Codicè et al. Journal of Cheminformatics (2025) 17:33
https://doi.org/10.1186/s13321-025-00972-y

Journal of Cheminformatics

RESEARCH

Open Access

The specification game: rethinking the evaluation of drug response prediction for precision oncology

Francesco Codicè[1*], Corrado Pancotti[1], Cesare Rollo[1], Yves Moreau[2], Piero Fariselli[1] and Daniele Raimondi[3]

Abstract

Precision oncology plays a pivotal role in contemporary healthcare, aiming to optimize treatments for each patient based on their unique characteristics. This objective has spurred the emergence of various cancer cell line drug response datasets, driven by the need to facilitate pre-clinical studies by exploring the impact of multi-omics data on drug response. Despite the proliferation of machine learning models for Drug Response Prediction (DRP), their validation remains critical to reliably assess their usefulness for drug discovery, precision oncology and their actual ability to *generalize* over the immense space of cancer cells and chemical compounds.

Scientific contribution In this paper we show that the commonly used evaluation strategies for DRP methods can be easily *fooled* by commonly occurring dataset biases, and they are therefore not able to truly measure the ability of DRP methods to generalize over drugs and cell lines ("specification gaming"). This problem hinders the development of reliable DRP methods and their application to experimental pipelines. Here we propose a new validation protocol composed by three Aggregation Strategies (Global, Fixed-Drug, and Fixed-Cell Line) integrating them with three of the most commonly used train-test evaluation settings, to ensure a truly realistic assessment of the prediction performance. We also scrutinize the challenges associated with using IC50 as a prediction label, showing how its close correlation with the drug concentration ranges worsens the risk of misleading performance assessment, and we indicate an additional reason to replace it with the Area Under the Dose-Response Curve instead.

Keywords Cancer, Precision medicine, Drug response prediction, Deep learning, Validation protocol

Introduction

One of the main goals of precision oncology is to deliver the right drugs in the right doses, on the basis of the specific characteristics of each patient [1]. Biological challenges such as tumor heterogeneity, where diverse cellular subpopulations drive conflicting drug responses, and acquired resistance mechanisms present significant barriers to achieving consistent precision oncology outcomes. These barriers persist even with comprehensive molecular profiling, as tumors dynamically evolve through genetic and epigenetic adaptations or interactions with the microenvironment. In order to improve this aspect of the clinical practice we are in need of reliable preclinical models [2]. Large datasets containing drug response measurements on cancer cell lines have been published, such as the National Cancer Institute 60 (NCI60) [3], the Cancer Cell Line Encyclopedia (CCLE) [4], the Genomics of Drug Sensitivity in Cancer (GDSC) [5], and the Cancer Therapeutics Response Portal (CTRP) [6]. The cell lines in these datasets represent

*Correspondence:
Francesco Codicè
francesco.codice@unito.it
[1] Department of Medical Sciences, University of Torino, 10123 Torino, Italy
[2] ESAT-STADIUS, KU Leuven, Leuven 3001, Belgium
[3] Institut de Génétique Moléculaire de Montpellier, Université de Montpellier, 34293 Montpellier, France

different types of cancer and are usually characterized by various omics data, including sequencing, transcriptomics, proteomics, and methylation data [7].

These data can be used to design computational models that serve as in silico alternatives to in vitro cell viability screenings [7, 8], providing tailored predictions of drug response across various cell lines. Such Drug Response Prediction (DRP) models would be particularly useful if they could generalize to unobserved drugs or cancer types [9], providing information about potential drug candidates for further analysis, thereby expediting the cancer drug discovery process [9].

Community challenges, such as DREAM, have contributed to DRP modeling, as demonstrated by the 2014 NCI-DREAM Drug Sensitivity Prediction Challenge which evaluated 44 models on unseen breast cancer cell lines and showed that non-linear approaches and multi-omics integration improved predictive performance [10]. Among the various approaches for DRP on cancer cell lines that have been explored, there are traditional models like Bayesian Matrix Factorization [11], Random Forests [12], and Support Vector Machines [13] as well as more recent Neural Networks and Deep Learning techniques [14–16]. These methods, including Convolutional NNs, Graph NNs and Multimodal DL architectures are capable of handling complex, high-dimensional data and have been used to model drugs, genetic features, and to integrate multiomics data [14, 17–23].

Aside from the sophistication of the models employed for this task, in this paper, we investigate two crucial but often overlooked aspects of DRP, which are (1) the validation approach used to evaluate the prediction performance and (2) the prediction label used to train the models.

Regarding the first point, the validation approaches used to evaluate the performance of DRP methods, which involve the combination of how the train and test sets are created (Splitting Strategy) and the specific approach we proposed in this paper for aggregating the prediction results to compute the prediction scores (Aggregation Strategy). Our in silico experiments show how the subtleties hidden in the validation of DRP methods can lead to completely misleading performance scores, depending on the characteristics of the datasets used, and how the proper combination of the right Splitting and Aggregation strategies can overcome these issues, by evaluating the model exactly on the kind of task it is designed to solve. Since on the most important DRP datasets, including GDSC, CCLE, and CTRP, the type of drug tested is the main driver of the variability in IC50 values, simply learning which drugs are generally strong or weak allows any DRP predictor to fool any global evaluation metrics computed on the entire test dataset. Regrettably, this misleading evaluation setting remains prevalent across all current DRP methods. Our study highlights that despite the seemingly impressive global performance metrics, DRP models may still lack any real capability to accurately predict the outcomes for novel (previously unseen) cell lines or drugs. A recent study [24] has proposed z-scored drug response values to address drug-driven variability, but this approach cannot handle new compounds. Our differentiated Aggregation Strategies offer a more flexible solution for evaluating model generalization across both drugs and cell lines. Another recent study has addressed a similar issue in the context of gene essentiality predictions, showing how biases in data can mislead model evaluation and emphasizing the need for bias-aware validation frameworks [25].

The second point concerns most DRP methods since they focus on the regression of IC50 values to measure drug response [8]. The IC50 value corresponds to the drug concentration necessary to inhibit the viability of 50% of the cells, which is obtained by dose-response curves from cell viability experiments. These experiments are performed within specific concentration ranges, which are chosen based on the existing knowledge on the target drug [26]. In the paper, we show that the final IC50 values are highly dependent on these ranges, and in particular on the Maximum Concentration (MC) tested. This scenario leads to DRP models that struggle to generalize to new drugs, and they cannot even guess the expected concentration ranges. We therefore endorse the use of the Area Under the Dose-Response Curve (AUDRC) as an alternative target label, an approach explored in several studies [27–30].

These puzzling results show that, unless stricter evaluation criteria are put in place, specifically targeted for the type of generalization ability that we want to test (on unseen drugs or cell lines), the model is able to bypass the conventional evaluation metrics, similarly to what has been shown in other contexts, such as image recognition, cancer driver prediction and Reinforcement Learning [31–33]. This leads to a situation of specification gaming[33] (also known as reward hacking [34]), in which the model satisfies the evaluation criteria without achieving the desired outcome.

The high performance scores are reached instead by exploiting an unfortunate combination of a peculiar data structure and evaluation metrics that are too generic to be robust to data loopholes [32, 33].

To prevent the model from obtaining high performance scores by just gaming the validation specifications of DRP, in this paper, we propose different ways to aggregate the predictions (Aggregation Strategies) in order to compute more meaningful evaluation metrics. Each of them is specifically meant to measure a particular

generalization ability (towards drugs or cell lines). We also show how the choice of Aggregation Strategy critically depends on how the data are split into training and test subsets (Split Strategy). The novel validation protocols for DRP can be specifically targeted to the type of generalization expected from the model under scrutiny (prediction of novel drugs or novel cell lines).

Background: splitting strategies in DRP

In cheminformatics literature has been highlighted that the choice of splitting in a dataset of compounds, which may have internal structural relationships, can significantly alter both the effectiveness of modeling and the reliability of validation [35, 36]. More generally, observed performances of ML methods depend on the way in which samples are allocated between training and test sets within the validation procedure of choice. In particular, the choice of splitting strategy primarily depends on the assumptions that the researchers want to test (e.g., generalization to structurally novel compounds), and secondarily on the availability of sufficient data to enable the training of functional models [8, 32, 37, 38].

To investigate which aspects of the cancer DRP are more challenging and what level of performance can be realistically expected in real-life settings, in this paper, we analyze the DRP validation problem. Here we start by listing the increasingly stringent strategies available to define the training and test sets (see Fig. 1C for an overview of the splitting strategies):

1 *Random splits:* This approach is also called Mixed-Set in [8, 39], and it is generally the least challenging, leading to the highest observed performance scores. In this scenario, a randomly selected subset of drug-cell line pairs is excluded from the training set and used as the test set. This train-test Splitting Strategy quantifies how accurate a model is in filling the gaps in a drug-cell lines matrix containing some unobserved values. Practically, this would correspond to filling a non-exhaustive screening on a panel of otherwise known cell lines and drugs. In this scenario, the model is not evaluated in terms of its ability to generalize to cell lines or drugs for which we completely lack drug response measurements.
2 *Unseen cell lines:* In this case, the train and test splits are made by ensuring that the cell lines in the training set are not present in the test. The test set is constructed by randomly selecting a subset of cell lines and all of their IC50 values from the entire dataset. To achieve high performance scores in this validation, the models need to be able to generalize to unseen cell lines. With respect to the Random Splits, this therefore increases the difficulty of the prediction task.
3 *Unseen drugs:* The train and test splits are made to ensure that the drugs that appear in the test set are not present in the training set. To perform well in this setting, the model must be able to generalize well to completely unseen drugs.
4 *Unseen cell line-drug pairs:* This is the most stringent validation setting. In this case, the training and test splits are built to ensure that each of the cell lines and drugs present in the test set are both absent from the training set. This setting therefore evaluates the ability of the model to generalize at the same time to unseen drugs and cell lines, which should be the ultimate goal of the cancer drug sensitivity prediction field. However, until now, generalization in this setting has been nearly impossible, and as such, it is infrequently utilized in evaluations [9].

These different Splitting Strategies are generally a standard in DRP literature [8, 9, 40], however, there is noticeable variability in their actual use [8]. For example it is common practice to pay special attention to issues like this, with scaffold splitting being one such technique used to avoid bias from specific molecular structures [41, 42]. For instance, the DREAM Challenge on drug sensitivity prediction emphasized the importance of rigorous splitting by evaluating models on a blind test set of entirely unseen cell lines [10]. In the context of QSAR modeling, the choice of train-test split impacts performance estimates, model internal optimization, and generalizability [43]. Figure 3A visually illustrates the generalization difficulty that the DRP models need to overcome, when they are assessed using these strategies.

Results

Beyond global metrics: aggregation strategies for robust model evaluation

DRP methods need careful evaluation to assess their real-world applicability. In this section, we introduce a novel framework that examines prediction performance from multiple perspectives, moving beyond the traditional "global" evaluation approach commonly used in the field.

Once the predictions are computed with one of the train-test splitting strategies described above, some performance metrics must be computed on the predictions to evaluate their level of agreement with the ground truth used as labels.

Typically, in ML model validation, metrics are applied "globally", namely they are computed across all predicted values in the test set, and this is indeed the approach usually adopted by DRP methods. But what is precisely lost, in terms of analyzing what the predictions mean,

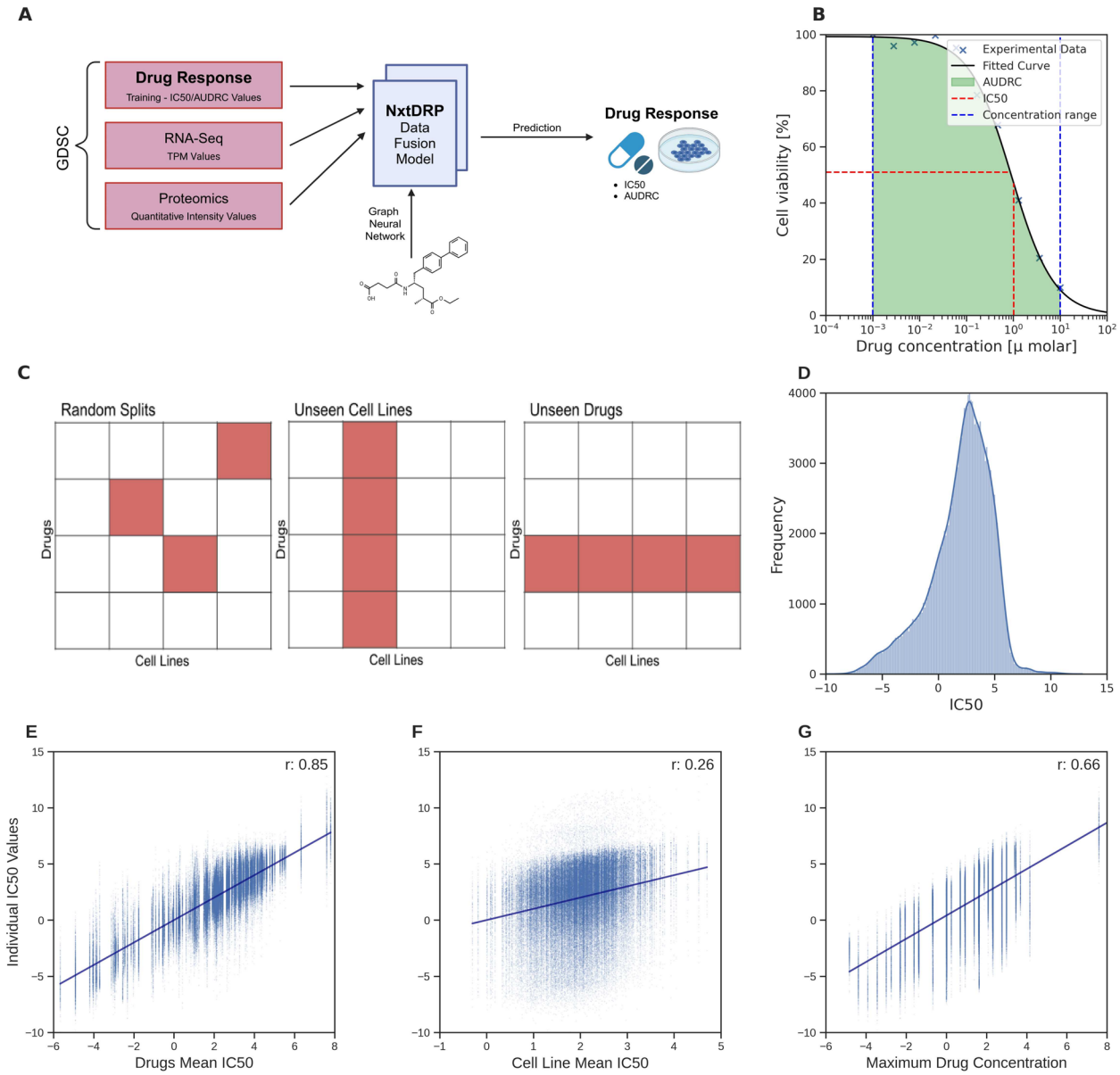

Fig. 1 **A** Schematic representation of the complete NxtDRP model pipeline. **B** Representative dose-response curve illustrating the definition of IC50 and the Area Under the Dose-Response Curve (AUDRC). **C** Data splitting strategies employed for model evaluation: Random Splits, Unseen Cell Lines, and Unseen Drugs (red cells indicate the test set). **D** Distribution of IC50 values derived from the GDSC dataset. **E** Distribution of individual IC50 values normalized by their corresponding drug's mean IC50. **F** Distribution of individual IC50 values normalized by their corresponding cell line's mean IC50. **G** Individual IC50 values relative to the maximum drug concentration tested in each cell viability experiment. All IC50 values are expressed on the natural logarithm (ln) scale

by using just this "global" averaging? Here we look at the predictions from two other angles as well, computing the performance scores also across drugs and cell lines separately, obtaining three prediction Aggregation Strategies:

1. *Global*: This is the most common approach. The performance metrics are calculated over the entire test set (i.e., overall correlations).
2. *Fixed-drug*: In this Aggregation Strategy, the performance metrics are computed individually for each drug, and the resulting Fixed-Drug performance scores are then averaged over the entire test-set (across all the drugs). This Aggregation Strategy enables us to analyze the prediction quality for individual drugs independently, thereby highlighting the mod-

el's ability to discern between the potentially unique behaviors of different cell lines.

3. *Fixed-cell line*: In the third Aggregation Strategy, the metrics are calculated individually for each cell line in the test set. The resulting Fixed-Cell Line performance scores are then averaged over the entire test set. This approach allows us to analyze the performance on each cell line independently from the drugs used, emphasizing the model's ability to distinguish between different drugs when a fixed cell line is considered.

In the scientific literature, Global aggregation is prevalent. However relying solely on this Aggregation Strategy may result in unreliable (i.e. inflated) prediction performance, depending on the dataset characteristics. The role that these Aggregation Strategies play in terms of what they precisely measure is also tightly intertwined with the Splitting Strategy being used. For a mathematical definition of these strategies, see Suppl. Section S2. Figure 3B visually illustrates how reliable these aggregation strategies are when it comes to evaluate the generalization ability of DRP models on drugs and cell lines. Throughout the rest of the paper, we delve into this phenomenon in detail.

A novel non-linear multi-omics drug response prediction method

To showcase the relevance of various combinations of train-test splits and prediction aggregation strategies for the validation of DRP methods on cancer cell lines, we developed a novel multiomics prediction method, called NxtDRP. It is based on the NXTfusion non-linear data fusion library proposed in [44], and allows total flexibility in testing the relevance of different types of omics data, making it particularly suitable for our analyses. The NXTfusion library generalizes the classical Matrix Factorization approach to perform inference over heterogeneous sources of information represented as Entity-Relation (ER) graphs. Each Entity in the ER graph corresponds to a class of objects (i.e. Cell Lines and Drugs), and it is internally represented by a set of latent variables that are optimized to accurately predict the target labels (i.e. IC50 values). Each -omic data matrix is added to the ER graph as a Relation connecting two Entities. For example, the Proteomics and RNA-Seq data are respectively represented as relations between the Cell Lines and the Proteins and between the Cell Lines and the Genes entities (see Fig. 6 and Methods "Entity-relation graph inference for DRP" for more details).

To benchmark NxtDRP, we used the GDSC dataset, which is the most commonly used database for this task [45], due to its size and the availability of various omics to characterize cell lines.

As mentioned above, this dataset exhibits greater variability between drugs than between cell lines. This peculiarity is shared by the other major DRP datasets, such as CCLE and CTRP (see Fig. 4).

We extracted 948 cell lines and 223 drugs, totaling 172,114 drug response values in the form of IC50 (see Fig. 1D). We followed the same pre-processing steps proposed in [16] (see Methods Datasets for more details). Figure 1A, shows an overview of NxtDRP and the multi-omics data it integrates, such as RNA-Seq expression and Proteomics profiles from GDSC.

We compared the performance of NxtDRP with different ER graphs, to evaluate how the integration of different multi-omics data contribute to the prediction. The most basic ER graph involves only the Main Task (MT), namely the matrix containing the IC50 values corresponding to the Cell Line-Drug pairs available in GDSC (Fig. 6B). We then added to this minimal ER graph the available omics, one at a time, including Proteomics (PR) and RNA-Seq (EX) matrices as additional relations (see Fig. 6C–E), in an attempt to better characterize the cell lines. Throughout the paper, we refer to these relations respectively as MT, PR and EX, to indicate which are included in each model.

To make it more suitable for DRP, we also extended the original NXTfusion library [44] by adding a Graph NN to incorporate the molecular details of target drugs. See the Methods "An entity-relation data fusion model to predict the cancer cell lines drug sensitivity" and "Drug structure representation for machine learning" for more details.

We benchmarked NxtDRP with two previously published DRP methods: tCNN [16] and GraphDRP [15]. To ensure a fair comparison, we adopted the same iterated train-test design they used, as described in the tCNN paper [16]. This means that we trained and tested NxtDRP 40 times, each time with a new selection of the train and test sets (90–10% proportion), according to the chosen train-test Splitting Strategy.

To measure the performance of the methods, we adopted the Root Mean Squared Error (RMSE) and Pearson correlation metrics (r), averaging them over the train-test iterations (see Suppl. Section S1 for the details).

A detailed analysis of performance in the random splits validation

The Random Splits validation strategy measures how good a predictor is at filling the gaps in untested drugs-cell lines pairs. Practically, this corresponds to filling an incomplete screening on a panel of otherwise known cell lines and drugs. When validating a model following a

Random Splits strategy, developers and users must know that the model is not actually tested on its ability to generalize to cell lines or drugs for which we completely lack drug response measurements.

The results in Fig. 2A are obtained with the Random Splits validation strategy. The Global performance column indicates that NxtDRP, tCNN [16] and GraphDRP[15] are able to achieve high scores, with a Pearson correlation up to $r = 0.93$ over the entire dataset. The main reason why this setting does not present a substantial challenge is that the same drugs and cell lines can be present in both the training and test set (just not the same cell lines and drug pairs).

In Fig. 2A there is no substantial difference in performance, across all Aggregation Strategies, between the model that incorporates multi-omics data and the one that does not (see NxtDRP$_{MT}$ and NxtDRP$_{MT+PR+EX}$ models). Previous research has highlighted the uncertain

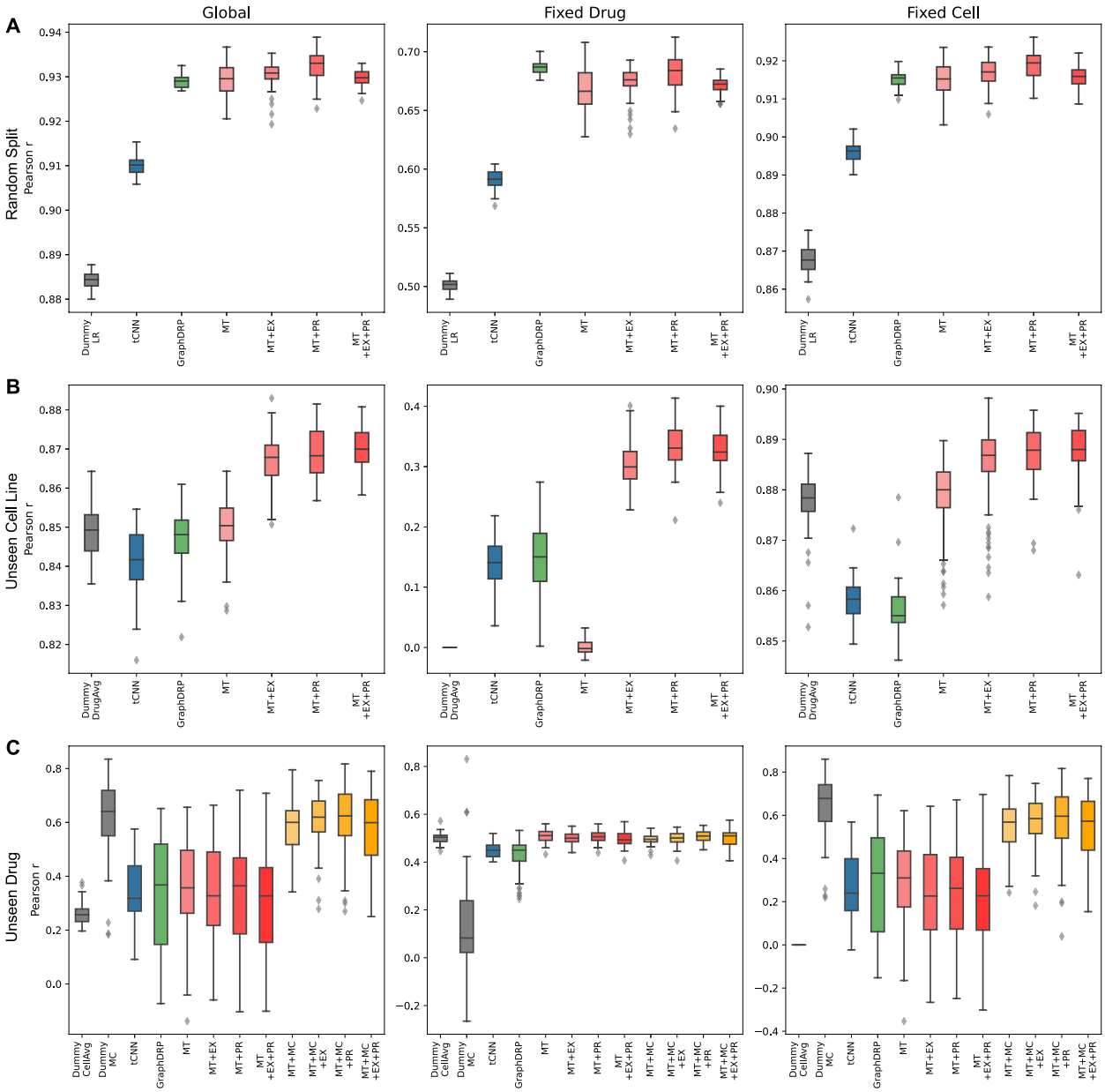

Fig. 2 Boxplots showing the distribution of the prediction performances, measured by Pearson's *r* values, for the tCNN [16], GraphDRP [15], and the NxtDRP models on the GDSC dataset across three Aggregation Strategies (columns) and three Splitting Strategies (rows A,B, and C). The NxtDRP variants MT, PR, and EX denote the omics data utilized: none, Proteomics, and Transcriptomics, respectively (for further details, refer to Fig. 6B–E)

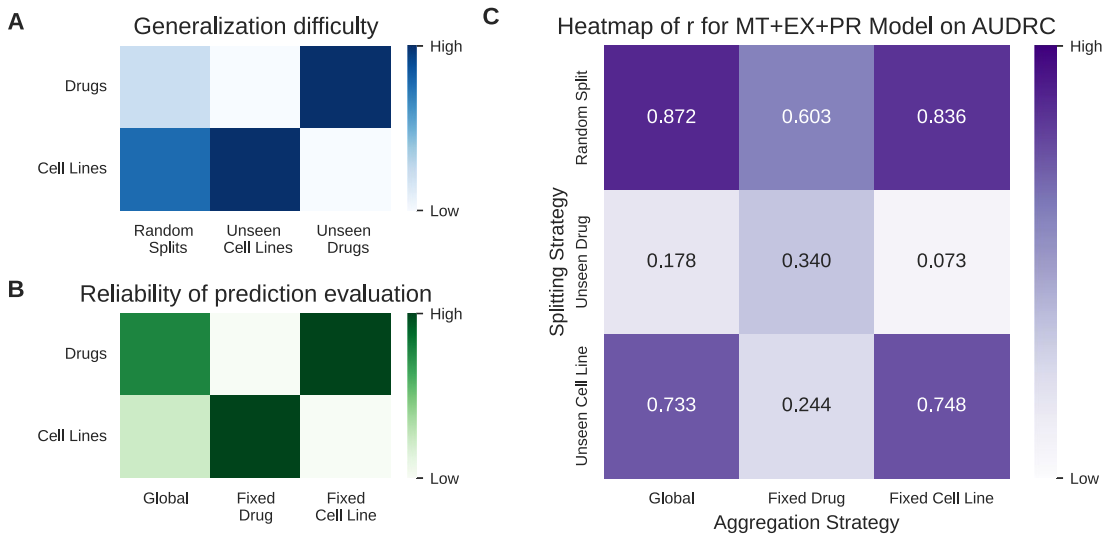

Fig. 3 **A** Heatmap qualitatively showing the difficulty of generalizing to either Drugs or Cell Lines with respect to different Splitting Strategies. **B** Heatmap qualitatively showing how various Aggregation Strategies underscore the level of generalization achieved for different entities, such as drugs or cell lines. **C** Heatmap quantitatively summarizing the Pearson *r* correlations of the predictions of $NxtDRP_{MT+PR+EX}$ on AUDRC in relation to both Splitting Strategy and Aggregation Strategy

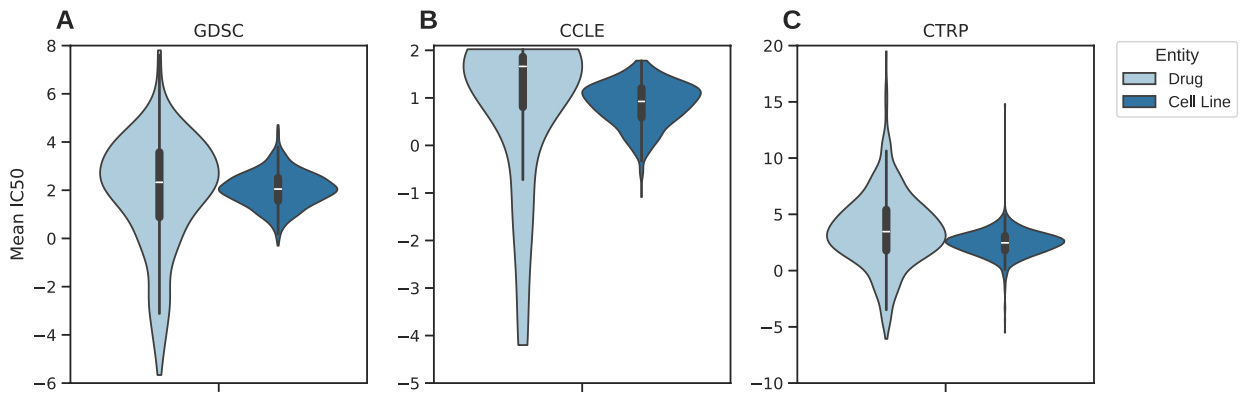

Fig. 4 Plot showing the distribution of drugs and cell lines mean IC50 values on GDSC (**A**), CCLE (**B**), CTRP (**C**) datasets. All the IC50 values are expressed in logarithmic scale

relevance of omics data in this context [46, 47]. This can be explained by the fact that, in terms of information content, all the IC50 values that the model observes for a given cell line and drug are sufficient to saturate the information it can learn, and therefore the additional omics data have no added value. In the Random Splits strategy, the contextual information provided by multi-omics measurements does not aid in characterizing the involved entities (Drugs and Cell Lines) more than what the drug response data alone ($NxtDRP_{MT}$) already achieves.

If we focus on the other columns of Fig. 2A, we see that the Fixed-Drug performances are lower than Global performances. One probable cause of this behavior is that the model is trained *globally*, and, by minimizing its loss function, it therefore tends to capture primarily the main source of variance present in the data. As shown in Fig. 4A, in GDSC, this variance is indeed primarily driven by drugs (this is true even for CCLE and CTRP, see Fig. 4B,C). This is a possible reason for the poorer modeling of the cell line variability highlighted by the Fixed-Drug performance. At the same time, we can observe that Global performance and Fixed-Cell Line performance are comparable. This is explained by the fact that Global performance reflects variance both within drugs and cell lines; since the

primary source of variance in the data is *globally* due to the drugs anyway, the two measurements are indeed extremely similar. Instead, Fixed-Cell Line aggregation considers only variance within drugs.

The importance of prediction aggregation strategies in the unseen cell lines validation

What happens instead if we populate the test sets used during validation only with samples coming from cell lines that are *not* present in the corresponding training sets? In real-life scenarios, this would evaluate the ability of our models to use the multi-omics data to generalize to unseen cell lines, without having observed any drug response values on them.

Why are global performances of unseen cell lines high even without multi-omics data?

The most noticeable aspect in Fig. 2B is that the Global performance of NxtDRP$_{MT}$, which does not include any multi-omics information able to contextualize the cell lines, is already able to reach state-of-the-art performance ($r = 0.85$).

To study this perplexing behaviour we looked at the distribution of the drugs IC50 values over the cell lines. As mentioned in Sect. "A detailed analysis of performance in the Random Splits validation", in Fig. 4A, we see that the variability among drugs ($\sigma^2 = 2.38$) is greater than that among cell lines ($\sigma^2 = 0.72$). This means that, on GDSC, the predictors can already accurately model much of the variability in the data by just observing the drug responses available in the training set, regardless of the cell lines involved. This means that, if we only consider the Global aggregation of the predictions, like most of the state-of-the-art predictors do [14–16, 48–51], we remain completely unaware of the model ability to truly generalize over cell lines, which was supposed to be the goal of this Splitting Strategy.

The fact that the high Global performance in Fig. 2B are misleading becomes evident when we compare the Global and Fixed-Drug (which directly assesses the model's capacity to generalize to unseen cell lines) performance of NxtDRP$_{MT}$. The Pearson correlation sharply decreases from $r = 0.85$ to $r = 0$, as is supposed to be, since NxtDRP$_{MT}$ with the Unseen Cell Lines Splitting Strategy could not learn anything about the cell lines drug response in the test set. This drop dramatically highlights how the Global aggregation of predictions can lead to inflated and therefore meaningless performance scores: we see a Global Pearson correlation of 0.85 despite the actual model inability to discriminate unseen cell lines (NxtDRP$_{MT}$ Fixed-Drug $r = 0$).

Another noticeable aspect is that when additional relations are added to characterize the cell lines (NxtDRP$_{MT+EX+PR}$) the model reaches a Global $r = 0.87$, outperforming tCNN and GraphDRP respectively by 3.44% and 2.71% (see Suppl. Table S2). On the other hand in the Fixed-Drug settings of Fig. 2B we see that adding additional relations to the model (NxtDRP$_{MT+EX+PR}$) becomes highly relevant for the performance, providing an increase in Pearson correlation, from 0.00 to 0.33, while the multi-omics contribution was only a 2% r increase in the Global aggregation.

A dummy model highlights the bias induced by Global Aggregation

To make the analysis of this behavior clearer, we introduce the DummyDrugAvg model, which predicts the IC50 for a drug d as the average IC50 of d as observed on the training set (see Methods "Dummy models" for more details). DummyDrugAvg extremizes the situation already observed with NxtDRP$_{MT}$, since by construction it cannot model or recognize different cell lines. Nonetheless, it achieves a Global Pearson correlation of 0.85 (see Fig. 2B), which is in line with all the real DRP methods benchmarked in the same settings. This value corresponds indeed to the correlation we measured in Fig. 1E. Similarly to NxtDRP$_{MT}$, DummyDrugAvg suddenly drops to 0.0 correlation when the Aggregation Strategy switches to Fixed-Drug since it cannot predict how the same drug performs on different cell lines, but just the Global drugs trends over the entire test-set.

On the other hand, DummyDrugAvg and all the NxtDRP variants reach a high correlation ($r \simeq 0.88$) in the Fixed-Cell Line aggregation performance, since the global ranking of the drugs effectiveness is mostly conserved also within each individual cell line (see Suppl. Figure S3). This means that generally strong drugs will be stronger than generally weaker drugs, with an average correlation across cell lines of $r = 0.885$.

Prediction and evaluation challenges in the unseen drugs validation

Analysing the poor generalization in the unseen drugs validation

Mirroring the previous Splitting Strategy on Unseen Cell Lines, here we predict the IC50 values for drugs that are entirely absent from the training set. To perform well in the Unseen Drugs Splitting Strategy, a DRP model should be able to generalize to drugs never seen before, relying only on biologically relevant information such as its structure. In the case of NxtDRP, the drug structure is extracted from PubChem [52] and fed to the model via a Graph NN (see Methods "Drug structure representation for machine learning" and "Entity-relation graph inference for DRP" for more details).

The accurate prediction of IC50 values for the Unseen Drugs Splitting Strategy is currently a challenge, as shown by the lack of reliable in silico methods for this type of validation [14, 16]. This problem is challenging because the chemical space covered by the dataset is both sparse and highly diverse, making it difficult for deep learning models to generalize, as they would require a much more comprehensive representation of the chemical space to make accurate predictions; structurally similar drugs may be locally well-modeled, but predicting sensitivity of unseen compounds remains highly unreliable.

The difficulty of this task is indeed confirmed by the generally low prediction performance shown in Fig. 2C. Moreover, the variability is very high, making it impossible to directly compare models. This is mainly due to the high variability of the IC50 values, caused by the differing biochemical mechanisms of the drugs in the dataset and their relative effective concentration range, leading to heterogeneous subsets in each split and affecting prediction accuracy.

The difficulty of the problem posed by this Splitting Strategy is also evident from the other prediction Aggregation Strategies. In the Fixed-Cell Lines aggregation, which evaluates the models' ability to predict the dynamics between different drugs acting on the same cell line, the performance of NxtDRP$_{MT}$ has a 22% drop ($r = 0.28$) with respect to Global performance.

In Fig. 2C, the Global performance is relatively similar to the Fixed-Cell Line performance. As already mentioned in Sect. "A detailed analysis of performance in the Random Splits validation", this is because the former, due to higher variability among drugs (see Fig. 4), tends to measure mostly how well the model approximates the dynamics of drug-related IC50 values, which is the same that the Fixed-Cell Line aggregation exclusively does.

To better understand these results, we compared it with two baseline dummy predictors. The first is DummyCellAvg, which simply computes the mean of the IC50 values associated to each cell line c on the training set, and uses this mean as prediction value for any drug applied on c in the test set. This method achieves a Global correlation of $r = 0.26$. This value is substantially lower compared to the DummyDrugAvg performance on the unseen cell lines, which was $r = 0.85$ (see Suppl. Table S2). This is due to the fact that they both just exploit the variance in the dataset (respectively among cell lines and drugs), and the second is substantially higher than the first, as shown from the comparison of Fig. 1E and F. These plots show that the Global correlation achieved by DummyCellAvg corresponds exactly to the correlation in Fig. 1F, showcasing what a Global Aggregation Strategy truly measures when it comes to quantifying the prediction performance of DRP methods on datasets presenting a substantial structure in the data.

In the previous Sect. "The importance of prediction Aggregation Strategies in the unseen cell lines validation", we showed that to properly assess the models' ability to predict previously Unseen Cell Lines, the Fixed-Drug Aggregation Strategy is the most meaningful. Analogously, here we see that to reliably evaluate the predictions on Unseen Drugs, we should use the Fixed-Cell Lines Aggregation Strategy, since it explicitly measures the ability of the model to characterise the effect of different drugs in the same condition (the cell line). Indeed we see from Fig. 2C that the DummyCellAvg baseline achieves zero correlation in this Aggregation Strategy, while NxtDRP$_{MT}$ with just the drug information fed through the Graph NN reaches $r = 0.283$.

Conversely, the Fixed-Drug performance of DummyCellAvg and NxtDRP$_{MT}$ in Fig. 2C is higher to both Global and Fixed-Cell Line performance ($r \simeq 0.50$). The reason for this is that the IC50 values of cell lines are similarly distributed among different drugs ($r = 0.51$, see Suppl. Figure S3). Analogously to the evidence in Sect. A dummy model highlights the bias induced by Global Aggregation, this confirms that certain cell lines tend to be more sensitive to drugs, while others are generally less sensitive, however this difference is less pronounced than the analogous effect among drugs.

The addition of omics data in the Unseen Drugs splitting provides little to no noticeable improvement in terms of performance. As already highlighted in Sect. "A detailed analysis of performance in the Random Splits validation" given a drug-cell line pair, if the training set contains enough drug response values for that cell line, the model will be able to characterize it without relying on omics data.

The maximum drug concentration alone outperforms DRP methods

The drastic drop in performance from Fig. 2A and C confirms the general observation [8, 15, 16, 28] that DRP predictors struggle to generalize to unseen drugs [14, 16], with high errors on the IC50 predictions.

Part of this issue might be related to the fact that different drugs are tested at different concentration ranges, resulting in IC50 values that are expressed within these ranges. These concentration ranges are typically selected *a priori* based on existing in vitro and clinical data associated with each drug [26].

This behavior is also visible in Fig. 1G, where we show that the Maximum Concentration (MC) at which the drugs are tested strongly correlates ($r = 0.66$) with the IC50.

To gauge how this might affect the models predictions, we ran an additional test in which the model is provided with the MC at which each drug has been experimentally tested in order to infer the IC50.

We first integrated the MC as a feature in NxtDRP (see box NxtDRP$_{MT+PR+MC}$ in Fig. 2C), by concatenating it with the drug representation generated by the Graph NN. This value alone increases the Global Pearson correlation by 67% ($r = 0.60$).

To further investigate the role of MC in relation to the other inputs fed to NxtDRP$_{MT+MC}$, we designed an additional baseline model, called DummyMC. For each drug *d*, DummyMC predicts its IC50 by simply outputting the MC tested for *d*. This trivial approach outperforms any other DRP predictor in the Global predictions aggregation ($r = 0.61$) without even requiring any type of modeling of the cell line or the drug.

If we look at the Fixed-Drug performance in Fig. 2C, we see that the DummyMC Pearson correlation drops to almost zero, since it always outputs the same IC50 value to each drug *d*, regardless of the cell line involved. Conversely, NxtDRP$_{MT+PR+MC}$ on Fixed-Drug performance achieves $r = 0.56$, because the cell lines are characterized in this case.

DummyMC simply reflects the assumptions on which the *in vitro* experiments that DRP models wish to emulate, and it could therefore considered a *baseline* for benchmarking DRP performance. Unfortunately, as we show in Fig. 2C, current DRP methods are unable to surpass this baseline.

Discussion

In this paper we highlight two issues that, to the best of our knowledge, have not been adequately addressed in the DRP field. We believe that the development of increasingly sophisticated DRP models cannot ignore the need for a validation protocol that truly assesses their generalization capability. The critical points we have addressed concern the use of IC50 as the target label and the standard approach adopted for the validation of DRP models.

Beyond IC50: the area under the dose-response curve as an alternative prediction label

We showed how IC50 values strongly correlate with MC (see Fig. 1G), and how this becomes problematic since a baseline represented by the mere MC value (DummyMC) results in higher performance than advanced DRP models.

What makes this correlation even more problematic is the fact that it is subtly influenced by the assumptions underlying IC50, which is based on the premise that exists a concentration inhibiting 50% of cells [27]. In reality, 62% of the IC50 values in GDSC are higher than the actual MC tested for the target drug, meaning that most of the IC50 values have never been actually experimentally observed, and it is interpolated from the dose-response sigmoid instead. This is even more striking on the CCLE dataset, since when the IC50 exceeded the MC tested (the 55% of the cases), the IC50 was approximated by using the MC value instead (see Suppl. Figure S2).

An option in this sense is to use the Area Under the Dose-Response Curve (AUDRC) (see Fig. 1B), that is provided in drug response datasets. AUDRC, also known as AUC in chemistry literature, of 1 means that the drug is not able to inhibit the viability of cancer cell, even at the MC tested, while an AUDRC value close to zero indicates that the drug is extremely effective against cancer cells, even at the lowest concentrations tested. The AUDRC, commonly reported in its normalized form, effectively decouples the concentration ranges tested for each drug from their effectiveness as anticancer agents. This is because these ranges define the boundaries of the calculated area, as shown in Fig. 1B.

AUDRC has already been proposed as the preferred target label for DRP methods [8, 27, 30]. In this paper, we add further reasons and evidence to endorse the use of AUDRC as a more reliable target label. From a real-world clinical perspective, AUDRC represents a prediction at least as useful as IC50, as it can be used similarly to prioritize drug-cell line pairs for in vitro testing. At the same time, because AUDRC is normalized with respect to the concentration ranges specific to each drug, it also enables improved comparability among different drugs by abstracting away from the assumptions (e.g., cytotoxicity thresholds) made for that particular drug. It is robust in reflecting the drug's effectiveness across the concentration ranges at which the drug is active and providing a more accurate measure of the drug-cell interaction compared to the simple IC50 value [53].

The long-overlooked role of the aggregation strategies

Casting aside the discourse related to the target label of choice, our study shows that the validation of DRP methods is hindered by the fact that the only Aggregation Strategy currently used in literature (Global) biases the performance towards the entity showing greater variance, which happens to be the drug entity in GDSC, CCLE and CTRP (see Figs. 4 and 5). This leads to models that reach completely misleading high Global performance, even though they are actually unable to generalize on these entities.

The differentiation of the Aggregation Strategies that we propose addresses this issue. In addition to the Global aggregation, Fixed-Drug and Fixed-Cell Line aggregations must be used to assess the true performance of DRP

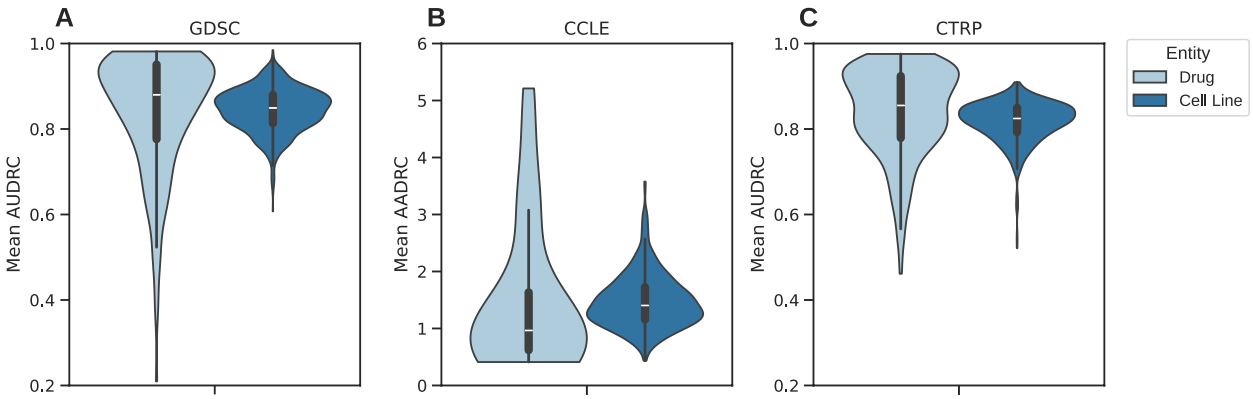

Fig. 5 Plot showing the distribution of drugs and cell lines mean AUDRC values on GDSC (**A**), CCLE (**B**), CTRP (**C**) datasets. All the IC50 values are expressed in logarithmic scale. For CCLE the area above the dose-response curve is reported

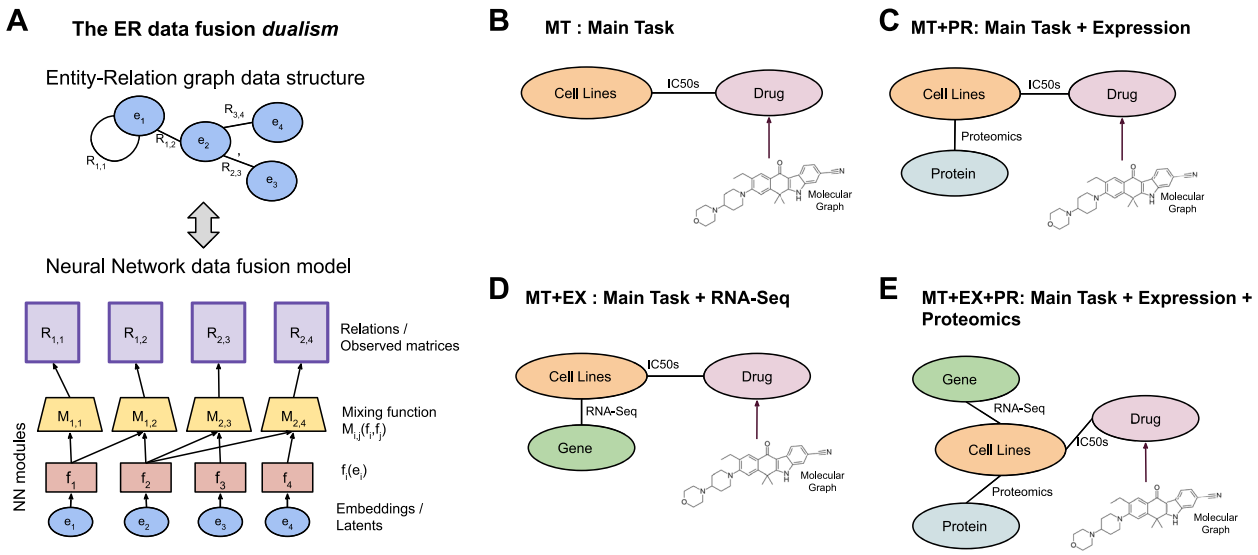

Fig. 6 **A** Representation of the dualism of the abstract representation of the data in an Entity-Relation graph and the NXTFusion schematization of the non-linear inference over that ER graph. **B** ER Graph representing only the Main Task that is the DRP on IC50 values. **C** ER Graph representing the Main Task with the addition of RNA-Seq expression relation. **D** ER Graph representing the Main Task with the addition of Proteomics relation. **E** ER graph representing the Main Task with both Proteomics and RNA-Seq relations

methods. They assess the ability of DRP models to generalize to unseen cell lines (Fixed Drug aggregation) and to unseen drugs (Fixed Cell Line aggregation), as we showed in Sects. "The importance of prediction aggregation strategies in the unseen cell lines validation" and "Prediction and evaluation challenges in the unseen drugs validation".

Aggregation Strategies are also relevant to evaluate the predictions in the Random Splits setting, as they provide an overview of the degree of generalization achieved independently for Drugs and Cell Lines.

Similar issues have been highlighted in [24], where the authors propose to predict a z-scored AUDRC value to address the variability problem. However, this approach becomes inapplicable in the Unseen Drugs setting, since mean and variance cannot be computed for never seen before drugs. Our proposed differentiation of Aggregation Strategies offers a more flexible solution, allowing us to avoid being misled by dataset variance while specifically analyzing the model's predictive capabilities with respect to either drugs or cell lines.

A more robust validation protocol for DRP methods

When using AUDRC, there is still a difference in variability between drugs and cell lines, albeit to a lesser extent (see Fig. 5A). For this reason, it is also necessary to couple it with the proper prediction Aggregation Strategies,

to avoid the issues highlighted in Sects. "The importance of prediction aggregation strategies in the unseen cell lines validation" and "Prediction and evaluation challenges in the unseen drugs validation", since also the AUDRC Global performance could be biased by the fact that it mostly describes the model's ability to explain variance among drugs rather than among cell lines.

To provide a complete solution to these issues, here we propose a novel validation protocol for DRP methods, which is free from the issues we highlighted so far:

1. Use AUDRC as prediction label, instead of the IC50.
2. Aggregate the predictions using the Fixed-Drug and Fixed-Cell Lines strategies (see Sect. "Beyond global metrics: Aggregation Strategies for robust model evaluation"). In particular, the Fixed-Drug aggregation describes the model ability to discriminate between cell lines, and vice versa, the Fixed-Cell Line performance indicates how well the model correctly distinguishes between different drugs.
3. In the case of a validation with Unseen Drugs train-test splits, the most relevant Aggregation Strategy will be Fixed-Cell Line. Conversely, to evaluate the generalization ability in the Unseen Cell Lines splits, the main prediction Aggregation Strategy should be Fixed-Drug.

To showcase the novel validation procedure we propose, we tested NxtDRP on AUDRC as target label in these settings. The results are shown in Suppl. Section S4.

Conclusion

In this study, we demonstrated that conventional evaluation practices for DRP methods result in misleading performance scores due to inherent dataset biases, particularly the disproportionate variability among drugs compared to cell lines in widely used datasets like GDSC, CCLE, and CTRP. By relying solely on global aggregation metrics, existing DRP models appear to achieve high performance by exploiting drug-specific trends rather than generalizing to unseen cell lines or drugs. This "specification gaming" phenomenon masks critical limitations in model generalizability.

To address this, we introduced three Aggregation Strategies—Global, Fixed-Drug, and Fixed-Cell Line—that allow us to evaluate performance independently across the entities involved in the prediction, namely cell lines and drugs. These strategies revealed stark discrepancies: models achieving high global scores often failed to generalize to novel cell lines (Fixed-Drug aggregation) or drugs (Fixed-Cell Line aggregation). For instance, with Unseen Cell Line splits, dummy models leveraging drug averages achieved near-state-of-the-art global performance, but collapsed to zero correlation when evaluated under task-specific aggregation, underscoring the necessity of tailored validation protocols, depending on the assumptions that need to be validated.

Furthermore, we highlighted fundamental limitations of IC50 as a prediction label. Its strong dependency on drug concentration ranges, particularly the Maximum Concentration (MC), enables trivial baselines (e.g. DummyMC) to outperform sophisticated DRP models. We advocate for replacing IC50 with AUDRC, which provides a more robust and interpretable measure of drug efficacy across concentrations.

By integrating AUDRC with our proposed validation framework-combining task-aligned Splitting and Aggregation Strategies-we establish a rigorous protocol to ensure DRP models are evaluated on their true ability to generalize, advancing their reliability for preclinical drug discovery and precision oncology applications.

Methods

Datasets

In this study, we used GDSC v. 6.0 [5] as our main drug-response dataset. It contains 1,074 cell lines and 224,510 IC50 measurements of 265 drugs. For each cell line, the following omics are available: whole genome sequencing, transcriptomics, proteomics, copy number variation and methylation data. Drugs are identified by their PubChem IDs, facilitating the retrieval of their chemical structures.

GDSC dataset contains drug dose-response data measurements. These measurements are obtained using fluorescence signal intensities, testing 9 different concentrations per drug with 2-fold dilution series [5]. The dose-response curve is then fitted by a nonlinear mixed effect model on a sigmoid curve. The IC50 values we use as prediction labels are therefore the results of the interpolation of the dose-response curve with the sigmoid curve (see Fig. 5A). These IC50 values are therefore subject to noise due, for example, to high experimental variability that may cause a poor fitting, which is estimated by an RMSE value. We excluded IC50 values with an RMSE > 0.3 from the analysis.

To facilitate a fair comparison with existing approaches, we followed the pre-processing steps outlined in [16]. First, only compounds with a PubChem ID were considered, resulting in a final dataset of 223 drugs and 948 cell lines, with a total of 172,114 drug response IC50 values. Next, we rescaled these values to the [0,1] range using the formula $y = \frac{1}{1+y^{-0.1}}$ [16].

Additionally, we retrieved quantitative proteomics intensity values from GDSC and scaled them to the [0,1] range. This data matrix comprises 4,538,041 data points across 874 cell lines and 8,457 proteins. We also obtained

RNA-Seq data containing 20,080,264 transcript per million (TPM) measurements, which were scaled to the [0,1] range. This data matrix involves 912 cell lines and 36,447 genes. To reduce dimensionality and retain the most informative features, we selected the top 1000 most variable genes, resulting in a total of 369,072 Cell Line-Gene pairs.

For further analyses, we employed two additional datasets—CCLE and CTRP—to investigate the variance among drugs and cell lines, as shown in Figs. 4 and 5. CCLE dataset contains 11,670 drug-cell line pairs, comprising 504 cell lines and 24 drugs, while the CTRP dataset contains 482,528 pairs, including 887 cell lines and 544 drugs.

Drug structure representation for machine learning

Each drug is uniquely identified by a PubChem ID. From PubChem we retrieved their Simplified Molecular Input Line Entry System (SMILES) representation [52]. However, this representation cannot be used as input in a DL model as is. We therefore transformed them into graphs that represent the molecular structures of the drugs.

Each drug is represented by a molecular graph in which the nodes corresponds to an atom and is described by the following set of features: atomic symbol (one-hot encoding), atomic number, atomic degree, atomic formal charge, atom in a ring, atom radical electrons, atom hybridization state, and atom aromaticity. Each edge in this graph correspond to a chemical bond between atoms.

Each molecular graph is used as input to a Graph NN, that consists of 4 graph attention layers [54], followed by a final global sum pooling step. The final pooling is necessary to provide a final latent representation of each drug that is independent of the actual number of atoms of each drug.

An entity-relation data fusion model to predict the cancer cell lines drug sensitivity

Predicting the response of cancer cell lines to anticancer drugs requires integrating heterogeneous sources of information, such as the omics available for each cell line and the molecular information related to each drug. To build a model able to do that, we started from our data fusion framework NXTFusion [44], which allows us to describe these heterogeneous data sources within an Entity-Relation (ER) graph (see Main Fig. 6A), which can be intuitively thought as a relational database on which it is possible to perform inference.

An ER graph consists of two key elements: a set of entities denoted as E, which represent classes of objects, and a set of relations R that specify how pairs of entities are interconnected. Each entity has a specific cardinality that corresponds to the number of instances present in the available data. For example, the Cell Line entity on GDSC data has 948 instances.

From a modeling perspective, the entities are represented by a set of trainable latent variables e_i. Each observed data matrix R_{ij} (for example proteomics data) corresponds to a relation between a pair of entities (i, j).

The training is performed globally on the ERG by finding the embedding sets e_i and e_j that minimize the reconstruction error on each observed matrix (relation) R_{ij}. The loss function to be minimized for each R_{ij} is therefore $L_{ij}(R_{ij}, M_{ij}(f_i(e_i), f_j e_j))$ where L_{ij} is the relation-specific loss function, f_i is an entity specific differentiable function (such as a feed-forward NN), and $M_{ij}(.,.)$ a relation-specific mixing function (such as bilinear layer) [44]. The loss function L_{ij} can match the type of data contained in each relation R_{ij}, depending on whether it is a classification, regression, or multi-class prediction task (see Fig. 6A).

Since the ER graph might contain an arbitrary number of relations between entities, the global objective function to be optimized is the following

$$\underset{W}{\operatorname{argmin}} \sum_{R_{ij} \in R} \omega_{ij} L_{ij}(R_{ij}, M_{ij}(f_i(e_i), f_j e_j))$$

where W is the set of trainable parameters of each NN module in f and M functions. ω is a scale factor meant to ensure that different relations have a comparable weight in the global loss (see [44] for more details).

Essentially, NXTfusion achieves data fusion through learning several tasks concurrently. The auxiliary tasks help the inference of the main task by providing additional context, which acts as an informed regularization [44], therefore helping the generalization on the main task by ensuring the convergence to more informative latent representations of the entities involved in the ER graph [37].

The prediction is performed by considering, for the relation of interest R_{ij}, all pairs of embeddings corresponding to entities e_i and e_j for which the interaction label has to be predicted. Clearly, the value of the relation for those entities pairs must not have been included in the training set. At inference time, the outcome is computed as $Y = M_{ij}(f_i(e_i), f_j(e_j))$. That is, the predictions rely on the embeddings as well as the weights of f and M that were learned during the training phase. In case of an unseen entity instance (cell line or drug) for the main task R_{ij} the embedding for that instance is learned through the reconstruction of auxiliary task. If no auxiliary task is available, the model is not able to learn a meaningful embedding (this is the reason of random performance of

NxtDRP_{MT} on Unseen Cell Lines split and Fixed-Drug aggregation).

Entity-relation graph inference for DRP

In the context of our DRP task, the ER graph comprises four entities: cell lines, drugs, proteins, and genes. The corresponding relations include the Cell Line-Drug relation (built on drug response values from GDSC and serving as the target for predictions during testing), the Cell Line-Protein relation (based on quantitative proteomics data), and the Cell Line-Gene relation (derived from RNA-Seq data).

Each entity-specific module f_i consists of a linear transformation, followed by a normalization (LayerNorm) and a hyperbolic tangent (Tanh) activation. For all relations, the mixing function M_{ij} is implemented as a bilinear layer, succeeded by LayerNorm, a Dropout with a probability of 0.1, a Tanh activation, and a final linear transformation.

The last crucial piece in the picture is the contextualization of the drugs. We therefore decided not to use a latent representation (embedding) to represent them, but to extend the concept of side information (i.e. features in the Matrix Factorization jargon [44, 55]) by directly connecting a Graph NN to the Cell Line-Drug relation (Fig. 6B–E). This Graph NN operates on the molecular graph representation of each drug (as described in Sect. "Drug structure representation for machine learning"), dynamically encoding structural and functional properties into relational embeddings. By design, this eliminates the need for drug's entity embeddings, allowing molecular features to propagate through the ER graph during both training and inference.

During the training phase the model jointly learns all the observed relations in the ER graph, except for the Cell Line-Drug interaction pairs reserved for the test set; these held-out pairs are masked during optimization. The predictions for the Cell Line-Drug pairs of the test set rely on two key components: (1) the learned embeddings for cell lines (reflecting their multi-omics and drug response profiles), and (2) the Graph NN-generated representations of drug molecules. The final predictions are computed by the mixing function layer M that combines these elements.

To assess the impact of integrating multiple omics, we compared different ER graphs with different combinations of relations (see Fig. 6B–E). The simplest graph, denoted as the Main Task (MT) graph, includes only the Cell Line-Drug relation (i.e. the IC50 matrix from GDSC, which is the target for predictions). We then progressively augmented this graph with additional relations: one for proteomics (PR) and another for RNA-Seq (EX) data, to enrich the cell line characterization (see Fig. 6B–E).

The method along with the code for its implementation is available at github.com/codicef/NxtDRP to reproduce the results.

Dummy models

The dummy models in this paper are categorized into two groups. The first group is based on simple transformations of the IC50 values from the dataset, while the second utilizes external information, such as the maximum concentration (MC) at which the drug was tested.

For the first category, these models employ IC50 data based on the specific validation performed. For Unseen Cell Lines, the average IC50 values for each drug (DummyDrugAvg) are used, excluding data from cell lines in the test set. Similarly, for Unseen Drugs, the average IC50 values for cell lines (DummyCellAvg) are utilized, excluding data from drugs in the test set. In the Random Splits strategy, a basic linear regression model (DummyLR) trained on concatenated one-hot-encoded identifiers for drugs and cell lines.

As for the dummy model based on MC, it simply uses the MC value at which each drug was tested (DummyMC). This information is known a priori from the screening experiments that the DRP models aim to replicate.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00972-y.

Supplementary material 1.

Acknowledgements

FC thanks E. Salvadori for her support and constructive discussions. FC is also grateful to T. Sanavia and G. Birolo for their help and scientific insights during the project. Special thanks to S. Borgato and S. Zanella for assistance with the graphics. DR is grateful to A. L. Mascagni.

Author contributions

FC, DR and PF wrote the manuscript. FC and DR implemented the method and performed the experiments. FC, DR, PF, CP and CR analyzed and discussed the results. DR, PF and YM supervised the project. All authors read and approved the final manuscript.

Funding

European Union's Horizon 2020 project GenoMed4All (Grant Agreement ID:101017549). DR was funded by an FWO senior post-doctoral fellowship (grant number 12Y5623N) and he is now funded by a CNRS Chaire de Professeur Junior grant (PROJET N° ANR-23-CPJ1-0171-01). CP was supported by a AIRC fellowship for Italy.

Availability of data and materials

The code and models developed in this study are available at https://github.com/codicef/NxtDRP. The preprocessed data are available https://github.com/codicef/NxtDRP/releases/tag/dataset. The GDSC dataset analyzed during the current study is available at https://www.cancerrxgene.org/, CCLE dataset is available at https://sites.broadinstitute.org/ccle/ and CTRP dataset is available at https://portals.broadinstitute.org/ctrp/.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 28 October 2024 Accepted: 13 February 2025
Published online: 14 March 2025

References

1. Schwartzberg L, Kim ES, Liu D, Schrag D (2017) Precision oncology: who, how, what, when, and when not? Am Soc Clin Oncol Educ Book 37:160–169
2. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A (2020) Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 4(1):19
3. Shoemaker RH (2006) The nci60 human tumour cell line anticancer drug screen. Nat Rev Cancer 6(10):813–823
4. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D et al (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483(7391):603–607
5. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Richard Thompson I et al (2012) Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res 41(D1):D955–D961
6. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, Ebright RY, Stewart ML, Ito D, Wang S et al (2013) An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 154(5):1151–1161
7. Firoozbakht F, Yousefi B, Schwikowski B (2021) An overview of machine learning methods for monotherapy drug response prediction. Brief Bioinform 23(1):bbab408, 10
8. Partin A, Brettin TS, Zhu Y, Narykov O, Clyde A, Overbeek J, Stevens RL (2023) Deep learning methods for drug response prediction in cancer: predominant and emerging trends. Front Med 10:1086097
9. Baptista D, Ferreira PG, Rocha M (2021) Deep learning for drug response prediction in cancer. Brief Bioinform 22(1):360–379
10. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Ammad-Ud-Din M, Hintsanen P, Khan SA et al (2014) A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 32(12):1202–1212
11. Ammad-Ud-Din M, Khan SA, Malani D, Murumägi A, Kallioniemi O, Aittokallio T, Kaski S (2016) Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics 32(17):i455–i463
12. Riddick G, Song H, Ahn S, Walling J, Borges-Rivera D, Zhang W, Fine HA (2011) Predicting in vitro drug sensitivity using random forests. Bioinformatics 27(2):220–224
13. Majumdar A, Liu Y, Yaoqin L, Shaofeng W, Cheng L (2021) kesvr: An ensemble model for drug response prediction in precision medicine using cancer cell lines gene expression. Genes 12(6):844
14. Liu Q, Zhiqiang H, Jiang R, Zhou M (2020) DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 36(Supplement 2):i911–i918
15. Nguyen T, Nguyen GTT, Nguyen T, Le D-H (2021) Graph convolutional networks for drug response prediction. IEEE/ACM Trans Comput Biol Bioinf 19(1):146–154
16. Liu P, Li H, Li S, Leung K-S (2019) Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinform 20(1):1–14
17. Nguyen GTT, Vu HD, Le D-H (2021) Integrating molecular graph data of drugs and multiple-omic data of cell lines for drug response prediction. IEEE/ACM Trans Comput Biol Bioinf 19(2):710–717
18. Chu T, Nguyen TT, Hai BD, Nguyen QH, Nguyen T (2022) Graph transformer for drug response prediction. IEEE/ACM Trans Comput Biol Bioinf 20(2):1065–1072
19. Jiang L, Jiang C, Xinyu Yu, Rao F, Jin S, Liu X (2022) Deeptta: a transformer-based model for predicting cancer drug response. Brief Bioinform 23(3):bbac100
20. Sharma A, Lysenko A, Boroevich KA, Tsunoda T (2023) Deepinsight-3d architecture for anti-cancer drug response prediction with deep-learning on multi-omics. Sci Rep 13(1):2483
21. Li Y, Guo Z, Gao X, Wang G (2023) Mmcl-cdr: enhancing cancer drug response prediction with multi-omics and morphology images contrastive representation learning. Bioinformatics 39(12):btad734
22. Taj F, Stein LD (2024) Mmdrp: drug response prediction and biomarker discovery using multi-modal deep learning. Bioinform Adv 4(1):vbae010
23. Manica M, Oskooei A, Born J, Subramanian V, Sáez-Rodríguez J, Martinez MR (2019) Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. Mol Pharm 16(12):4797–4806
24. Ovchinnikova K, Born J, Chouvardas P, Rapsomaniki M, Julio MK (2024) Overcoming limitations in current measures of drug response may enable AI-driven precision oncology. NPJ Precis Oncol 8(1):95
25. Bence S, Imre G, Valér K, László M, Milán S, Szalay, Kristóf Z (2023) The effect benchmark suite: measuring cancer sensitivity prediction performance-without the bias. bioRxiv, pp 2023–10
26. Vis DJ, Bombardelli L, Lightfoot H, Iorio F, Garnett MJ, Wessels LFA (2016) Multilevel models improve precision and speed of ic50 estimates. Pharmacogenomics 17(7):691–700
27. Sock JI, Chaibub NE, Justin G, Friend Stephen H, Margolin Adam A (2014) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. pp 63–74
28. Pozdeyev N, Yoo M, Mackie R, Schweppe RE, Tan AC, Haugen BR (2016) Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. Oncotarget 7(32):51619
29. Fallahi-Sichani M, Honarnejad S, Heiser LM, Gray JW, Sorger PK (2013) Metrics other than potency reveal systematic variation in responses to cancer drugs. Nat Chem Biol 9(11):708–714
30. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, Hon C, Mammoliti A, Nair SK, Mer AS, Ester M, Haibe-Kains B (2021) How much can deep learning improve prediction of the responses to drugs in cancer cell lines? Brief Bioinform 22(6):bbab294
31. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R (2019) Unmasking clever hans predictors and assessing what machines really learn. Nat Commun 10(1):1096
32. Raimondi D, Passemiers A, Fariselli P, Moreau Y (2021) Current cancer driver variant predictors learn to recognize driver genes instead of functional variants. BMC Biol 19:1–12
33. Victoria K, Jonathan U, Vladimir M, Matthew R, Tom E, Ramana K, Zac K, Jan L, Shane L (2020) Specification gaming: the flip side of ai ingenuity, 4 2020. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/
34. Skalse J, Howe N, Krasheninnikov D, Krueger D (2022) Defining and characterizing reward gaming. Adv Neural Inf Process Syst 35:9460–9471
35. Hawkins DM, Kraker J (2010) Deterministic fallacies and model validation. J Chemom 24(3–4):188–193
36. Hawkins DM, Basak SC, Mills D (2003) Assessing model fit by cross-validation. J Chem Inf Comput Sci 43(2):579–586
37. Mazzone E, Moreau Y, Fariselli P, Raimondi D (2023) Nonlinear data fusion over entity-relation graphs for drug-target interaction prediction. Bioinformatics 39(6):btad348, 05
38. Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW et al (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 36(5):513–523
39. Branson N, Cutillas PR, Bessant C (2023) Comparison of multiple modalities for drug response prediction with learning curves using neural networks and XGBoost. Bioinform Adv 4(1):vbad190
40. Chen Y, Zhang L (2021) How much can deep learning improve prediction of the responses to drugs in cancer cell lines? Brief Bioinform 23(1):bbab378
41. Robinson MC, Glen RC, Lee AA (2020) Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. J Comput Aided Mol Des 34(7):717–730
42. Jun X (2002) A new approach to finding natural chemical structure classes. J Med Chem 45(24):5311–5320

43. Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, Tropsha A (2012) Does rational selection of training and test sets improve the outcome of qsar modeling? J Chem Inf Model 52(10):2570–2578
44. Raimondi D, Simm J, Arany A, Moreau Y (2021) A novel method for data fusion over entity-relation graphs and its application to protein-protein interaction prediction. Bioinformatics 37(16):2275–2281
45. Singh DP, Kaushik B (2023) A systematic literature review for the prediction of anticancer drug response using various machine-learning and deep-learning techniques. Chem Biol Drug Des 101(1):175–194
46. Cortes-Ciriano I, Van Westen GJP, Bouvier G, Nilges M, Overington JP, Bender A, Malliavin TE (2016) Improved large-scale prediction of growth inhibition patterns using the nci60 cancer cell line panel. Bioinformatics 32(1):85–95
47. Wang C, Lye X, Kaalia R, Kumar P, Rajapakse JC (2021) Deep learning and multi-omics approach to predict drug responses in cancer. BMC Bioinform 22(Suppl 10):632
48. Li M, Wang Y, Zheng R, Shi X, Li Y, Fang-Xiang W, Wang J (2019) Deepdsc: a deep learning method to predict drug sensitivity of cancer cell lines. IEEE/ACM Trans Comput Biol Bioinf 18(2):575–582
49. Chu T, Nguyen TT, Hai BD, Nguyen QH, Nguyen T (2022) Graph transformer for drug response prediction. IEEE/ACM Trans Comput Biol Bioinf 20(2):1065–1072
50. Zuo Z, Wang P, Chen X, Tian L, Ge H, Qian D (2021) Swnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. BMC Bioinform 22:1–16
51. Tang Y-C, Gottlieb A (2021) Explainable drug sensitivity prediction through cancer pathway enrichment. Sci Rep 11(1):3128
52. Kim S, Thiessen PA, Bolton EE, Chen J, Gang F, Gindulyte A, Han L, He J, He S, Shoemaker BA et al (2016) Pubchem substance and compound databases. Nucleic Acids Res 44(D1):D1202–D1213
53. Yadav B, Pemovska T, Szwajda A, Kulesskiy E, Kontro M, Karjalainen R, Majumder MM, Malani D, Murumägi A, Knowles J et al (2014) Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. Sci Rep 4(1):1–10
54. Shaked B, Uri A, Eran Y (2021) How attentive are graph attention networks? arXiv preprint arXiv:2105.14491
55. Jaak S, Adam A, Pooya Z, Tom H, Wegner Jörg K, Vladimir C, Hugo C, Yves M (2015) Macau: scalable bayesian multi-relational factorization with side information using mcmc. arXiv preprint arXiv:1509.04610

Publisher's Note