Drug response prediction: A critical systematic review of current datasets and methods

Nguyen Khoa Tran, Gunnar W. Klau

Suggested Citation:
Tran, N. K., & Klau, G. (2025). Drug response prediction: A critical systematic review of current datasets and methods. Pattern Recognition Letters, 199, 21–26. https://doi.org/10.1016/j.patrec.2025.10.016

Pattern Recognition Letters 199 (2026) 21–26

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Drug response prediction: A critical systematic review of current datasets and methods

Nguyen Khoa Tran *, Gunnar W. Klau

Heinrich Heine University Düsseldorf, Universitätsstraße 1, Düsseldorf, 40225, North Rhine-Westphalia, Germany
Center for Digital Medicine, Düsseldorf, Germany

ARTICLE INFO

Editor: Doulaye Dembele

Dataset link: https://github.com/AlBi-HHU/Drug_Response_Prediction

Keywords:
Drug response prediction
Multi-output regression
Molecular graphs
Multi-omics
Multilayer perceptron
Graph neural network

ABSTRACT

Predicting drug response is a critical task in personalized medicine. Several recent studies have reported promising improvements in predictive performance with deep learning models trained on molecular characterizations of cell lines and drugs. However, our baseline tests suggest that little to no meaningful biological or chemical information is being learned from multi-omics data in the publicly available large-scale datasets GDSC and DepMap Public or molecular graphs, respectively. In our experiments, even gene expression data, commonly regarded as highly predictive, failed to deliver satisfactory drug response predictions. This raises the possibility that drug response measures or patterns observed in multi-omics data may not arise from underlying biological mechanisms. To investigate this, we identified and examined inconsistencies within and across the GDSC2 and DepMap Public 24Q2 datasets. We found that IC_{50} and AUC values of replicated experiments in GDSC2 had an average Pearson correlation coefficient of only 0.563 ± 0.230 and 0.468 ± 0.358, respectively. Additionally, somatic mutations shared between cell lines in the two datasets showed a Pearson correlation coefficient of only 0.180. Even in cases where TGSA, the current best-performing method to our knowledge, exceeded baseline performance, it still did not surpass a simple baseline multi-output multilayer perceptron (MMLP). Moreover, MMLP is not only more easily adaptable to new datasets but also significantly faster, making it a viable baseline for comparisons. In conclusion, our findings suggest that current cell-line and drug data are insufficient for existing modeling approaches to effectively uncover the biological and chemical mechanisms underlying drug response. Therefore, improving data quality or focusing on different data types is crucial before proposing novel methods.

1. Introduction

Precision medicine aims to tailor cancer therapies to individual patients, yet predicting a patient's response to a drug based on their biological characteristics remains challenging. This difficulty stems from the complex nature of cancer and the limited availability of clinical data.

To address these challenges, large-scale initiatives such as the Genomics of Drug Sensitivity in Cancer (GDSC) [1] and the Dependency Map (DepMap) project [2] have emerged. These projects involve testing numerous anti-cancer drugs across diverse cancer cell lines using high-throughput screening technologies. Additionally, these datasets include detailed gene profiles from omics data such as somatic mutations (MUT), copy number variations (CNV), and gene expression (EXP).

Drug response prediction models aim to find a mapping $f(x) \approx y$ from molecular characterizations of cell lines and drugs x to drug response values y. These response values are commonly measured as the half-maximal inhibitory concentration (IC_{50}) or the area under the curve (AUC). Machine learning models designed to predict drug response often include dense neural networks (DNNs) [3], convolutional neural networks (CNNs) [4–6], autoencoders [7], and attention mechanisms [8]. Other approaches involve random forests [9]. To further improve predictive accuracy, additional biological data like pathway information or protein-protein interaction (PPI) networks can be integrated through fully connected neural networks (FCNNs) [10, 11] or graph neural networks (GNNs) [12]. Drug features, including molecular fingerprints, drug targets, SMILES, and molecular graphs, have also been incorporated using DNNs [10], CNNs [4,5,8], and GNNs [6,12].

Eckhart et al. [13] have shown that among gene profiles, EXP generally offers better predictive power than MUT or CNV, though

* Corresponding author at: Heinrich Heine University Düsseldorf, Universitätsstraße 1, Düsseldorf, 40225, North Rhine-Westphalia, Germany.
E-mail addresses: nguyen.tran@hhu.de (N.K. Tran), gunnar.klau@hhu.de (G.W. Klau).

https://doi.org/10.1016/j.patrec.2025.10.016
Received 2 April 2025; Received in revised form 18 August 2025; Accepted 25 October 2025
Available online 30 October 2025

it is less robust to biological and technical variability [14]. Furthermore, dimensionality reduction can significantly improve the performance of both simple and complex models [13]. When incorporating additional biological data, pathway information has shown limited benefit [11], while PPI networks have been reported to improve predictive accuracy [12]. In terms of chemical representation, drug target data has been shown to be inferior to fingerprints [11], while fingerprints [10] and SMILES [8] have been shown to be inferior to molecular graphs [15,16]. Regarding model types, black-box models like multilayer perceptrons (MLPs) tend to outperform more interpretable models like random forests [11] or attention mechanisms [15].

In compliance with all these findings, twin graph neural networks with similarity augmentation (TGSA) [12] employs black-box GNNs for both molecular graphs and PPI networks and has emerged as the top-performing model in recent evaluations [15,16]. TGSA is reported to perform well in the *leave pairs out* (LPO) scenario, where drug responses for missing cell line-drug pairs are predicted when all cell lines and drugs are present in both training and test data. However, its performance is noted to decline in both blind test scenarios—a trend observed across multiple models [6,11,12,15]. These scenarios are the *leave cell lines out* (LCO) scenario, where cell lines are present in test data only, and the *leave drugs out* (LDO) scenario, where drugs are present in test data only. Both scenarios reflect real-world clinical settings, where predicting responses for previously unseen cell lines or drugs is essential.

To explore the reasons for this performance drop, we designed several baseline models for comparison with TGSA. Surprisingly, in the LDO scenario, TGSA did not exceed baseline performance, and in the LCO scenario, it did not surpass a simple multi-output multilayer perceptron (MMLP). Strikingly, even in the LPO scenario, TGSA failed to surpass MMLP. To investigate further, we conducted an ablation study on the gene profiles (MUT, CNV, and EXP) as well as the molecular graphs to assess their individual contributions to predictive performance. Additionally, we analyzed the GDSC2 dataset (version 8.5) and compared it with the DepMap Public dataset (version 24Q2). Our analysis reveals that within and across the datasets, MUT, CNV, and drug reponse values exhibit only low to moderate concordance. We conclude that with current data, existing models are unable to effectively uncover meaningful biological and chemical mechanisms underlying drug response, emphasizing the critical importance of improving data quality or generating and focusing on alternative data types. Existing models can then be adjusted to them and reevaluated.

In order to facilitate the verification of future datasets and models in LPO/LCO/LDO scenarios, we provide a user-friendly benchmark environment using Snakemake [17], which includes all baseline and ablation tests and MMLP. Since MMLP closely resembles a standard MLP, it can be easily adapted to various kinds of data and offers a low runtime, and thus it can serve as an additional baseline model that future models need to surpass.

2. Methodology

This section provides a description of the datasets, the two models TGSA and MMLP, and the experimental setup used in this study.

2.1. Datasets

The input data for drug response prediction models consists of at least two matrices: a feature matrix $x \in \mathbb{R}^{n\times d}$ and a target matrix $y \in \mathbb{R}^{n\times p}$, where n is the number of cancer cell lines, d is the number of gene features, and p is the number of cancer drugs. Regarding the GDSC2 dataset, we derived four feature matrices and one target matrix from Cell Model Passports [18]. We excluded the proteomics matrix due to 38.6% missing values, and we chose not to impute with 0 since missing entries correspond to high q-values, signaling unreliable protein abundance measurements. The three remaining feature matrices are the somatic mutation matrix $x_{\text{MUT}} \in \{0,1\}^{934\times 23189}$, the copy number variation matrix $x_{\text{CNV}} \in \{-2,-1,0,1,2\}^{934\times 20669}$, and the gene expression matrix $x_{\text{EXP}} \in \mathbb{R}_{+}^{934\times 37005}$. The target matrix is the $\log_{10}(\text{IC}_{50})$ matrix $y \in \mathbb{R}^{934\times 184}$, with 9.7% of the cell line-drug pairs missing. We filtered out all drugs without PubChem ID to enable running TGSA on the SMILES retrieved from PubChem [19] using PubChemPy (version 1.0.4) [20]. Also, since TGSA's PPI network requires feature selection for feasible runtime, we adopted the list compiled by Zhu et al. [12] containing cancer-related genes according to COSMIC [21], resulting in reduced input matrices: $x_{\text{MUT}}^{\text{COSMIC}} \in \{0,1\}^{934\times 658}$, $x_{\text{CNV}}^{\text{COSMIC}} \in \{-2,-1,0,1,2\}^{934\times 658}$, and $x_{\text{EXP}}^{\text{COSMIC}} \in \mathbb{R}_{+}^{934\times 658}$. Note that although the input matrices are of different types (binary, discrete, and continuous), both TGSA and MMLP are able process them simultaneously, owing to the flexible trainability of neural network edge weights. Further details on the data and preprocessing are provided in the Supplementary Material (Sections 1, 2, and 3).

2.2. Twin graph neural networks with similarity augmentation (TGSA)

Zhu et al. [12] designed TGSA to integrate both fine-grained (gene-level and atom-level) and coarse-grained (sample-level) information through two main steps: twin graph neural networks for drug response prediction (TGDRP) and a similarity augmentation module.

The TGDRP step involves two components. The first is a PPI network built from the feature matrices $x_{\text{MUT}}^{\text{COSMIC}}$, $x_{\text{CNV}}^{\text{COSMIC}}$, and $x_{\text{EXP}}^{\text{COSMIC}}$, using the detailed protein links file (version 11.0) from STRING [22], containing 400 GB gene interaction data. Each node in the PPI network represents a gene with the node features being taken from the feature matrices, and edges are drawn between nodes if the detailed protein links file contains an interaction between two genes above a predefined threshold (default threshold is 0.95). The second component is a molecular graph of the form $G = (V, E)$ built from the SMILES of a drug using RDKit (version 2022.09.1) [23]. Predictions for cell line-drug pairs are generated by training one GNN on the PPI network and another on the molecular graph, combining their results using a fully connected neural network (FCNN).

The second optional step, the similarity augmentation module, requires one cell line graph and one drug graph that are built as follows: Each cell line or drug is represented as a node, where each cell line is connected to the five most similar cell lines according to the Pearson correlation coefficient (PCC) of their EXP values, whereas each drug is connected to the five most similar drugs according to the Jaccard similarity of their fingerprints. To compute the final drug response prediction for each cell line-drug pair, the model parameters of the earlier trained TGDRP model are used for initializing the cell line graph and the drug graph. After that, one GNN is trained on the cell line graph and another GNN on the drug graph, with their results subsequently combined in an FCNN.

2.3. Multi-output multilayer perceptron (MMLP)

Because training a separate MLP for every single drug is time-consuming [13] and disregards potential correlations among targets [24], our MMLP predicts a fixed number of targets simultaneously for given cell lines. However, it cannot be used in the LDO scenario since the same fixed number of targets is needed for training and testing. For its input, MMLP uses one feature matrix along with the target matrix. Multiple feature matrices with the same cell lines can be concatenated into a single matrix, e.g., x_{MUT}, x_{CNV}, and x_{EXP} can be concatenated into $x \in \mathbb{R}_{+}^{934\times 80863}$.

MMLP follows the architecture of a standard MLP, with an input layer of d nodes (one per feature), a configurable number of hidden layers with h nodes (h is a hyperparameter), and an output layer of p nodes (one per drug). Two simple modifications are introduced: (1) a sigmoid layer between the input and first hidden layer, assigning a weight between 0 and 1 to each feature, interpretable as a feature

importance score [25]; (2) an imputation mask that sets gradients to zero during backpropagation for missing values. Modification (1) allows for more granularity, especially for binary features, and also supports feature selection if desired, while modification (2) ensures that imputed values (occurring due to missing values in the target matrix and/or due to the LPO scenario) do not affect parameter updates.

After performing hyperparameter tuning using grid search, we selected the best hyperparameters for the GDSC2 dataset, which were used in all experiments and are as follows: number of hidden layers = 1, hidden size $h = 2048$, batch size = 8, activation function = LeakyReLU, dropout ratio = 0.5, optimization algorithm = Adam, learning rate = 0.0001, weight decay = 0, loss function = mean squared error (MSE), maximum epochs = 300, with early stopping after 10 epochs of no improvement.

2.4. Test configuration

To ensure reliable results, we applied k-fold cross-validation (CV) with $k = 5$. For each split, the training set was used to optimize model parameters, selecting the best performance on the validation set before measuring performance on the test set. Hyperparameters for TGSA were taken from Zhu et al. [12], while hyperparameters for MMLP were preselected, see Section 2.3.

We employed three types of CV: record-wise (LPO), subject-wise (LCO), and target-wise (LDO). LPO CV tests the ability to predict missing cell line-drug pairs when all cell lines and drugs are present in the training data, LCO CV evaluates predictions for unseen cell lines, and LDO CV assesses predictions for unseen drugs. See Fig. 1 for an illustration of the three data splitting options.

While Shen et al. [15] and Menden et al. [26] also train and test models on a single drug at a time, we omitted this approach for TGSA because training on the drug features of a single drug would be redundant since there is no difference between the drug features among all datapoints.

Note that due to missing data (9.7%) and the fact that the number of cell lines/drugs is not always divisible by 5, the CV splits do not always contain exactly 20% of the data.

2.5. Baseline tests

We conducted the following baseline tests:

1. Mean predictor: For each unseen cell line-drug pair (i, j), the drug response prediction is computed as the mean of:
   - LPO: the average response of cell line i to all seen drugs and the average response of all seen cell lines to drug j.
   - LCO: the average response of all seen cell lines to drug j.
   - LDO: the average response of cell line i to all seen drugs.

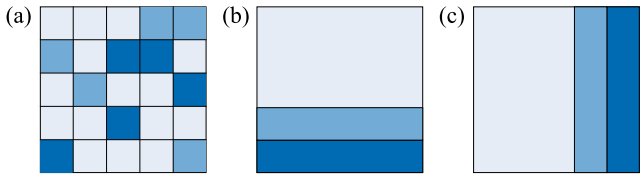

Fig. 1. An illustration of splitting target matrix y in three different ways: (a) record-wise (LPO), (b) subject-wise (LCO), or (c) target-wise (LDO). In this example, the training set (light blue) consists of 3/5 of y, while the validation set (medium blue) and test set (dark blue) consist of 1/5 each. Rows are cell lines and columns are drugs.

2. Biological information test: For models directly training on a feature matrix (e.g., MMLP), we replaced the feature matrix with an identity matrix of size $n \times n$ to eliminate any coincidental patterns or similarities between cell lines. This test is not applicable for LCO.
3. Chemical information test: For models training on molecular graphs (e.g., TGSA), we replaced the molecular graph of each drug i with a single node containing $p = 184$ node attributes. All attributes were set to 0 except for the ith attribute, ensuring the elimination of any coincidental patterns or similarities between drugs. This test is not applicable for LDO.
4. Shuffled feature matrix: We shuffled the entries of the feature matrix to check if any biological information is learned. As the shuffled data may introduce coincidental patterns that either favor or disadvantage the models, we preferred the second baseline test when testing for biological information and only ran this test for LCO.

3. Results

We implemented a benchmarking workflow for drug response prediction models using Python 3, PyTorch [27], and the Snakemake workflow management system [17]. The root mean square error (RMSE) was selected as the evaluation metric, as the commonly used coefficient of determination (R^2) is unsuitable for nonlinear models [28].

Figs. 2, 3, and 4 present the results for LPO, LCO, and LDO CV, respectively. For each model, the five RMSEs (one per test set) are displayed in a boxplot. Runtime for the TGSA models varied between 10 h and 3 days, while MMLP models took between 1 and 4 h. In the following, we distinguish between TGDRP (the first step of TGSA) and TGSA (both steps of TGSA). MMLP_{ALL} refers to MMLP trained on all features, i.e., the concatenation of x_{MUT}, x_{CNV}, and x_{EXP}. For models trained on individual feature matrices, MMLP_i represents MMLP trained on x_i, where $i \in \{\text{MUT}, \text{CNV}, \text{EXP}\}$. The superscript $^{\text{COSMIC}}$ is added when training on COSMIC genes only, e.g., $\text{MMLP}_{\text{ALL}}^{\text{COSMIC}}$. The second baseline test, the biological information test, is denoted as MMLP_{id}, while the third baseline tests, the chemical information tests, with TGDRP and TGSA are written as TGDRP_{id} and TGSA_{id}, respectively. The fourth baseline test, the shuffled feature matrix, was conducted on MMLP_{EXP} only and is referred to as $\text{MMLP}_{\text{EXP}}^{\text{shuffled}}$.

In the LPO scenario (Fig. 2), the best-performing model is $\text{MMLP}_{\text{ALL}}^{\text{COSMIC}}$, outperforming TGDRP and TGSA when trained on the same COSMIC genes. Selecting only COSMIC genes seems to generally have a positive effect on the performance of MMLP in the LPO scenario. Even though $\text{MMLP}_{\text{ALL}}^{\text{COSMIC}}$ and other models demonstrate low RMSEs, none significantly outperform the baseline tests. TGDRP and TGSA achieve mean RMSEs of 0.936 and 0.937, respectively, slightly worse than the baseline TGSA_{id} (mean RMSE: 0.935), suggesting no meaningful chemical information was learned. $\text{MMLP}_{\text{ALL}}^{\text{COSMIC}}$ with a mean RMSE of 0.930 shows a marginal improvement of 4% ($= 1 - \frac{\text{mean RMSE of model}}{\text{mean RMSE of baseline test}}$) over MMLP_{id} (mean RMSE: 0.970), indicating limited biological information is learned. TGDRP and TGSA perform worse than $\text{MMLP}_{\text{ALL}}^{\text{COSMIC}}$, suggesting they also fail to capture significant biological insights from the PPI network. SAURON-RF [9], which is, to the best of our knowledge, the best-performing random forest model, was evaluated by the original authors in the LPO scenario with 5-fold CV, but only on a large subset of all cell line-drug pairs of the GDSC dataset and only on 20 to 100 EXP features due to scalability limitations. The authors reported an MSE of 1.96 (corresponding to an RMSE of 1.4), which is significantly worse than all other methods except the mean predictor. This is consistent with the findings in [11], where black-box models such as MLPs generally outperformed more interpretable models like random forests. Hence, we excluded random forests from further analyses.

In the LCO scenario (Fig. 3), MMLP_{EXP} achieves the best performance with a mean RMSE of 1.283. Once again, TGDRP and TGSA perform worse than the baseline TGSA_{id} (mean RMSE: 1.341), reinforcing

the hypothesis that no chemical information is learned from molecular graphs. However, as many models outperform the first and fourth baselines (mean predictor and shuffled x_{EXP}), with $MMLP_{EXP}$ outperforming by 16%, some biological information seems to be captured, especially from EXP data. In contrast to the LPO scenario, COSMIC gene selection in LCO negatively impacts $MMLP_{ALL}$ and $MMLP_{EXP}$.

In the LDO scenario (Fig. 4), TGDRP emerges as the best-performing model. However, Shen et al. [15] reported a significantly higher mean RMSE for TGSA (approximately 2.7, with the exact value provided upon request being 2.6642) compared to our obtained mean RMSE. Therefore, we repeated 5-fold CV five times. Overall, the resulting RMSEs of LPO, LCO, and LDO for TGDRP and TGSA indicate that TGSA performs similarly to TGDRP, but not better. The results from LPO and LCO repetitions were consistent with Figs. 2 and 3, so they

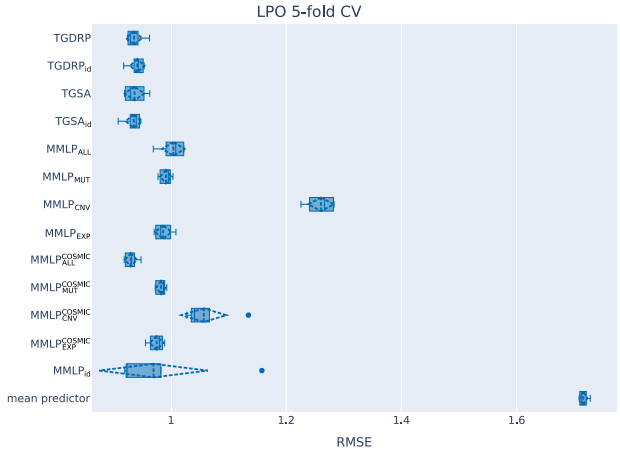

Fig. 2. LPO 5-fold CV results for GDSC2. Each boxplot is computed from the five RMSEs. The additional dashed lines are mean and standard deviation.

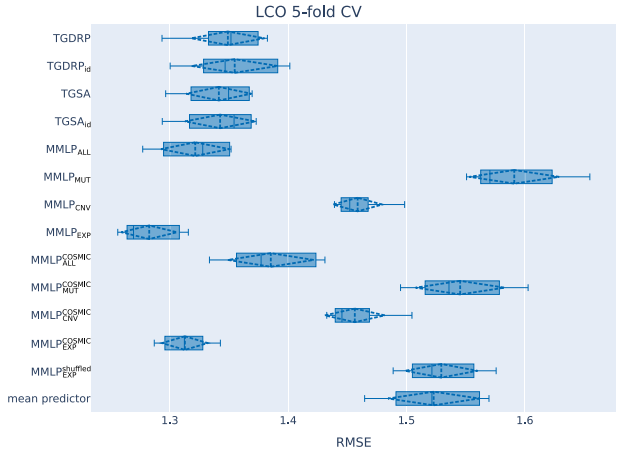

Fig. 3. LCO 5-fold CV results for GDSC2. Each boxplot is computed from the five RMSEs. The additional dashed lines are mean and standard deviation. Note that for the fourth baseline test, only the result for $MMLP_{EXP}^{shuffled}$ is shown because $MMLP_{EXP}$ performed best.

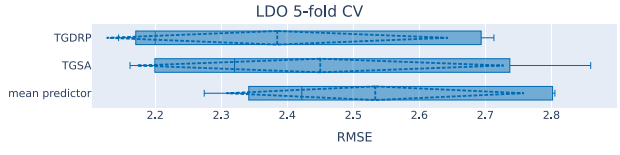

Fig. 4. LDO 5-fold CV results for GDSC2. Each boxplot is computed from the five RMSEs. The additional dashed lines are mean and standard deviation.

are omitted. However, the LDO results showed considerable variation across data splits. The average RMSE and standard deviation across five repetitions for TGDRP, TGSA, and the mean predictor are 2.487 ± 0.280, 2.538 ± 0.280, and 2.536 ± 0.242, respectively. Strikingly, the mean predictor outperformed TGSA and was only 2% worse than TGDRP while having a 14% more stable standard deviation. Given that the RMSEs for TGDRP and TGSA are approximately 2.5, and considering that the drug response values are $\log_{10}$-transformed, this translates to predicted values that are roughly 300 ($\approx 10^{2.5}$) times the original drug response or about $\frac{1}{300}$ of it. Hence, we conclude that the LDO scenario remains inadequately addressed by current methods.

For the smaller DepMap Public 24Q2 dataset with 474 cell lines and 24 drugs, TGDRP is clearly outperformed by MMLP (LPO and LCO) and the mean predictor (LDO), see the Supplementary Material (Section 4).

4. Discussion

We selected GDSC2 for our experiments due to its improved screening methodologies compared to GDSC1 [1], and its larger set of cell lines and drugs than DepMap Public 24Q2. Given the poor predictive performance, we explored whether data quality might also play a role. We used the concordance correlation coefficient (CCC) instead of the widely used Pearson correlation coefficient (PCC) to evaluate data reproducibility, because CCC accounts for both correlation and agreement, unlike PCC, which measures only linear correlation. The CCC is defined as $\text{CCC} = \frac{2\text{PCC}\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}$, where x and y are the two variables with means μ_x and μ_y as well as variances σ_x^2 and σ_y^2, respectively. While the PCC is usually reported with a p-value representing the probability that, if the true correlation were zero, a dataset would yield a PCC at least as far from zero as the observed value in either direction (i.e., a two-sided test) [29], the CCC is usually reported with a 95% confidence interval (CI), often obtained via bootstrap; in our analysis, we used 100 repetitions. The CI reflects the range of plausible values for the true concordance between two datasets and is used instead of a p-value because, unlike a PCC of zero resulting only from no linear correlation, a CCC of (practically) zero can result from substantial bias (e.g., $x = y + 9999$), scale differences (e.g., $x = 1000 \cdot y$), perfect negative correlation, or a combination of these factors, making it difficult to define a single, meaningful null hypothesis.

First, we examined the drug response values, i.e., IC_{50} or AUC values, by comparing duplicate cell line-drug experiments in GDSC2. 9 drugs were tested twice on up to 792 cell lines, resulting in 6288 duplicate cell line-drug pairs. For each drug, we calculated the CCC among IC_{50} values and among AUC values of duplicate experiments, with an average CCC of $0.563\% \pm 0.230$ for IC_{50} and $0.468\% \pm 0.358$ for AUC. While four drugs showed a CCC above 0.7 for both IC_{50} and AUC, three drugs had CCCs below 0.3, even down to 0.013 (CI: [0.010, 0.016]), suggesting that the drug screenings are inconsistent when reproduced. This is further supported by comparisons between IC_{50} values from 368 cell lines and 9 drugs GDSC2 and DepMap Public 24Q2 have in common, which yield a CCC of only 0.409 (CI: [0.388, 0.430]), although this outcome might be influenced as GDSC2 and DepMap Public 24Q2 use different concentration ranges in the drug screenings. Coupled with criticisms of IC_{50} and AUC as drug response metrics such as incomparability across drugs, dependence on concentration range, and ignoring proliferation rate of cell lines [30], these findings highlight the need for different measures, which are currently not included in the GDSC2 and DepMap Public 24Q2 datasets. Potential alternatives include growth rate inhibition (GR) [30] or metrics derived from live-cell imaging data. Like traditional drug screening, live-cell imaging incubates cell lines at a few specific drug concentrations, but instead of providing cell viability at a single time point, it captures images over a range of time points, allowing cell viability to be indirectly inferred by counting cells and offering greater robustness due to the availability of more data points.

Next, we investigated the omics data. Comparing the MUT file and the pan-cancer gene feature file, both from GDSC2 (920 cell lines and 285 genes in common), revealed low overlap, with only 6.0% of mutations shared, and a CCC of 0.084 (CI: [0.082, 0.087]). A similar low overlap is between MUT data of GDSC2 and DepMap Public 24Q2 (1213 cell lines and 18 300 genes in common), with an overlap of 11.7% and a CCC of 0.180 (CI: [0.179, 0.180]). This is consistent with the findings by Ben-David et al. [31] that, although cell lines are generally considered clonal, they are in fact highly genetically heterogeneous, leading to variable drug responses. As DepMap Public 24Q2 measures the relative copy number whereas the CNV data of GDSC2 is categorical, CNV data between the two datasets (941 cell lines and 19 097 genes in common), could not be directly compared with the CCC, so instead we calculated a PCC of 0.519 (p-value $<5 \cdot 10^{-324}$), indicating moderate correlation. Surprisingly, gene expression data measured in transcripts per million showed high agreement (CCC: 0.951, CI: [0.951, 0.952]) between GDSC2 and DepMap Public 24Q2 (998 cell lines and 19 061 genes in common), despite being considered a less reliable biomarker than mutation data due to biological and technical factors [14]. This finding should be explored in future studies. Given these results, we opted not to merge GDSC2 and DepMap Public 24Q2 datasets unlike Zhu et al. [12], as their feature and target matrices lacked strong correlations overall. Altogether, these observations align with our LCO results (Fig. 3), where MMLP_{EXP} and $\text{MMLP}_{\text{EXP}}^{\text{COSMIC}}$ performed best while MMLP_{MUT} and $\text{MMLP}_{\text{MUT}}^{\text{COSMIC}}$ performed worst overall, implying that biological information can be learned from EXP but not MUT. Nevertheless, in the LPO results (Fig. 2), neither MMLP_{EXP} nor $\text{MMLP}_{\text{EXP}}^{\text{COSMIC}}$ surpassed the baseline MMLP_{id}, which uses no biological information at all. This suggests that while EXP provides some valuable information, it is insufficient on its own to give accurate drug response predictions, as even in the LCO results, MMLP_{EXP} predicts values roughly 19 ($\approx 10^{1.28}$) times the original drug response or about $\frac{1}{19}$ of it. Instead, predictions seem mostly influenced by patterns across the drug responses for a cell line, which would also explain the poor results in the LDO scenario (Fig. 4).

Most drugs exert their effects by binding to target proteins [32]. Transcriptomic EXP data serve as a proxy for gene function by capturing how actively a gene is transcribed into messenger RNA (mRNA), an intermediate in protein production. As such, EXP data may outperform genomic MUT and CNV data for drug response prediction. Proteomics data may be even more effective, as they provide direct insights into produced proteins and cellular processes [33]. However, the current usability of proteomics is hindered by missing values, as noted in Section 2.1, highlighting the urgent need to develop methods that ensure comprehensive measurement of proteomic data.

To capture the full complexity of cell line-drug interactions, both suitable biological and chemical features must be effectively integrated. Biological features should be modeled using methods suited to their inherent structure. For example, MUT, CNV, and EXP are unordered, making CNN-based approaches such as tCNNS [5] and GraphDRP [6] inappropriate. Additionally, these data are non-sequential, making recently highly successful sequential models such as recurrent neural networks (RNNs) or transformers inappropriate as well. Regarding chemical data, all current methods struggle to effectively incorporate drug features into drug response prediction models. Although the use of GNNs on molecular graphs to derive drug embeddings appears conceptually sound at first, the poor performance of TGSA suggests that intramolecular signaling among atoms either lacks meaningful chemical basis or fails to contribute useful information. One possible explanation is that most drugs exert their effects by binding to specific target proteins expressed by the cell, a process that is highly dependent on the three-dimensional (3D) structures of both the drug and the protein. Molecular graphs, however, represent only two-dimensional (2D) connectivity and lack explicit 3D structural information, preventing GNNs from capturing drug-target binding interactions, which would also explain why PPI networks did not improve predictive performance. This fundamental limitation also extends to other currently used drug representations such as fingerprints or SMILES, which lack 3D spatial data as well. We therefore propose that future work should focus on predicting drug-target binding directly by leveraging experimentally resolved or predicted protein structures (e.g., via AlphaFold [34,35]) alongside known or computed chemical structures (e.g., via RDKit [23]). These structural representations can serve as inputs to drug-target prediction tools such as DiffDock [36], enabling the prediction of whether and how a drug is likely to bind to a given target, and subsequently, whether such binding leads to a specific cellular response such as growth inhibition, growth stimulation, or cell death. An alternative direction is to revisit the integration of drug-target data. Although it has been reported to perform worse than molecular fingerprints [11], this may be attributed to the large number of unknown or poorly characterized drug-target interactions, as well as the discontinuation of support for the STITCH database [37] since 2015.

Lastly, we observed in the results section that selecting only COSMIC genes improved performance in the LPO scenario but achieved the opposite in the LCO scenario. The improvement in LPO suggests that excluding non-COSMIC genes may prevent overfitting or reduce noise by focusing on important genes only. However, since training on all genes resulted in higher RMSEs for the validation sets, overfitting can be ruled out. The decline in LCO suggests that limiting the model to known cancer-related genes may exclude important predictive features. These two seemingly contradictory findings suggest that a more refined feature selection strategy could enhance performance in both scenarios by balancing feature inclusion and exclusion. However, the excessive runtime of TGSA prevented a comparison with MMLP in feature selection experiments. Furthermore, given the previously discussed inherent data incongruence, such a comparison would likely provide limited insight.

5. Conclusions

We benchmarked the state-of-the-art TGSA model against several baseline tests and MMLP, a simple multi-output multilayer perceptron, across different CV scenarios. While both TGSA and MMLP struggled to learn meaningful biological and/or chemical information, MMLP consistently outperformed TGSA in terms of RMSE in the LPO and LCO scenarios with much shorter runtimes. However, the LDO scenario remains inadequately addressed by current models.

Our findings emphasize the critical need to refine both the acquisition and selection of gene input data, employing more reliable drug response metrics, and improving methods for incorporating chemical information before proposing complex, innovative models. Current and new models can then be adapted or developed for these data sources and should be systematically (re)assessed. For future benchmarking efforts, we recommend conducting baseline tests to ensure no spurious biological or chemical information influences results and to use MMLP as an additional baseline model.

CRediT authorship contribution statement

Nguyen Khoa Tran: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gunnar W. Klau:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Max J. Ried for the provided computational infrastructure and support as well as Torben Schmitz and Marvin Gooßens for the implementation of the imputation mask during backpropagation and the sigmoid layer after the input layer.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patrec.2025.10.016.

Data availability

Code, data, and links to data are found on https://github.com/AlBi-HHU/Drug_Response_Prediction.

References

[1] W. Yang, J. Soares, P. Greninger, et al., Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 (D1) (2012) D955–D961.
[2] R. Arafeh, T. Shibue, J. Dempster, et al., The present and future of the cancer dependency map, Nat. Rev. Cancer 25 (1) (2025) 59–73, URL: https://doi.org/10.1038/s41568-024-00763-x.
[3] Y. Chiu, H. Chen, T. Zhang, et al., Predicting drug response of tumors from integrated genomic profiles by deep neural networks, BMC Med. Genom. 12 (2019) 143–155.
[4] Y. Chang, H. Park, H. Yang, et al., Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature, Sci. Rep. 8 (1) (2018) 8857.
[5] P. Liu, H. Li, S. Li, et al., Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network, BMC Bioinformatics 20 (2019) 1–14.
[6] T. Nguyen, G. Nguyen, T. Nguyen, et al., Graph convolutional networks for drug response prediction, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (1) (2021) 146–154.
[7] M. Li, Y. Wang, R. Zheng, et al., DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (2) (2019) 575–582.
[8] M. Manica, A. Oskooei, J. Born, et al., Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders, Mol. Pharm. 16 (12) (2019) 4797–4806.
[9] K. Lenhof, L. Eckhart, N. Gerstner, et al., Simultaneous regression and classification for drug sensitivity prediction using an advanced random forest method, Sci. Rep. 12 (1) (2022) 13458.
[10] B. Kuenzi, J. Park, S. Fong, et al., Predicting drug response and synergy using a deep learning model of human cancer cells, Cancer Cell 38 (5) (2020) 672–684.
[11] Y. Li, D. Hostallero, A. Emad, Interpretable deep learning architectures for improving drug response prediction performance: myth or reality? Bioinform. 39 (6) (2023) btad390.
[12] Y. Zhu, Z. Ouyang, W. Chen, et al., TGSA: protein–protein association-based twin graph neural networks for drug response prediction with similarity augmentation, Bioinform. 38 (2) (2022) 461–468.
[13] L. Eckhart, K. Lenhof, L. Rolli, et al., A comprehensive benchmarking of machine learning algorithms and dimensionality reduction methods for drug sensitivity prediction, Brief. Bioinform. 25 (4) (2024) bbae242.
[14] S. Mourragui, M. Loog, M. van Nee, et al., Percolate: an exponential family jive model to design dna-based predictors of drug response, in: International Conference on Research in Computational Molecular Biology, Springer, 2023, pp. 120–138.
[15] B. Shen, F. Feng, K. Li, et al., A systematic assessment of deep learning methods for drug response prediction: from in vitro to clinical applications, Brief. Bioinform. 24 (1) (2023) bbac605.
[16] C. Lin, Y. Guan, H. Li, Artificial intelligence approaches for molecular representation in drug response prediction, Curr. Opin. Struct. Biol. 84 (2024) 102747.
[17] F. Mölder, K. Jablonski, B. Letcher, et al., Sustainable data analysis with snakemake, F1000Research 10 (2021).
[18] D. van der Meer, S. Barthorpe, W. Yang, et al., Cell model passports—a hub for clinical, genetic and functional datasets of preclinical cancer models, Nucleic Acids Res. 47 (D1) (2019) D923–D929.
[19] S. Kim, J. Chen, T. Cheng, et al., PubChem 2023 update, Nucleic Acids Res. 51 (D1) (2023) D1373–D1380.
[20] P. Walters, PubChemPy: A python wrapper for the PubChem PUG REST API, 2014, https://github.com/mcs07/PubChemPy.
[21] J. Tate, S. Bamford, H. Jubb, et al., COSMIC: the catalogue of somatic mutations in cancer, Nucleic Acids Res. 47 (D1) (2019) D941–D947.
[22] D. Szklarczyk, R. Kirsch, M. Koutrouli, et al., The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic Acids Res. 51 (D1) (2023) D638–D646.
[23] G. Landrum, P. Tosco, B. Kelley, et al., Rdkit/rdkit: 2022_09_1 (Q3 2022) release (release_2022_09_1), 2022, http://dx.doi.org/10.5281/zenodo.7235579.
[24] N. Tran, L. Kühle, G. Klau, A critical review of multi-output support vector regression, Pattern Recognit. Lett. 178 (2024) 69–75.
[25] V. Borisov, J. Haug, G. Kasneci, CancelOut: A layer for feature selection in deep neural networks, in: Artificial Neural Networks and Machine Learning–ICANN 2019: Deep Learning: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part II 28, Springer, 2019, pp. 72–83.
[26] M. Menden, F. Iorio, M. Garnett, et al., Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties, PLoS One 8 (4) (2013) e61318.
[27] A. Paszke, S. Gross, F. Massa, et al., PyTorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[28] A. Spiess, N. Neumeyer, An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach, BMC Pharmacol. 10 (2010) 1–11.
[29] P. Virtanen, R. Gommers, T. Oliphant, et al., SciPy 1.0: Fundamental algorithms for scientific computing in python, Nature Methods 17 (2020) 261–272, URL: https://doi.org/10.1038/s41592-019-0686-2.
[30] M. Hafner, M. Niepel, M. Chung, et al., Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs, Nature Methods 13 (6) (2016) 521–527.
[31] U. Ben-David, B. Siranosian, G. Ha, et al., Genetic and transcriptional evolution alters cancer cell line drug response, Nat. 560 (7718) (2018) 325–330.
[32] B. Bryant, K. Knights, Pharmacology for Health Professionals, Elsevier Australia, 2011.
[33] N. Branson, P. Cutillas, C. Bessant, Comparison of multiple modalities for drug response prediction with learning curves using neural networks and XGBoost, Bioinform. Adv. 4 (1) (2023) vbad190, http://dx.doi.org/10.1093/bioadv/vbad190.
[34] J. Jumper, R. Evans, A. Pritzel, et al., Highly accurate protein structure prediction with AlphaFold, Nat. 596 (7873) (2021) 583–589.
[35] M. Varadi, S. Anyango, M. Deshpande, et al., AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Res. 50 (D1) (2022) D439–D444.
[36] G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, Diffdock: Diffusion steps, twists, and turns for molecular docking, 2022, arXiv preprint arXiv:2210.01776.
[37] D. Szklarczyk, A. Santos, C. Von Mering, et al., STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data, Nucleic Acids Res. 44 (D1) (2016) D380–D384.