various features and HPs suggests that DRP with DL exhibits no significant difference when utilizing ECFP, Mordred, or Dragon representations as drug features (87).

---

3 RDKit: Open-source cheminformatics. Last accessed: January (2023). https://www.rdkit.org.

5.3. Graph structures

Graphs are powerful representations where complex systems can be represented using nodes and edges. Recent advancements in GNNs have opened promising research directions allowing efficient learning of predictive representations from graph data (88–90), and contributing to the widespread interest in GNN among AI researchers and application domain experts (91, 92). Graph-based representations have emerged as a new trend in computational drug development and discovery (93, 94). In graph notation, each molecule is described with a unique graph of nodes and edges, where each atom (i.e., node) and each bond (i.e., edge) can be represented with multiple features characterizing physical and chemical properties. Resources and examples for constructing graphs are well documented in popular chemoinformatics toolkits, as well as software packages dedicated for GNNs (95–97).

By borrowing methodologies from related fields such as drug property prediction and drug design, developers of DRP models exploit graph molecular structures for representing drugs combined with GNN-based architectures. We observe concordant results among those papers investigating molecular graphs with GNNs, reporting superior performance of their models compared to baselines that use non-graph representations. Based on these recent results and the assumption that graphical description is arguably a more natural way to represent drugs as compared to the aforementioned alternatives, molecular graphs combined with GNNs should be a default modeling choice. However, only few papers actually report results with rigorous ablation analysis with extensive HP tuning comparing molecular graphs with other representations such as SMILES, FPs, and descriptors for the application of DRP.

6. Deep learning methods for drug response prediction

We review existing approaches for modeling DRP with DL in terms of two perspectives: (1) NN modules that are used as building blocks for constructing DRP architectures, (2) learning schemes that are used to train models with the goal to improve prediction generalization. The diversity of methods is highlighted in Figure 6.

6.1. Neural network modules

Constructing a NN architecture requires choosing appropriate components, their properties, and the way these components are organized and connected (i.e., network topology). These components are often available as modules in DL frameworks and range from simple dense layers to more advanced structures such as attention. Certain structures, however, are not generally available as is and should be constructed manually using the available modules (e.g., residual connections). While there is an absence of rigorous methodologies for designing network topology, the choice of certain modules can be driven by the characteristics of input data.

6.1.1. Dense layers

Menden et al. (34) is a pioneering work where cell line and drug features were used to train an FC-NN with a single hidden layer to predict IC50. Recently, FC-NN models with multiple dense hidden layers have been proposed to predict response in cell lines (71) and humans (67). Several other models with only dense layers were explored, combining advanced NN attributes. MOLI utilizes triplet loss function and late integration of multiomic inputs to generalize across cancer models by training with cell lines and predicting in PDXs and patients (39). DrugOrchestra is multi-task model that jointly learns to predict drug response, drug target, and side effects (98). RefDNN explores data preprocessing techniques to generate more predictive cell line and drug representations (99). For cell line representation, multiple ElasticNets produce a vector representing drug resistance of a cell line to a set of reference drugs, while for drug representation, a structure similarity profile is computed for each input drug with respect to the set of reference drugs. Another model with late integration of multiomic data is proposed for predicting drug response and survival outcomes which also uses neighborhood component analysis for feature selection of multiomic data (69). PathDSP utilizes multi-modal data preprocessed *via* pathway enrichment analysis and integrated into an FC-NN for DRP (45). All the aforementioned models demonstrate promising prediction performance while utilizing dense layers only in their architectures.
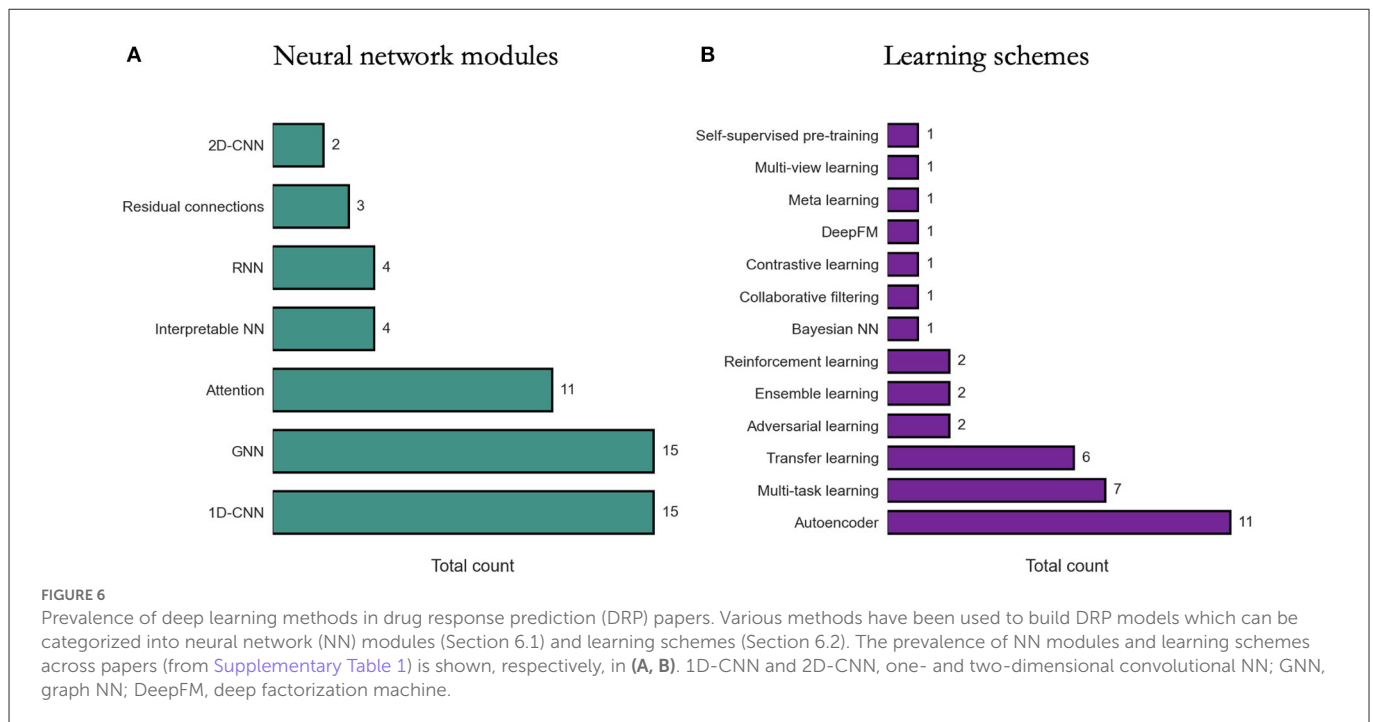
6.1.2. Convolutional layers

Due to state-of-the-art performance in various applications (100) and fewer learning parameters required as compared to dense layers, 1-D CNNs are a popular choice in DRP models, particularly for processing omics data. DRP models utilizing CNNs have been published every year since 2018.

DeepIC50 is a multi-class classifier exploiting an early integration of cell line mutations and drug features (descriptors and FPs) passed to a three-layer CNN and followed by an FC-NN prediction module (66). tCNNs exploits late integration of two parallel CNN subnetworks with three layers each (101). Binary features of cell lines (mutations and CNV) and drugs (one-hot encoded SMILES) are propagated through the respective subnetworks, concatenated, and passed through an dense layer for the prediction of IC50. GraphDRP is very similar to tCNNs in terms of CNN utilization, with the primary difference being the utility of molecular graphs and GNN layers for learning drug representations. CDRScan is an ensemble of five different architectures with varying design choices (late and early integration, shallow and deep NNs, with and without dense layers) which all contain CNNs and learn to predict IC50 from mutation and FP data (102).

DeepCDR takes the design further by exploiting late integration of drug and multiomic features, including genomic mutation, gene expression, and DNA methylation, where only the mutations are passed through a CNN layers (103). Following DeepCDR, two similar models were published, GraTransDRP (104) and GraOmicDRP (105) which both use late integration of multiomics and molecular graphs for drug representation. While DeepCDR uses CNNs only for genomic features, all three subnetworks for multiomic inputs consist of CNNs in GraTransDRP and GraOmicDRP. SWNet is another model utilizing late integration of cell lines and drugs with three CNN layers in the cell-line subnetwork and one such layer in the prediction subnetwork (106).

We may gather that 1-D CNNs carry substantial benefits in the context of DRP, considering their prevalence (Figure 6). Based

FIGURE 6
Prevalence of deep learning methods in drug response prediction (DRP) papers. Various methods have been used to build DRP models which can be categorized into neural network (NN) modules (Section 6.1) and learning schemes (Section 6.2). The prevalence of NN modules and learning schemes across papers (from Supplementary Table 1) is shown, respectively, in **(A, B)**. 1D-CNN and 2D-CNN, one- and two-dimensional convolutional NN; GNN, graph NN; DeepFM, deep factorization machine.

on our search, however, only two papers explicitly report their empirical findings assessing CNNs vs. alternative learning modules. Interestingly, both papers suggest that CNNs underperform dense layers in their respective architectures. Zhao et al. (71) reports that in a twelve-layer model predicting response from gene expression, dense layers outperform CNNs or RNNs. Manica et al. (73) present extensive analyzes exploring various architectures for encoding drug representations directly from SMILES, including CNNs, bidirectional RNNs (bRNNs), and various attention modules. Their results suggest that a combination of convolutional and attentions modules produces the most predictive model while a CNN-only encoding subnetwork performs the worst as compared to the explored variants.

With the objective to leverage 2-D CNNs, two algorithms were recently published for converting tabular data into images and utilizing CNNs for predicting drug response. Both algorithms, REFINED (76) and IGTD (77), convert gene expressions and drug descriptors into images as a preprocessing step, and then learn to predict drug response with late integration of 2-D CNN subnetworks. The papers report superior performance of proposed algorithms as compared to various baseline models.

6.1.3. Attention mechanism

Presenting PaccMann (80), Oskooei et al. were the first to report the use of attention-based NN for DRP, exploiting late feature integration with attention mechanisms incorporated both in the cell line and drug subnetworks. On the cell line path, gene expressions are encoded with self-attention producing a gene attention (GA) vector. On drug path, SMILES embeddings are combined with GA *via* contextual-attention, where the GA vector serves as the context. This design which was further described in Manica et al. (73) improves generalization as compared to baselines as well as facilitates interpretability *via* attention weights.

Attention-based designs have been further explored in recent years. CADRE is a collaborative filtering model with contextual-attention and pre-trained gene embeddings (gene2vec), designed to recommend treatments based on cell line gene expressions (107). Ablation analyzes against simpler models suggest that CADRE's attributes yield improved generalization. Moreover, the authors show how attention weights could be used to identify biomarker genes, partially addressing the notorious black-box property associated with many DL models. In AGMI, attentions are used for aggregating heterogeneous feature types, including raw multiomics as well as engineered features *via* protein-protein interaction (PPI), gene pathways, and gene correlations with PCC (Pearson correlation coefficient), contributing to improved generalization. Overall, papers report the predictive benefits of attentions (106–111), while a few papers also explore attention weights for interpretability (107, 110).

Transformer is an attention-based architecture that have recently been explored in DRP models for encoding drug representations (70, 104). In GraTransDRP (104), the authors propose to extend an existing model, GraOmicDRP (105), by modifying the drug subnetwork to include graph attention network (GAT), graph isomorphism network (GIN), and a graph transformer which learns from graph data. In DeepTTA (70), a transformer module encodes drug information represented as text data [ESPF substructures (112)]. Both models report significant improvement in generalization thanks to transformer modules. Transformers have shown immense success first in language and later vision applications, and expected to gain further attention from the DRP community.

6.1.4. Graph neural network layers

Designed to efficiently learn from graph data, GNNs have been extremely popular in applications where information can be represented as graphs. With at least 15 GNN-based DRP models published since 2020 (Supplementary Table 1), it is apparent that

the use of GNNs with graph data becomes an emerging alternative to some of the more traditional approaches discussed above. In DRP papers, a graph, $G = (V, E)$, is often characterized by a set of nodes V (a.k.a. vertices), a set of edges E (a.k.a. links), an adjacency matrix W which contains values representing associations between the nodes, and an attribute matrix A containing feature vectors representing node attributes (113). A primary challenge that researchers are facing when designing GNN-based models lies in how to transform components of the DRP problem into graphs. Luckily, accumulated knowledge and methodologies allow viewing biological and chemical systems as networks.

Due to substantial progress in applying GNNs to drug discovery and development (94, 95), it was relatively straightforward to adopt similar techniques to DRP. Papers propose models that convert drug SMILES into graphs and utilize modern GNN layers to encode latent drug representations (49, 103–106, 114). The common approach is constructing a unique graph per drug with nodes and edges representing atoms and bonds, respectively, where node features are the properties of each atom. The subnetworks consist of several (usually 3–5) GNN layers such as GCN and GIN.

Another approach is constructing graphs using biological information of cancer samples, where genes are the graph nodes and gene relationships are the edges. There is slightly more diversity of approaches in this space as compared to drugs partially due to the use multiomic data which allow constructing graphs with heterogeneous node attributes. The multiomics can be utilized to encode gene relationships and attributes using one or more data modalities, including correlations between genes (109, 111, 115, 116), known protein interactions (i.e., PPI) (109, 111, 117, 118) using STRING database (119), and relationships based on known gene pathways (109) using GSEA dataset (120). Recently, novel approaches have been explored such as heterogeneous graphs where both cell lines and drugs are encoded as graph nodes (108, 116, 118, 121), and a model that utilizes diverse data types for building graphs, including differential gene expressions, disease-gene association scores and kinase inhibitor profiling (111). Despite the diversity of the different approaches, most papers report superior performance as compared baseline models.

6.2. Learning schemes

Various learning schemes have been proposed to improve drug response prediction in cancer models. These techniques can be used in different configurations, usually regardless of the fundamental NN building blocks that are used to construct the architecture.

6.2.1. Autoencoders

When considering the large feature space of cancer and drug representations (Sections 4, 5) and the size of drug response datasets, the number of input features outnumbers the number of response samples which potentially leads to model overfitting (6). Learning predictive representations with high-dimensional data manifests a primary strength of multi-layer NNs. A prominent example are autoencoders (AEs) which are NNs that are trained to compress and then reconstruct data in an end-to-end unsupervised learning fashion. Given that an AE can reliably reconstruct the input data, the compressed latent representation is characterized by reduced

data redundancy which improves the feature to sample ratio and can be subsequently used as inputs in downstream drug response prediction task.

The primary application of AEs in DRP models is dimensionality reduction. The compressed representation is used as input to a DRP model which is usually a NN or in some cases a classical ML model. DeepDSC reduces gene expressions from about 20,000 genes down to 500 and concatenates the downsampled data with drug FPs as inputs to a FC-NN (122). VAEN exploits variational AEs (VAEs) to learn a low-dimensional representation of gene expressions which are fed into ElasticNet (48). DEERS compresses cancer features (gene expression, mutation, and tissue type) and drugs features (kinase inhibition profiles) into 10 dimensions which are concatenated for a FC-NN (44). Dr.VAE learns pre- and post-treatment embeddings of gene expressions and leverages those embeddings for DRP (123). AutoBorutaRF combines AEs with a feature selection method followed by Boruta algorithm and RF (124). Ding et al. (125) utilized learned representations of multiple hidden layers of the encoder as input features, rather than using just the latent representation.

AEs have also been utilized with external datasets as part of model pretraining. DeepDR uses gene expressions and mutations from the TCGA database to train separate AEs for each omic type (126). The pre-trained encoders are extracted from the individual AEs, concatenated, and passed as inputs to a FC-NN. The combined model was trained to predict drug response in cell lines. Another model uses separate AEs to encode cancer and drug features (127). A GeneVAE encodes gene expressions from CCLE dataset, and a Junction-Tree VAE (JT-VAE) (128) encodes drug molecular graphs from the ZINC database. Similar to DeepDR, the pre-trained encoders are concatenated and combined for DRP.

Dimensionality reduction and model pre-training are common applications of AEs. Yet, AEs have also been utilized in less conventional ways. Xia et al. (41) proposed UnoMT, a multi-task learning (MTL) model that predicts drug response in cell lines *via* late-integration of multiomic features and drug representations. In addition to the primary DRP task, the model predicts several auxiliary tasks, where one of them is a decoder that reconstructs gene expressions. TUGDA exploits advanced learning schemes such as domain adaption and MTL to improve generalization in datasets with limited sample size (51). Each prediction task in the MTL corresponds to a different drug. With the goal to mitigate the influence of unreliable tasks and reduce the risk of negative transfer, a regularization AE takes the model output and reconstructs an intermediate hidden layer.

6.2.2. Transfer learning

Transfer learning refers to learning schemes designed to improve generalization performance in a target domain T by transferring knowledge acquired in a source domain S (129). The most common scenario is to transfer learned representations from a source domain with large amounts of data into a target domain which suffers from insufficient data. Transfer learning with DL have shown remarkable success in various applications (130) and have recently been applied to DRP (14, 42, 50, 51, 78, 131).

Zhu et al. proposed an ensemble transfer learning (ETL) that extends the classical transfer learning by combining predictions from multiple models, each trained on a cell-line dataset and fine-tuned on a different cell-line dataset. Considering three application scenarios,

a NN with late-integration of gene expression and drug descriptors outperformed baselines in most cases, while the LightGBM showed superior performance in certain experimental settings. AITL was proposed to improve generalization across biological domains (50). The model was trained using gene expressions and drug responses from source and target domain data, where abundant cell line data serve as source and the relatively scarce data from either PDXs or patients serves as target. A primary component for accomplishing knowledge transfer is a feature extractor subnetwork that learns shared representations using source and target features, which are passed to multi-task subnetwork for response prediction of source and target samples. Ma et al. (131) proposed TCRP, a two-phase learning framework with meta-learning as a pre-training step followed by a few-shot learning for context transfer. With the goal to obtain transferable knowledge, a meta-learning is applied iteratively by training a simple NN with different subsets of cell lines. The same NN was trained to predict drug response in both phases but with different data. With a few samples from the target context, the pre-trained NN was further trained for a single iteration, demonstrating performance with patient-derived cell line (PDTC) and PDXs.

ETL and AITL can be categorized as inductive transfer learning, where labels from both the source and target data are used to improve generalization on target data (130). However, abundance of per-clinical and clinical data are not labeled, and therefore, remain unused with classical transfer learning. Velodrome is a semi-supervised model that exploits labeled cell lines and unlabeled patient data to improve out-of-distribution (OOD) generalization on datasets that remained unused during training. Generalization has been demonstrated with PDX, patient, and cell lines from non-solid tissue types. TUGDA is an MTL and domain adaptation model which relies on cell line data and advanced learning schemes to improve predictions in data-limited cancer models (51). The model takes gene expressions and learns a latent shared representation that is propagated to predict drug response in a multi-task fashion with each task corresponding to a different drug. To enable domain adaptation from cell lines to other cancer models such as PDXs and patients, adversarial learning is used where a discriminator is branched from the shared layer to classify the type of cancer model. The discriminator is employed in supervised and unsupervised steps, with source- and target domain samples, respectively.

6.2.3. Multi-task learning

As transfer learning and multi-task learning (MTL) are related approaches (129, 130), several models have integrated both methods in their DL-based DRP models (41, 50, 51). However, utilizing MTL without transfer learning is also common. DrugOrchestra is an original work where multiple heterogeneous data sources were curated to develop a multi-task model for simultaneously learning to predict drug response, drug target, and side effects while utilizing advanced weigh sharing approaches (98). SWnet proposed a gene weight layer which scales the contributions of mutations when combined with gene expressions as a multiomic input (106). In its simplest form, the weight layer is trained for the entire dataset. The authors also explored a weight matrix where each vector is determined for a single drug. Considering N drugs in the dataset, the model was trained in an MTL fashion with N prediction tasks, one task for each drug, outperforming the single weight layer configuration. Zhu et al. (118) used MTL in a pre-training step

to accumulate chemical knowledge by training a GNN to predict more than a thousand biochemical properties of molecules, a strategy adopted from another work (132). The pre-trained model was integrated into the complete DRP model called TGSA.
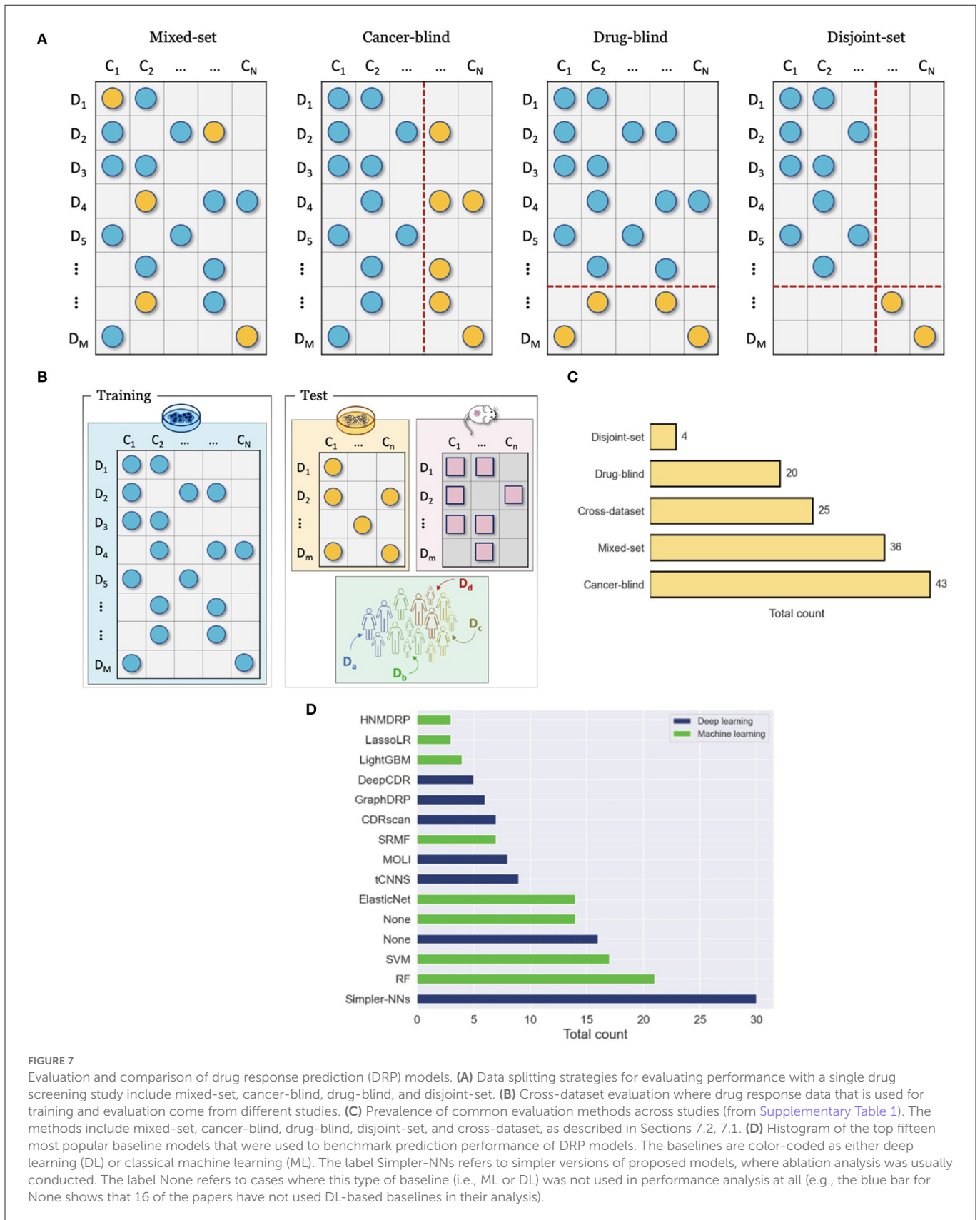
7. Model evaluation and comparison

Papers proposing DRP models explore advanced approaches for data representation, NN architecture design, and learning schemes. To motivate the increasing complexity of methods, it is essential to demonstrate their utility for cancer medicine and advantages against existing approaches (i.e., baseline models). Model performance in these papers is primarily assessed *via* prediction generalization, i.e., prediction accuracy on a test set of unseen data samples, with several common metrics used for regression, classification, and ranking. The choice of the test set is critical as it may illustrate the potential utility of the model for a given application in cancer medicine and significantly affect the observed performance. To make a comparison between proposed and existing models compelling, consistent evaluations should be utilized in terms of drug response datasets, training and test sets, evaluation metrics, and optimization efforts (e.g., HP tuning). This section presents common evaluation approaches used for prediction generalization in terms of drug screening datasets (single or multiple datasets), data splitting strategies (train and test), and baselines models (Figure 7).

7.1. Single-dataset evaluation

A single drug screening study is most commonly used to develop and assess performance of DRP models. Regardless of specific cancer models, the screening data space can be characterized by the involved drugs and cancers. The 2-D matrix in Figure 3 illustrates such a space composed of a finite number of known cancer cases and drugs, where the marked coordinates symbolize that treatment responses are available for these combinations. Considering this 2-D space, four data splitting strategies are commonly used, where each can imitate a different application scenario of drug response prediction (Figure 7A).

7.1.1. Known cancer and drugs (mixed-set)

To ensure a valid assessment of prediction generalization, training and test sets should contain unique cancer-drug pairs without overlap. The overlap may occur, however, on cancer or drug (but not both) in which case a model encounters drug or cancer features during testing that also showed up during the training phase. For example, consider two drug response samples $r_{11} = \{d_1, c_1\}$ included in the training set and $r_{12} = \{d_1, c_2\}$ in the test set. While the pairs are unique, the drug features of d_1 show up during both training and testing. The cancers or drugs that show up in test and training set are deemed as *known* or *seen*. This splitting strategy is also known as mixed-set, where drugs and cancers are mixed in training and test sets. Randomly splitting cancer-drug pairs naturally results in a mixed-set analysis, where drug and cancer features can show up in both training and test sets. Due to the overlap on both feature dimensions, this analysis naturally leads to the highest performance, and it is also the most common and straightforward to implement.

FIGURE 7
Evaluation and comparison of drug response prediction (DRP) models. **(A)** Data splitting strategies for evaluating performance with a single drug screening study include mixed-set, cancer-blind, drug-blind, and disjoint-set. **(B)** Cross-dataset evaluation where drug response data that is used for training and evaluation come from different studies. **(C)** Prevalence of common evaluation methods across studies (from Supplementary Table 1). The methods include mixed-set, cancer-blind, drug-blind, disjoint-set, and cross-dataset, as described in Sections 7.2, 7.1. **(D)** Histogram of the top fifteen most popular baseline models that were used to benchmark prediction performance of DRP models. The baselines are color-coded as either deep learning (DL) or classical machine learning (ML). The label Simpler-NNs refers to simpler versions of proposed models, where ablation analysis was usually conducted. The label None refers to cases where this type of baseline (i.e., ML or DL) was not used in performance analysis at all (e.g., the blue bar for None shows that 16 of the papers have not used DL-based baselines in their analysis).

Because of the simplicity of this strategy, almost all papers include it in their analysis. An application of prediction models developed with the mixed-set validation is repurposing of known anticancer

drugs for unexplored cancer conditions that have not been treated with these drugs. When considering the 2-D space of known cancer cases and drugs in a given dataset, not all cancer cases and drugs

have been screened against each other. If a DRP model exhibits high generalization, it can be deployed as a virtual screening tool which helps in identifying highly sensitive cancer-drug pairs and thereby assisting in the design of experiments or treatments.

7.1.2. Unknown cancers and known drugs (cancer-blind)

When a dataset is split such that an overlap occurs only on drugs but not on cancer cases, it is said that drugs are known but the cancer cases are unknown with respect to the test set. This splitting strategy is also sometimes referred to as cancer-disjoint, cancer-blind, unseen-cancer, and a special case called leave-n-cancers-out where a certain number of cancers is excluded from the training set. Splitting a drug response dataset this way requires extra work which is, therefore, less common than random splitting but still prevalent, especially in recent years. The performance is usually worse than in the mixed-set analysis due to increased complexity in generalizing across the omics feature space of cancers. Alternatively, this type of analysis is perhaps the most adequate in simulating the utility of DRP models for the design of personalized cancer treatment. When considering a potential workflow of utilizing a DRP model for personalized cancer treatment, the patient tumor data is not expected to be included in the training set that was used for model development. However, the drug treatments would be chosen from a collection of known compounds (approved or investigational) that may have been screened against other cancer cases, the drug features of which were used to train the prediction model during model development. Models that exhibit high generalization in this scenario are considered more suitable for designing personalized cancer treatment.

7.1.3. Known cancers and unknown drugs (drug-blind)

Analogously to cancer-blind analysis, the drug-blind refers to data partitioning where an overlap between the training and test sets may occur on cancers but not on drugs. This data splitting strategy is also sometimes referred to as drug-disjoint, unseen-drugs, and a special case called leave-n-drugs-out where a certain number of drugs is excluded from the training set. Interestingly, the generalization in drug-blind analysis as evaluated by common performance metrics is significantly worse than in mixed-set and cancer-blind analysis and, in some cases, models completely fail to make effective predictions. The worse performance can be attributed to the immense chemical space of drug compounds which imposes a challenge on predictive models of learning generalizable feature embeddings with just a few hundreds of drugs that were screened in typical drug screening studies. In addition, it has been shown empirically that drug diversity contributes to a majority of response variation (41, 133), which can explain the performance drop in drug-blind analysis. Models that excel in generalizing for unknown drugs can be useful for repurposing of non-cancer therapies to cancer indications and development of novel drugs for cancer treatment. Existing drugs approved for non-cancer diseases can be repurposed for cancer treatment, thereby decreasing expenses and speeding up the time to market. A DRP model, exhibiting high performance in drug-blind scenario, can be utilized for *in silico* drug screening across cancer types and libraries of approved compounds.

7.1.4. Unknown cancers and drugs (disjoint-set)

By extending the drug-blind and cancer-blind analysis to both dimensions, generalization analysis can be performed where both drugs and cancers remain disjoint between training and test sets. This analysis exhibits the worst generalization performance. It is primarily used to assess the capacity of models to generalize in this challenging scenario where the application of such model in clinical or preclinical setting is not essentially obvious. Models exhibiting significantly better performance as compared to baselines could be interesting cases for further exploration of the model architecture to scenarios more relevant in pre-clinical and clinical settings.

7.2. Cross-dataset evaluation

Another validation scheme is to assess generalization across datasets, where response data for model development and model testing are derived from different drug screening studies (Figure 7B). The source dataset (D_S), used for model development, and the target dataset (D_T), used for evaluation, can be of the same or different cancer models. Several selected works have been summarized in Sharifi-Noghabi et al. (14), showing that cell lines constitute the primary cancer model for D_S while cell lines, PDX, and patient response data are all common options for D_T. Cross-dataset analysis is less prevalent as compared to single-dataset analysis. Primary challenges are related to data preparation and arise when standardizing metadata across datasets, including annotations of drugs, cancer samples, and omics features, as well as utilizing consistent data preprocessing steps for feature normalization (e.g., addressing batch effect) and computation of response metrics.

Inconsistencies in drug screening data across cell line studies, partially due to different experimental setups, are an acknowledged reality which impedes naive integration of data from multiple sources into a single dataset (62, 134). A large-scale empirical study focused on assessing generalization of DL and ML models across five cell line datasets suggests that generalization may depend on multiple attributes of D_S and D_T datasets, including the number of unique drugs and cell lines in D_S and D_T, the drug and cell line overlap between D_S and D_T, and the sensitivity assays used for measuring drug response. Although this type of analysis requires adequate knowledge and substantial effort in data preparation and model development, demonstrating the ability to make accurate predictions across datasets or cancer models can be essential for a DRP model to become a favorable candidate for preclinical and clinical studies.

7.3. Baselines

Proposing novel DRP models could be seen unnecessary unless it leads to better performance as compared to existing models. In predictive modeling, baseline models usually refer to models that serve as a point of reference. Figure 7D shows the number of times each model was used as a baseline across papers, where ML-based and DL-based models are color-coded to distinguish between the two types. The label *None* refers to cases where this type of baseline (i.e., ML or DL) was not used in performance analysis at all (e.g., the blue bar for *None* shows that 16 of the papers have not used DL-based baselines in their analysis).

The most popular choice for a baseline is to use a *simpler* version of a proposed model, where simplicity is defined in the context of a given model. For example, in the case of MOLI which advocates for multi-omics late integration with triplet loss function, simpler versions include combinations of single- instead multi-omics, early instead of late integration, and binary cross-entropy loss function instead of triplet loss function (39). Using simpler versions of a model as baselines can be both straightforward and insightful. In many cases, modifying a model to construct a baseline involves minor changes of little complexity where model attributes are replaced with simpler alternatives followed by ablation analysis demonstrating the necessity of the proposed attributes. However, the actual baselines and ablation experiments that are chosen to benchmark models are a critical aspect that affects credibility. Some DL models explore alternative data representations, such as images (76, 77) or graphs (105, 118). In these cases, ablation studies with common vector-based representations (e.g., FPS, descriptors) should be considered. Certain models incorporate structural biological information such as pathways into NN architectures with the goal to produce interpretable DRP models. With these models, a baseline may need to incorporate random pathway information to serve as a reference point to any claimed pathway interpretability, as discussed in Li et al. (135).

The second most popular baseline model is random forest (RF) presumably due to its availability, simplicity, and predictive power. Boosting algorithms such XGBoost and LightGBM are less prevalent although poses similar characteristics as RF yet faster and often more predictive. Tree-based models and boosting algorithms have shown tremendous success in various prediction tasks and data science competitions. HP tuning is often an essential step to squeeze the maximum performance from these algorithms. When these models are used as baselines in DRP papers, very little (if any) HP tuning is performed. When the focus of proposed models is on prediction performance, extensive HP tuning should be performed with the baselines the reported.

Another common approach to demonstrate model performance is to compare it with other community models. The main problem with this approach is that many papers extract reported performance scores from original publications without actually implementing the models or using the same dataset. Model performance is likely to depend on feature types and specific samples allocated for training and testing. Thus, this approach makes sense in cases where benchmark datasets have been established with clearly designated training and testing samples. However, this is not yet the case with DRP tasks. Reproducing results from papers is challenging, time-consuming, and often contradicts the reported scores (26). Alternatively, there are papers that diligently reproduce models and use these as baselines. So long as benchmark datasets have not been established, reproducing models and benchmarking with consistent datasets should be considered the standard if one decides to compare a model with another published community model.

Oftentimes, simple analysis can lead to useful insights. For example, in the same study that analyzed baselines for interpretable pathway-based architectures (135), the authors also explored a *naïve* baseline which simply reports the average of drug response values across all samples. While very easy to implement, this baseline performed surprisingly well. This simple and model-agnostic baseline can be very informative not only with respect to a specific model,

but more generally for the entire prediction task. Despite the extra amount of work which may require the design of appropriate baselines, rigorous and comprehensive benchmark analysis is likely to establish fidelity and attract the community.

8. Discussion

8.1. Practices

Development of DRP models is a comprehensive task, success of which depends on multiple factors, such as prepossessing and representation of data, model training, and performance evaluation. Good design choices related to each component of the DRP workflow (Section 2) can positively contribute to model performance, while bad practices can result in adverse effects, leading to poor performance. Here, we provide a short summary of practices that can guide researchers in developing DRP models.

Feature scaling is a common preprocessing step, usually applied to both cancer and drug features. However, when the dataset contains omics data from multiple sources or batches, active measures should be considered for mitigating systematic differences between the batches, which otherwise are likely to bias the downstream ML analysis (51, 136, 137). Possible measures include batch-correction (51) and architectures that are specifically designed to address discrepancies in the input feature space (50, 52). Notice that models can also benefit from standardizing data in the output space. For example, in order to combine response data from multiple sources, the drug response AUC values shall be calculated based on the same dose range (41). Still in the context of *in vitro* drug response, there is a consensus that the global measures of response such AUC and AAC are more robust and produce more predictive models as compared to models trained with IC50 (14, 41, 61).

Due to the diversity of NN architectures and learning schemes (Sections 6.1, 6.2), it is challenging to decipher best practices for DRP model development. Yet, some consensus seems to exist regarding several design principles. Late feature integration exhibits better generalization as compared to early integration approaches (39, 40, 78). Transfer learning and domain adaption significantly improve generalization across datasets and biological cancer models (50, 51). Similar to other application domains, DL for DRP usually requires extensive tuning of HPs to boost performance and integration of early stopping during the training process to avoid model over-fitting. Further comprehensive and systematic studies are required to produce unbiased conclusions regarding DL design methodologies for DRP.

Cross-validation with a DRP model can be performed through different data partitioning schemes, each simulating a different application of the model (Section 7). For example, if authors target the application of precision oncology, then cancer-blind analysis should be considered for model evaluation (Section 7.1.2). Alternatively, if mixed-set analysis is used (Section 7.1.1), the model will produce inflated scores, exhibiting overoptimistic results in the context of precision oncology. Similarly, drug-blind analysis should be used with models targeting drug development applications

which commonly leads to significantly lower performance scores as compared to mixed-set and cancer-blind analysis (Section 7.1.3). Whenever possible and regardless of the chosen evaluation scheme, extensive cross-validation should be performed with multiple data splits (e.g., k-fold cross-validation repeated several times) that would generate enough data points for statistical significance analysis and comparison with baseline models (i.e., mean and standard deviation).

The choice of adequate baseline models is critical for objectively assessing the true capabilities of proposed methods, where ablation analysis plays a central role. Claims made about the predictive power of promising data representations such as molecular graphs (Section 5.3) should be benchmarked against common alternatives such as descriptors and FPs (Section 5.2). Another reasonable baseline is a model trained on one-hot encoded labels of drugs, which utilize only drug identities while ignoring the features (133, 138). Similar one-hot encoding strategies can be applied to cancer features (138) as well as ablation analysis exploring various omics cancer representations (105). Novel NN modules that aim to better leverage DRP data should be benchmarked against alternative and generally less complex modules (e.g., attention modules vs. dense layer). For models integrating cancer pathway data, randomly generated pathways can be used to evaluate the gain obtained by leveraging the pathway information (135). Another baseline is to take the average of sample response values. While very easy to implement, this naïve baseline often exhibits surprisingly good results and can put the model performance into the appropriate perspective, allowing to control for inflated performance scores (135).

Finally, there are two more practices that could be highly beneficial for increasing visibility, but are often neglected. First, many papers finalize their performance analysis by demonstrating performance scores using various validation schemes. Instead, it would be highly beneficial to discuss how the proposed DRP model could be integrated in a larger patient care or drug development workflow and what are the primary challenges preventing these type of models from being deployed in real-world applications. Use case scenario could be highly relevant in this context. The second practice relates to software implementation. Many papers are accompanied by software and data that were used to build their models. However, the code repositories often miss certain pieces of information required to reproduce the paper results. Instead, it is recommended for papers to publish well-documented repositories with reproduce code, including data preprocessing scripts and installation instructions of the computational environment. This practice will improve the usability of models, their adoption as baselines, and visibility within the community.

8.2. Challenges

In recent years, AI approaches have taken a notable place in health care (139). Their promise of utilizing large data arrays to improve treatment protocols, aid drug development efforts, and increase diagnostic precision is attractive to the clinical research community. However, navigating a complex landscape of factors affecting clinical outcomes is a challenging task. We can see it in the number of emerging papers on the DRP problem (140). The research community put in monumental efforts to develop methods that leverage multiomics and drug data, explore the applicability of the existing methods, and bridge the gap between results on ubiquitously used cell lines and patient-derived models. However, this area is facing multiple challenges related to the lack of unified evaluation framework, generalizability, interpretability as well as challenges relating to computational representations of omic data, biological response, and drug representations.

8.2.1. Unified framework for model evaluation and comparison

Currently, there is no standard or accepted framework for evaluation or comparison of cancer drug response models. Model development and performance comparison against baselines is frequently accomplished with different compositions of datasets, inconsistent training and testing splits, and diverse scoring metrics, using varying protocols for hyperparameter optimization or none at all. Yet, the majority of papers report to outperform current state-of-the-art (SOTA) in the task of drug response prediction. In ML, SOTA refers to the best model for a specific task as evaluated by concrete performance criteria on a benchmark dataset of predetermined test set samples. Thousands of benchmark datasets and prediction tasks are publicly available for different applications ranging from vision and language to drug discovery and tumor segmentation. Whereas, most papers follow the same general model development workflow (Figure 2), benchmark datasets and agreed-upon evaluation criteria have yet been established and adopted in the community. It makes identifying the most promising research areas challenging and impedes the research community from making directed efforts to breach them. Creating an ImageNet moment for DRP *via* established benchmark datasets, consistent test sets, robust evaluation criteria, and a platform for publishing and monitoring SOTA models should be on the critical path for our community.

8.2.2. Generalizability

Despite the breadth of methods, the majority of papers have focused on improving predictions in cell lines, often demonstrating only a marginal improvement in prediction generalization. While cell lines remain a primary biological media to study cancer and conduct drug screenings, the potential utility of prediction models for improving patient care is not immediately evident, and several questions naturally arise. How well models trained with cell lines would generalize to xenografts or patients? How much one could rely on any given model prediction in a decision-making process? What are the potential clinical application for these models? Addressing questions important from the user perspective and suggesting use case scenarios have not been the driving mechanism behind the majority of published models. Yet, certain trends aiming to respond to these challenges emerge: transfer learning that utilize abundant cell line datasets to improve predictions in PDX and patients, uncertainty quantification allowing to estimate the confidence for each model prediction (51), and ranking learning models principally suitable for personalized treatment recommendation (72, 74, 141). Demonstrating the utility of DRP models, integrated

in a larger patient care workflow, could provide the needed user-centric view and navigate the community toward developing application-aware models.

8.2.3. Interpretability

Despite the demonstrated success of the DL on expansive datasets, delegating the decision-making process on patient well-being to the black box is widely contested (142). Thus, providing not only precise but also salient results is a task of paramount importance for the clinical community. To address this challenge, model-agnostic methods such as Integrated Gradients (143) and SHAP (144), designed to explain model predictions, have been applied post-training to DRP models (42, 45, 145), where the explanation is provided in the form of most important features attributing to model predictions. However, it has been shown that these methods can lead to highly misleading information in applications more comprehensible to humans such as image classification, and therefore, attempting to explain black boxes with *post-hoc* methods is perceived as dangerous in high-stakes decision-making domains (146). Instead of explaining predictions, the alternative is to design interpretable models that are understandable by domain experts or provide insights into the decision-making process (147, 148). Unfortunately, no single definition exists, and identifying a model as interpretable is considered domain-specific (146, 147). Several papers referring to their DRP models as interpretable integrate domain structural knowledge into the model form (145, 149–151). Currently, no clear definition of an interpretable DRP model yet exists, nor the extent to which it might improve cancer treatment is known. Therefore, it is challenging to evaluate whether existing efforts in this direction are in line with the views of the clinical community at large on this important matter. Addressing the interpretability issue will perhaps require better framing of what interpretable DRP models are and how much performance drop could be tolerated, if any, for improved interpretability. As discussed in Section 7.3, baseline models that can assess the quality and robustness of model interpretability can serve to bolster the claim regarding the benefits and usability of interpretable models (135). Collaboration with cancer biologists and clinical oncologists would be essential to advance this direction. And, as discussed in Section 7.3, baseline models which compare the quality and robustness of interpretability can serve to bolster claims regarding the improvement or introduction of model interpretability.

8.2.4. Variability and high dimensionality associated with omics data

Omics values differ substantially depending on the underlying biological model (e.g., cell line, organoid, xenograft), experimental protocols and selected platforms (e.g., Illumina, Nanopore), technical variations (batch effect), and computational pipeline for processing raw data. Identifying the potential sources of variation and taking directed measures to mitigate the undesired differences is critical, especially in cross-domain generalization scenarios where the similarity between the source and target domains is a fundamental assumption (50–52, 67). A major conclusion of the NCI-DREAM challenge was that gene expression modality provides most of the predictive power for cell line drug sensitivity prediction with additional improvement when combined with other omics types

(27). This collaborative effort contributed to the adoption of gene expression (GE; e.g., microarrays or RNA-Seq) or combining them with other omics in DRP models. Integrating multiomics into the learning process further exacerbates the already problematic feature size of single omics data. A few papers indeed demonstrate significant performance boost with multiomics (45, 49, 108) but the majority report only marginal improvement (104, 108, 118, 126, 141, 152). As opposed to DREAM participators which utilized agreed-upon datasets and scoring metrics, the DRP models in Supplementary Table 1 substantially differ among them as discussed earlier, largely contributing to discrepancies and mixed conclusions. Analysis across multiple models and omics types is required to evaluate the predictive capabilities of individual data types and their subsets and make unbiased and coherent conclusions. Such analysis should incorporate recent trends which encode biological information such as protein-protein interactions (PPI), gene correlations, and pathway information (45, 49, 109, 111).

8.2.5. Representation of biological treatment response

In vitro response data (e.g., cell line and organoid) is usually derived from a series of inherently noisy cell viability experiments and subsequent curve fitting. However, these curves do not always have a good fit, and usage of a single point on the fitted curve may result in a substantial information loss. While the use of IC50 as a prediction variable has been the most prevalent in regression models (Figure 4), recent assessment studies could shift this trend toward the use of global and generally more robust measures such as AUC and AAC (14, 41). With the objective of utilizing prediction models in decision-making situations, another option is converting continuous into discrete values. This is particularly common in cross-domain generalization from *in vitro* to *in vivo*, where *in vivo* response is generally encoded as discrete values. The third option is generating a rank list of items. This has been applied to a personalized treatment recommendation given cancer information. Surprisingly, despite the success of recommendation algorithms and the direct relevance to precision oncology, only few methods have been explored (72, 74, 141).

8.2.6. Vastness of chemical space and its representations

Reviewed papers consistently report the response prediction accuracy drop for compounds, previously unseen by the model. Inability to reliably overcome this limitation severely limits the practical value of DRP models to virtual drug screening or drug design applications. Representing drug molecules with graphs and using GNN for learning response prediction has recently inspired many DRP architectures. The use of graph structures is motivated by the proposition that molecular graphs better capture intrinsic chemical properties of molecules (49, 108, 153). Whereas, this is presumably expected to produce a better prediction of drug response, we have not found a comprehensive study that could assert this hypothesis. In fact, in a related field of molecular property prediction, a comparison with multiple datasets and prediction tasks suggests that on average models that use FPs or descriptors outperform graph-based models (154). We have found only one study comparing side-by-side the added value of molecular graphs

against SMILES, reporting <0.5% improvement as evaluated by Pearson correlation coefficient (155). As in the case of cancer representation, further studies are required across various models and datasets to assess the predictive capabilities of molecular graphs and other drug representations to DRP. Several underexplored but interesting directions include SMILES combined with transformer-based models, kinase inhibition profiles representing kinase inhibitor therapies (44), and 2D-CNNs that learn from FPs or descriptors transformed into images (77).

This list of challenges is in no way conclusive, as we refrained from studying even more complicated scenarios emerging in polypharmacy, even though the treatment of patients in the clinic often involves combination therapy. Meanwhile, most of the prediction models to date focus on single-drug treatments. This highlights a significant disconnect between the modeling community and the current patient standard of care. However, even this simplified problem raises many obstacles, and making progress in overcoming them would significantly benefit the scientific community because they are relevant to a much wider domain of AI-driven biological research. Each of the highlighted problems has multiple potential solutions, and we hope to see subsequent progress in the near future, allowing AI to deliver on its grand promise.

Author contributions

Writing the original draft: AP. Collection of prediction models and other references: AP, TB, YZ, and AC. Review and editing: AP, TB, YZ, ON, AC, JO, and RS. All authors have read and agreed to the published version of the manuscript.

Funding

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1086097/full#supplementary-material

References

1. Yancovitz M, Litterman A, Yoon J, Ng E, Shapiro RL, Berman RS, et al. Intra-and inter-tumor heterogeneity of BRAFV600EMutations in primary and metastatic melanoma. *PLoS ONE*. (2012) 7:e0029336. doi: 10.1371/journal.pone.0029336

2. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer*. (2013) 108:479–85. doi: 10.1038/bjc.2012.581

3. Collins DC, Sundar R, Lim JSJ, Yap TA. Towards precision medicine in the clinic: from biomarker discovery to novel therapeutics. *Trends Pharmacol Sci*. (2017) 38:25–40. doi: 10.1016/j.tips.2016.10.012

4. Hartl D, de Luca V, Kostikova A, Laramie J, Kennedy S, Ferrero E, et al. Translational precision medicine: an industry perspective. *J Transl Med*. (2021) 19:245. doi: 10.1186/s12967-021-02910-6

5. Prasad V, Kaestner V, Mailankody S. Cancer drugs approved based on biomarkers and not tumor Type-FDA approval of pembrolizumab for mismatch repair-deficient solid cancers. *JAMA Oncol*. (2018) 4:157–8. doi: 10.1001/jamaoncol.2017.4182

6. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol*. (2020) 4:19. doi: 10.1038/s41698-020-0122-1

7. Ballester PJ, Stevens R, Haibe-Kains B, Huang RS, Aittokallio T. Artificial intelligence for drug response prediction in disease models. *Brief Bioinform*. (2021) 23:bbab450. doi: 10.1093/bib/bbab450

8. De Niz C, Rahman R, Zhao X, Pal R. Algorithms for drug sensitivity prediction. *Algorithms*. (2016) 9:77. doi: 10.3390/a9040077

9. Piyawajanusorn C, Nguyen LC, Ghislat G, Ballester PJ. A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling. *Brief Bioinform*. (2021) 22:bbab312. doi: 10.1093/bib/bbab312

10. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Brief Bioinform*. (2021) 22:360–79. doi: 10.1093/bib/bbz171

11. Kumar V, Dogra N. A comprehensive review on deep synergistic drug prediction techniques for cancer. *Arch Comput Methods Eng*. (2022) 29:1443–61. doi: 10.1007/s11831-021-09617-3

12. Azuaje F. Computational models for predicting drug responses in cancer research. *Brief Bioinform*. (2017) 18:820–9. doi: 10.1093/bib/bbw065

13. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev*. (2019) 11:31–9. doi: 10.1007/s12551-018-0446-z

14. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, Hon C, Mammoliti A, Nair SK, et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief Bioinform*. (2021) 22:bbab294. doi: 10.1093/bib/bbab294

15. Güvenç Paltun B, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Brief Bioinform*. (2021) 22:346–59. doi: 10.1093/bib/bbz153

16. Rafique R, Islam SMR, Kazi JU. Machine learning in the prediction of cancer therapy. *Comput Struct Biotechnol J*. (2021) 19:4003–17. doi: 10.1016/j.csbj.2021.07.003

17. Zadorozhny K, Nuzhna L. Deep generative models for drug design and response. *arXiv:210906469 [cs]* (2021). doi: 10.48550/arXiv.2109.06469

18. Firoozbakht F, Yousefi B, Schwikowski B. An overview of machine learning methods for monotherapy drug response prediction. *Brief Bioinform.* (2022) 23:bbab408. doi: 10.1093/bib/bbab408

19. Chiu YC, Chen HIH, Gorthi A, Mostavi M, Zheng S, Huang Y, et al. Deep learning of pharmacogenomics resources: moving towards precision oncology. *Brief Bioinform.* (2020) 21:2066–83. doi: 10.1093/bib/bbz144

20. Tanoli Z, Vähä-Koskela M, Aittokallio T. Artificial intelligence, machine learning, and drug repurposing in cancer. *Expert Opin Drug Disc.* (2021) 16:977–89. doi: 10.1080/17460441.2021.1883585

21. An X, Chen X, Yi D, Li H, Guan Y. Representation of molecules for drug response prediction. *Brief Bioinform.* (2021) 23:bbab393. doi: 10.1093/bib/bbab393

22. Wu L, Wen Y, Leng D, Zhang Q, Dai C, Wang Z, et al. Machine learning methods, databases and tools for drug combination prediction. *Brief Bioinform.* (2022) 23:bbab355. doi: 10.1093/bib/bbab355

23. Wu Z, Lawrence PJ, Ma A, Zhu J, Xu D, Ma Q. Single-cell techniques and deep learning in predicting drug response. *Trends Pharmacol Sci.* (2020) 41:1050–65. doi: 10.1016/j.tips.2020.10.004

24. Chen J, Zhang L. A survey and systematic assessment of computational methods for drug response prediction. *Brief Bioinform.* (2021) 22:232–46. doi: 10.1093/bib/bbz164

25. Koras K, Juraeva D, Kreis J, Mazur J, Staub E, Szczurek E. Feature selection strategies for drug sensitivity prediction. *Sci Rep.* (2020) 10:9377. doi: 10.1038/s41598-020-65927-9

26. Chen Y, Zhang L. How much can deep learning improve prediction of the responses to drugs in cancer cell lines? *Brief Bioinform.* (2021) 23:bbab378. doi: 10.1093/bib/bbab378

27. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol.* (2014) 32:1202–12. doi: 10.1038/nbt.2877

28. Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun.* (2019) 10:2674. doi: 10.1038/s41467-019-09799-2

29. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539

30. Stevens R, Taylor V, Nichols J, Maccabe AB, Yelick K, Brown D. AI for science: report on the department of energy (DOE) town halls on artificial intelligence (AI) for science. In: *Argonne National Lab. (ANL)*, Argonne, IL (2020).

31. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling climate change with machine learning. *ArXiv:1906.05433 [cs, stat]* (2019). doi: 10.48550/arXiv.1906.05433

32. Cohen R, Zhang Y, Solomon O, Toberman D, Taieb L, van Sloun RJ, et al. Deep convolutional robust PCA with application to ultrasound imaging. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Brighton, UK: IEEE (2019). p. 3212–6.

33. Utyamishev D, Partin-Vaisband I. Multiterminal pathfinding in practical VLSI systems with deep neural networks. *ACM Trans Design Autom Electron Syst.* (2022) 2022:3564930. doi: 10.1145/3564930

34. Menden MP, Iorio F, Garnett M, Mcdermott U, Benes CH. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS ONE.* (2013) 8:61318. doi: 10.1371/journal.pone.0061318

35. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* (2013) 41:955–61. doi: 10.1093/nar/gks1111

36. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous systems. *arXiv:1603.04467 [cs.DC]* (2015). doi: 10.48550/arXiv.1603.04467

37. Chollet F, others. *Keras.* (2015). Available online at: https://keras.io

38. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32.* Vancouver, BC: Curran Associates, Inc. (2019). p. 8024–35.

39. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics.* (2019) 35:501–9. doi: 10.1093/bioinformatics/btz318

40. Partin A, Brettin T, Evrard YA, Zhu Y, Yoo H, Xia F, et al. Learning curves for drug response prediction in cancer cell lines. *BMC Bioinform.* (2021) 22:252. doi: 10.1186/s12859-021-04163-y

41. Xia F, Allen J, Balaprakash P, Brettin T, Garcia-Cardona C, Clyde A, et al. A cross-study analysis of drug response prediction in cancer cell lines. *Brief Bioinform.* (2022) 23:bbab356. doi: 10.1093/bib/bbab356

42. Prasse P, Iversen P, Lienhard M, Thedinga K, Herwig R, Scheffer T. Pre-Training on *in vitro* and fine-tuning on patient-derived data improves deep neural networks for anti-cancer drug-sensitivity prediction. *Cancers.* (2022) 14:3950. doi: 10.3390/cancers14163950

43. Mazumder M, Banbury C, Yao X, Karlaš B, Rojas WG, Diamos S, et al. DataPerf: benchmarks for data-centric AI development. *arXiv:2207.10062 [cs.LG]* (2022). doi: 10.48550/arXiv.2207.10062

44. Koras K, Kizling E, Juraeva D, Staub E, Szczurek E. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. *Sci Rep.* (2021) 11:15993. doi: 10.1038/s41598-021-94564-z

45. Tang YC, Gottlieb A. Explainable drug sensitivity prediction through cancer pathway enrichment. *Sci Rep.* (2021) 11:3128. doi: 10.1038/s41598-021-82612-7

46. Partin A, Brettin T, Zhu Y, Dolezal JM, Kochanny S, Pearson AT, et al. Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images. *ArXiv:2204.11678 [q-bio]* (2022). doi: 10.48550/arXiv.2204.11678

47. Hostallero DE, Wei L, Wang L, Cairns J, Emad A. A deep learning framework for prediction of clinical drug response of cancer patients and identification of drug sensitivity biomarkers using preclinical samples. *biorxiv.* (2021). doi: 10.1101/2021.07.06.451273

48. Jia P, Hu R, Pei G, Dai Y, Wang YY, Zhao Z. Deep generative neural network for accurate drug response imputation. *Nat Commun.* (2021) 12:1740. doi: 10.1038/s41467-021-21997-5

49. Ma T, Liu Q, Li H, Zhou M, Jiang R, Zhang X. DualGCN: a dual graph convolutional network model to predict cancer drug response. *BMC Bioinform.* (2022) 23:129. doi: 10.1186/s12859-022-04664-4

50. Sharifi-Noghabi H, Peng S, Zolotareva O, Collins CC, Ester M. AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics.* (2020) 36:i380–8. doi: 10.1093/bioinformatics/btaa442

51. Peres da Silva R, Suphavilai C, Nagarajan N. TUGDA: task uncertainty guided domain adaptation for robust generalization of cancer drug response prediction from *in vitro* to *in vivo* settings. *Bioinformatics.* (2021) 37:i76-i83. doi: 10.1093/bioinformatics/btab299

52. Sharifi-Noghabi H, Harjandi PA, Zolotareva O, Collins CC, Ester M. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nat Mach Intell.* (2021) 3:962–72. doi: 10.1038/s42256-021-00408-w

53. Mourragui S, Loog M, van de Wiel MA, Reinders MJT, Wessels LFA. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics.* (2019) 35:i510–9. doi: 10.1093/bioinformatics/btz372

54. Mourragui SMC, Loog M, Vis DJ, Moore K, Manjon AG, van de Wiel MA, et al. Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning. *Proc Natl Acad Sci USA.* (2021) 118:2118. doi: 10.1073/pnas.2106682118

55. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* (2012) 483:603–7. doi: 10.1038/nature11003

56. Iorio F, Knijnenburg TA, Vis DJ, Saez-Rodriguez J, Mcdermott U, Garnett MJ, et al. A landscape of pharmacogenomic interactions in cancer. *Cell.* (2016) 166:740–54. doi: 10.1016/j.cell.2016.06.017

57. Seashore-Ludlow B, Rees MG, Cheah JH, Coko M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* (2015) 5:1210–23. doi: 10.1158/2159-8290.CD-15-0235

58. Larsen BM, Kannan M, Langer LF, Leibowitz BD, Bentaieb A, Cancino A, et al. A pan-cancer organoid platform for precision medicine. *Cell Rep.* (2021) 36:109429. doi: 10.1016/j.celrep.2021.109429

59. Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat Med.* (2015) 21:1318–25. doi: 10.1038/nm.3954

60. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac Symp Biocomput.* (2013) 2014: 63–74. doi: 10.1142/9789814583220_0007

61. Yadav B, Pemovska T, Szwajda A, Kulesskiy E, Kontro M, Karjalainen R, et al. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci Rep.* (2014) 4:5193. doi: 10.1038/srep05193

62. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, et al. Inconsistency in large pharmacogenomic studies. *Nature.* (2013) 504:389–93. doi: 10.1038/nature12831

63. Knijnenburg TA, Klau GW, Iorio F, Garnett MJ, McDermott U, Shmulevich I, et al. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Sci Rep.* (2016) 6:36812. doi: 10.1038/srep36812

64. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci USA.* (2001) 98:10787–92. doi: 10.1073/pnas.191368598

65. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer.* (2015) 15:489. doi: 10.1186/s12885-015-1492-6

66. Joo M, Park A, Kim K, Son WJ, Lee HS, Lim G, et al. A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. *Int J Mol Sci.* (2019) 20:6276. doi: 10.3390/ijms20246276

67. Sakellaropoulos T, Vougas K, Narang S, Koinis F, Kotsinas A, Polyzos A, et al. A deep learning framework for predicting response to therapy in cancer. *Cell Rep.* (2019) 29:3367–73.e4. doi: 10.1016/j.celrep.2019.11.017

68. Daoud S, Mdhaffar A, Jmaiel M, Freisleben B. Q-Rank: Reinforcement Learning for Recommending Algorithms to Predict Drug Sensitivity to Cancer Therapy. *IEEE J Biomed Health Inform.* (2020) 24:3154–61. doi: 10.1109/JBHI.2020.3004663

69. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics.* (2021) 22:214. doi: 10.1186/s12864-021-07524-2

70. Jiang L, Jiang C, Yu X, Fu R, Jin S, Liu X. DeepTTA: a transformer-based model for predicting cancer drug response. *Brief Bioinform.* (2022) 23:bbac100. doi: 10.1093/bib/bbac100

71. Zhao Z, Li K, Toumazou C, Kalofonou M. A computational model for anti-cancer drug sensitivity prediction. In: *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS).* Nara: IEEE (2019). p. 1–4.

72. Prasse P, Iversen P, Lienhard M, Thedinga K, Bauer C, Herwig R, et al. Matching anticancer compounds and tumor cell lines by neural networks with ranking loss. *NAR Genomics Bioinform.* (2022) 4:lqab128. doi: 10.1093/nargab/lqab128

73. Manica M, Oskooei A, Born J, Subramanian V, Sáez-Rodríguez J, Rodríguez Martínez M. Toward explainable anticancer compound sensitivity prediction *via* multimodal attention-based convolutional encoders. *Mol Pharm.* (2019) 16:4797–806. doi: 10.1021/acs.molpharmaceut.9b00520

74. Su R, Huang Y, Zhang Dg, Xiao G, Wei L. SRDFM: siamese response deep factorization machine to improve anti-cancer drug recommendation. *Brief Bioinform.* (2022) 23:bbab534. doi: 10.1093/bib/bbab534

75. Schwartz LH, Litière S, Vries Ed, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1–Update and clarification: From the RECIST committee. *Eur J Cancer.* (2016) 62:132–7. doi: 10.1016/j.ejca.2016.03.081

76. Bazgir O, Zhang R, Dhruba SR, Rahman R, Ghosh S, Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nat Commun.* (2020) 11:4391. doi: 10.1038/s41467-020-18197-y

77. Zhu Y, Brettin T, Xia F, Partin A, Shukla M, Yoo H, et al. Converting tabular data into images for deep learning with convolutional neural networks. *Sci Rep.* (2021) 11:11325. doi: 10.1038/s41598-021-90923-y

78. Zhu Y, Brettin T, Evrard YA, Partin A, Xia F, Shukla M, et al. Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci Rep.* (2020) 10:1–11. doi: 10.1038/s41598-020-74921-0

79. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J Cheminform.* (2020) 12:56. doi: 10.1186/s13321-020-00460-5

80. Oskooei A, Born J, Manica M, Subramanian V, Sáez-Rodríguez J, Rodríguez Martínez M. PaccMann: prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. In: *NIPS.* Montreal, QC (2018).

81. Liu H, Zhao Y, Zhang L, Chen X. Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Mol Therapy Nucleic Acids.* (2018) 13:303–11. doi: 10.1016/j.omtn.2018.09.011

82. Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv:1703.07076 [cs.LG]* (2017). doi: 10.48550/arXiv.1703.07076

83. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond JL, et al. Randomized SMILES strings improve the quality of molecular generative models. *J Cheminform.* (2019) 11:71. doi: 10.1186/s13321-019-0393-0

84. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem.* (2010) 32:1466–74. doi: 10.1002/jcc.21707

85. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. *J Cheminform.* (2018) 10:4. doi: 10.1186/s13321-018-0258-y

86. Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON software: an easy approach to molecular descriptor calculations. *MATCH Commun Math Comput Chem.* (2006) 56:237–48.

87. Clyde A, Brettin T, Partin A, Shaulik M, Yoo H, Evrard Y, et al. A systematic approach to featurization for cancer drug sensitivity predictions with deep learning. *arXiv:200500095 [cs, q-bio].* (2020). doi: 10.48550/arXiv.2005.00095

88. Kipf TN, Welling M. Semi-Supervised classification with graph convolutional networks. In: *International Conference on Learning Representations.* Toulon (2017).

89. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. In: *International Conference on Learning Representations.* Vancouver, BC (2018).

90. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. *arXiv:170602216 [cs, stat].* (2018). doi: 10.48550/arXiv.1706.02216

91. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics.* (2018) 34:i457–66. doi: 10.1093/bioinformatics/bty294

92. Utyamishev D, Partin-Vaisband I. Knowledge graph embedding and visualization for pre-silicon detection of hardware Trojans. In: *2022 IEEE International Symposium on Circuits and Systems (ISCAS).* Austin, TX: IEEE (2022). p. 1–5.

93. Sun M, Zhao S, Gilvary C, Elemento O, Zhou J, Wang F. Graph convolutional networks for computational drug development and discovery. *Brief Bioinform.* (2020) 21:919–35. doi: 10.1093/bib/bbz042

94. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. *Brief Bioinform.* (2021) 23:bbab159. doi: 10.1093/bib/bbab159

95. Ramsundar B, Eastman P, Walters P, Pande V, Leswing K, Wu Z. *Deep Learning for the Life Sciences.* O'Reilly Media (2019). Available online at: https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837

96. Grattarola D, Alippi C. Graph neural networks in tensorflow and Keras with Spektral. *arXiv:2006.12138 [cs.LG].* (2020). doi: 10.1109/MCI.2020.3039072

97. Li M, Zhou J, Hu J, Fan W, Zhang Y, Gu Y, et al. DGL-LifeSci: an open-source toolkit for deep learning on graphs in life science. *ACS Omega.* (2021) 6:27233–8. doi: 10.1021/acsomega.1c04017

98. Jiang Y, Rensi S, Wang S, Altman RB. DrugOrchestra: Jointly predicting drug response, targets, and side effects *via* deep multi-task learning. *biorxiv.* (2020). doi: 10.1101/2020.11.17.385757

99. Choi J, Park S, Ahn J. RefDNN: a reference drug based neural network for more accurate prediction of anticancer drug resistance. *Sci Rep.* (2020) 10:1861. doi: 10.1038/s41598-020-58821-x

100. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: a survey. *Mech Syst Signal Process.* (2021) 151:107398. doi: 10.1016/j.ymssp.2020.107398

101. Liu P, Li H, Li S, Leung KS. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinform.* (2019) 20:408. doi: 10.1186/s12859-019-2910-6

102. Chang Y, Park H, Yang HJ, Lee S, Lee KY, Kim TS, et al. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci Rep.* (2018) 8:1–11. doi: 10.1038/s41598-018-27214-6

103. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics.* (2020) 36:i911–8. doi: 10.1093/bioinformatics/btaa822

104. Chu T, Nguyen TT, Hai BD, Nguyen QH, Nguyen T. Graph transformer for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform.* (2022) 1–9. doi: 10.1109/TCBB.2022.3206888

105. Nguyen GTT, Vu HD, Le DH. Integrating molecular graph data of drugs and multiple -Omic data of cell lines for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform.* (2022) 19:710–7. doi: 10.1109/TCBB.2021.3096960

106. Zuo Z, Wang P, Chen X, Tian L, Ge H, Qian D. SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. *BMC Bioinform.* (2021) 22:434. doi: 10.1186/s12859-021-04352-9

107. Tao Y, Ren S, Ding MQ, Schwartz R, Lu X. Predicting drug sensitivity of cancer cell lines *via* collaborative filtering with contextual attention. In: *Proceedings of the 5th Machine Learning for Healthcare Conference.* PMLR (2020). p. 660–84.

108. Liu X, Song C, Huang F, Fu H, Xiao W, Zhang W. GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Brief Bioinform.* (2021) 23:bbab457. doi: 10.1093/bib/bbab457

109. Ruiwei F, Yufeng X, Minshan L, Danny C, Ji C, Jian W. AGMI: attention-guided multi-omics integration for drug response prediction with graph neural networks. *arXiv:211208366 [cs, q-bio].* (2021). doi: 10.48550/arXiv.2112.08366

110. Jin I, Nam H. HiDRA: hierarchical network for drug response prediction with attention. *J Chem Inf Model.* (2021) 61:3858–67. doi: 10.1021/acs.jcim.1c00706

111. Pu L, Singha M, Ramanujam J, Brylinski M. CancerOmicsNet: a multi-omics network-based approach to anti-cancer drug profiling. *Oncotarget.* (2022) 13:695–706. doi: 10.18632/oncotarget.28234

112. Huang K, Xiao C. Explainable substructure partition fingerprint for protein, drug, and more. In: *NeurIPS.* Vancouver, BC (2019).

113. Yi HC, You ZH, Huang DS, Kwoh CK. Graph representation learning in bioinformatics: trends, methods and applications. *Brief Bioinform.* (2022) 23:bbab340. doi: 10.1093/bib/bbab340

114. Nguyen TT, Nguyen GTT, Nguyen T, Le DH. Graph convolutional networks for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform.* (2021) 19:146–54. doi: 10.1101/2020.04.07.030908

115. Ahmed KT, Park S, Jiang Q, Yeu Y, Hwang T, Zhang W. Network-based drug sensitivity prediction. *BMC Med Genomics.* (2020) 13:193. doi: 10.1186/s12920-020-00829-3

116. Peng W, Chen T, Dai W. Predicting drug response based on multi-omics fusion and graph convolution. *IEEE J Biomed Health Inform.* (2022) 26:1384–93. doi: 10.1109/JBHI.2021.3102186

117. Kim S, Bae S, Piao Y, Jo K. Graph convolutional network for drug response prediction using gene expression Data. *Mathematics.* (2021) 9:772. doi: 10.3390/math9070772

118. Zhu Y, Ouyang Z, Chen W, Feng R, Chen DZ, Cao J, et al. TGSA: protein-protein association-based twin graph neural networks for drug response prediction with similarity augmentation. *Bioinformatics*. (2021) 38:461–8. doi: 10.1093/bioinformatics/btab650

119. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. (2021) 49:D605–12. doi: 10.1093/nar/gkaa1074

120. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. (2005) 102:15545–50. doi: 10.1073/pnas.0506580102

121. Hostallero DE, Li Y, Emad A. Looking at the BiG picture: incorporating bipartite graphs in drug response prediction. *Bioinformatics*. (2022) 38:3609–20. doi: 10.1093/bioinformatics/btac383

122. Li M, Wang Y, Zheng R, Shi X, Li Y, Wu FX, et al. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans Comput Biol Bioinform*. (2021) 18:575–82. doi: 10.1109/TCBB.2019.2919581

123. Rampasek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr.VAE: improving drug response prediction *via* modeling of drug perturbation effects. *Bioinformatics*. (2019) 35:3743–51. doi: 10.1093/bioinformatics/btz158

124. Xu X, Gu H, Wang Y, Wang J, Qin P. Autoencoder based feature selection method for classification of anticancer drug response. *Front Genet*. (2019) 10:233. doi: 10.3389/fgene.2019.00233

125. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol Cancer Res*. (2018) 16:269–78. doi: 10.1158/1541-7786.MCR-17-0378

126. Chiu YC, Chen HIH, Zhang T, Zhang S, Gorthi A, Wang LJ, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics*. (2019) 12:18. doi: 10.1186/s12920-018-0460-9

127. Dong H, Xie J, Jing Z, Ren D. Variational autoencoder for anti-cancer drug response prediction. *arXiv:2008.09763 [cs.LG]* (2021). doi: 10.48550/arXiv.2008.09763

128. Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR (2018). p. 2323–32.

129. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. (2010) 22:1345–59. doi: 10.1109/TKDE.2009.191

130. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. *Proc IEEE*. (2021) 109:43–76. doi: 10.1109/JPROC.2020.3004555

131. Ma J, Fong SH, Luo Y, Bakkenist CJ, Shen JP, Mourragui S, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer*. (2021) 2:233–44. doi: 10.1038/s43018-020-00169-2

132. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre-training graph neural networks. In: *International Conference on Learning Representations*. Addis Ababa (2020).

133. Zhu Y, Brettin T, Evrard YA, Xia F, Partin A, Shukla M, et al. Enhanced co-expression extrapolation (coxen) gene selection method for building anti-cancer drug response prediction models. *Genes*. (2020) 11:1070. doi: 10.3390/genes11091070

134. Gupta A, Gautam P, Wennerberg K, Aittokallio T. A normalized drug response metric improves accuracy and consistency of anticancer drug sensitivity quantification in cell-based screening. *Commun Biol*. (2020) 3:1–12. doi: 10.1038/s42003-020-0765-z

135. Li Y, Hostallero DE, Emad A. Interpretable deep learning architectures for improving drug response prediction: myth or reality? *biorxiv*. (2022). doi: 10.1101/2022.10.03.510614

136. Dincer AB, Janizek JD, Lee SI. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*. (2020) 36:i573–82. doi: 10.1093/bioinformatics/btaa796

137. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, et al. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*. (2017) 33:2539–46. doi: 10.1093/bioinformatics/btx196

138. Xia F, Shukla M, Brettin T, Garcia-Cardona C, Cohn J, Allen JE, et al. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinform*. (2018) 19:486. doi: 10.1186/s12859-018-2509-3

139. Bohr A, Memarzadeh K. Chapter 2 - the rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K, editors. *Artificial Intelligence in Healthcare*. Academic Press (2020). p. 25–60. Available online at: https://www.sciencedirect.com/science/article/pii/B9780128184387000022

140. Clyde A. AI for science and global citizens. *Patterns*. (2022) 3:100446. doi: 10.1016/j.patter.2022.100446

141. Liu M, Shen X, Pan W. Deep reinforcement learning for personalized treatment recommendation. *Stat Med*. (2022) 41:4034–56. doi: 10.1002/sim.9491

142. Naik N, Hameed BMZ, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg*. (2022) 9:862322. doi: 10.3389/fsurg.2022.862322

143. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34 th International Conference on Machine Learning*. Sydney, NSW (2017).

144. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Neural Information Processing Systems*. Long Beach, CA (2017).

145. Snow O, Sharifi-Noghabi H, Lu J, Zolotareva O, Lee M, Ester M. Interpretable drug response prediction using a knowledge-based neural network. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21*. New York, NY: Association for Computing Machinery (2021). p. 3558–68.

146. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x

147. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci USA*. (2019) 116:22071–80. doi: 10.1073/pnas.1900654116

148. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. (2020) 58:82–115. doi: 10.1016/j.inffus.2019.12.012

149. Deng L, Cai Y, Zhang W, Yang W, Gao B, Liu H. Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. *J Chem Inf Model*. (2020) 60:4497–505. doi: 10.1021/acs.jcim.0c00331

150. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*. (2020) 38:672–84. doi: 10.1016/j.ccell.2020.09.014

151. Zhang H, Chen Y, Li F. Predicting anticancer drug response with deep learning constrained by signaling pathways. *Front Bioinform*. (2021) 1:639349. doi: 10.3389/fbinf.2021.639349

152. Park S, Soh J, Lee H. Super.FELT: supervised feature extraction learning using triplet loss for drug response prediction with multi-omics data. *BMC Bioinform*. (2021) 22:269. doi: 10.1186/s12859-021-04146-z

153. Hosseini R, Simini F, Clyde A, Ramanathan A. Deep surrogate docking: accelerating automated drug discovery with graph neural networks. In: *Workshop on AI for Science: Progress and Promises*. New Orleans, LA (2022).

154. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Cheminform*. (2021) 13:12. doi: 10.1186/s13321-020-00479-8

155. Yan X, Liu Y, Zhang W. Deep graph and sequence representation learning for drug response prediction. In: *Artificial Neural Networks and Machine Learning-ICANN*. Bristol: Springer (2022). p. 97–108.