Understanding the Sources of Performance in Deep Learning Drug Response Prediction Models

Nikhil Branson [1 2] **Pedro R. Cutillas** [2] **Conrad Besseant** [1 3]

Abstract

Motivation: Anti-cancer drug response prediction (DRP) using cancer cell lines plays a vital role in stratified medicine and drug discovery. Recently there has been a surge of new deep learning (DL) models for DRP that show promising performance improvements. However, different models use different input data modalities and neural network architectures making it hard to find the source of these improvements.

Results: We consider three DL DRP models with reported state-of-the-art performance and use genomics or transcriptomics profiles of cell lines and chemical profiles of drugs as input data. We also construct null hypothesis benchmarks that simply leverage patterns in truth values and thus do not use omics or chemical profiles. By experimenting with these DL models and comparing them with our null hypothesis benchmarks we show that for multiple testing types, much or all of the current reported performance is a property of the training truth values. Furthermore, we show that for cancer blind testing (testing on unseen cell lines) no performance comes from drug features, instead, performance is due to the transcriptomics cell line profiles.

1. Introduction

Anti-cancer drug response prediction (DRP) has three main aims that, if delivered, would each improve patient outcomes or decrease treatment costs. These aims are to: (a) repurpose existing drugs, (b) tailor more effective treatments to specific groups or individuals and (c) help discover novel drugs. Cancer cell lines are at the heart of most DRP studies because they give a proxy for patient data while offering an abundance of publicly available drug screening data (Sharifi-Noghabi et al., 2021b).

Several large-scale databases have publicly available drug response measurements and omics cell line profiles. These include the Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2012), the cancer cell line encyclopedia (CCLE) and the Cancer Therapeutic Response Portal (CTRP) (Seashore-Ludlow et al., 2015). See (Sharifi-Noghabi et al., 2021b; Baptista et al., 2021) for in-depth comparisons of the different databases. These databases typically contain hundreds of thousands of drug response measurements. For example, GDSC2 has 969 cell lines screened for up to 297 drugs (although not all cell lines have been screened for all drugs).

DRP models commonly use machine learning (ML) algorithms that take omics profiles of cancer cell lines and drug profiles as inputs to predict drug response. With the success of deep learning (DL) in many areas, with varying data types - including images, text and audio (Devlin et al., 2018; He et al., 2016; Liu et al., 2021; Purwins et al., 2019), it was natural for researchers to apply DL to DRP. Indeed, DL methods have been reported to show state-of-the-art performance for DRP (Jiang et al., 2022; Liu et al., 2022; Nguyen et al., 2021). See (Partin et al., 2023) for a review of DL methods in DRP.

DRP models typically feed each of the two inputs (cell line and drug) through separate branches, to encode these inputs. They then use the combined encoded inputs to predict how effective the input drug is for the input cell line. Earlier DRP models used the same subnetwork architectures for the drug and cell line branches. For example, tCNNS used two subnetworks made from convolutional layers (Liu et al., 2019). More recent models have used different subnetworks for the drug and cell line branches to leverage the different modalities of the input data.

Typically DRP models have used convolutional or dense layers for the cell line branches (Liu et al., 2019; Jiang et al., 2022; Liu et al., 2022; Nguyen et al., 2021; Branson et al., 2023; Kim et al., 2021; Liu et al., 2020; Partin et al., 2021; Xia et al., 2022). In addition to these layers, there have been DRP models that use transformers (Jiang et al., 2022), and

[1]School of Biological and Behavioural Sciences, Queen Mary University of London, London E1 4NS, United Kingdom [2]Centre for Genomics and Computational Biology, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, United Kingdom [3]Digital Environment Research Institute, Queen Mary University of London, London E1 1HH, United Kingdom. Correspondence to: Conrad Besseant <c.bessant@qmul.ac.uk>.

.

graph convolutional layers (Liu et al., 2020; Nguyen et al., 2021; Kim et al., 2021) for their drug branches. There is a large variation in the architectures of the drug branches because drugs are three-dimensional chemical substances made of atoms bonded together. Therefore, there are many different ways to both numerically represent drugs and extract features from these representations.

Due to the above, there are multiple differences in the representation of data and network structures of DRP models. Thus, when a new model outperforms an old one it is not clear if the improvement comes from enhancements to the model's architecture, a better representation of the input data or a combination of the two. To compound this problem, DRP models are typically trained and tested in three different ways, corresponding to the three different aims of DRP and when ablation studies are done they are typically only done using one of the testing types. Furthermore, the results and metrics can be stratified in two different ways, which we will show can also significantly impact the reported performance. However, this has been overlooked in previous studies.

Thus, in this paper, we look at how the different data types and subnetworks of three models, with reported state-of-the-art performance, impact their performance for DRP for multiple testing types. The three models we use are tCNNS (Liu et al., 2019), DeepTTA (Jiang et al., 2022), and GraphDRP (Nguyen et al., 2021). Each of these models uses different drug branch architectures. tCNNS uses convolutional layers, DeepTTA uses a transformer and GraphDRP uses graph convolutional layers. We also introduce three simple but effective null hypothesis benchmarks, one for each testing type. These benchmarks do not use any omics data or drug structures. The benchmarks test how much performance improvement using omics data and drug structures adds and show how much performance can be attributed to patterns in the training truth values. This study shows that, for multiple testing types and models, all or most of the performance can be attributed to patterns in the training truth values. Furthermore, we find that, for multiple testing types, no performance comes from input chemical drug structures and instead, performance is due to the transcriptomics profiles.

2. Background and Related Work

2.1. Drug response prediction definition

To make the definition of DRP more concrete, consider a training dataset of size n, $T_n = \{x_{c,i}, x_{d,i}, y_i\}_{i=1}^{n}$. Here $x_{c,i}$ and $x_{d,i}$ are the representation of the ith cell line and drug respectively. y_i is the truth value associated with the ith cell line drug pair, which describes how effective the ith drug is at killing the ith cell line. For example, y could be an IC50 value, the concentration of a drug needed to reduce the activity of a cell line by 50%.

A DRP model, M created by a learning algorithm takes $x_{c,i}$ and $x_{d,i}$ as inputs and predicts for the corresponding truth value $\hat{y}_i$ such that $M(x_{c,i}, x_{d,i}) = \hat{y}_i$.

2.2. Evaluating DRP models

The aim of drug response prediction is to find a model using T that performs well on unseen data. This can be done by evaluating the model on some held-out testing data E. There are three common ways of constructing T and E corresponding to the three different objectives for DRP as outlined in the introduction.

1. Mixed-set: any drug and any cell line can be in either T or E, but a drug cell line pair can only be in one set.

2. Cancer-blind: cell lines in T can not be in E, but drugs can be in either set.

3. Drug blind: drugs in T can not be in E, but cell lines can be in either set.

Generally moving down from 1 to 3 in the above list gives a data-splitting method that leads to more stringent testing. Therefore, the performance of models also tends to decrease as data splitting methods further down the list are used (Partin et al., 2023).

Using mixed-set testing shows if the model could be used for drug repurposing. This is because, for a given dataset a cell line in this dataset is typically not evaluated against all drugs in the dataset. Therefore, the model can be used to evaluate drug cell line pairs that have not yet been explored. This is a quick and inexpensive way to predict if a known drug (drug in the training set) can be repurposed for a known cancer subtype.

Cancer-blind testing shows if the model could be used to find candidate drugs that might be effective for cell lines, not in the training set. Furthermore, cancer-blind testing is a suitable method for simulating how good a model would be in a stratified medicine context. This is because, in a clinical context, a number of examples of patients' responses to a set of drugs will be available for training a model. Then this model would be used to predict which drugs, out of the previously used set of drugs, would be suitable for new patients.

A model that performs well using drug blind testing could be used to find novel anti-cancer drugs. This is because the model would be able to predict the responses of a candidate drug to known cell lines.

2.3. Stratification of results in metric calculations

DRP models performance metrics can be stratified in two different ways, by cell line or by drug. We consider cancer blind testing in the following to make our examples concrete. We define stratifying results by cell line as finding a metric for each cell line in the test set individually, across all drugs, before averaging the results. Similarly, we define stratifying the results by drug as finding a metric for each drug, across all cell lines in the test set, before averaging the results. In contrast, a non-stratified metric is found once across all cell line drug pairs in the testing set without averaging. Thus, there are three different ways performance metrics can be reported. In DRP studies this is typically not discussed and only stratification by cell line is reported. However, as we will show, how results are stratified makes a big impact on cancer blind testing results.

The two different methods for stratification correspond to evaluating the model for distinct problems the model could be applied to. Stratifying by cell line corresponds to simulating the performance in a clinical stratified medicine context which aims to recommend a drug out of all the drugs the model is trained on, for a new patient. A model that performs well using this testing would also be directly useful in a drug discovery context where cell lines are used to narrow down candidate drugs. Specifically, the model could be used to rank treatments from the set of all drugs the model has been trained on for an unseen cancer sub-type.

In contrast, stratifying by drug corresponds to simulating the performance in a clinical stratified medicine context where there is a set of new patients and you want to recommend if a given drug should be taken for each of these patients. This is typically the scenario in which DRP models are applied to (Gerdes et al., 2021; Shubham et al., 2024; He et al., 2022; Sharifi-Noghabi et al., 2021a). The corresponding application in a drug discovery context directly using cell lines is where you have a candidate drug that has already been screened for some cell lines and you can then use the model to screen a set of many unseen cancer sub-types for efficacy. We provide the equations for calculating the stratified metrics in Appendix A.

2.4. Neural Network Architectures in DRP

The mathematical properties of different neural network layers mean different layers and thus architectures are best suited to different problems. Typically the choice of layers is based on the modality of the input data. For DRP, because drugs can be mathematically represented in multiple different ways it leads to many viable options for types of neural network layers that can be used, as we will see below.

2.4.1. Dense layers

Due to their power and simplicity dense layers can be thought of as the default choice for drug and cell line branches in DRP and multiple studies have used multilayer perceptions (MLPs) for both branches (Menden et al., 2013; Partin et al., 2021; Sakellaropoulos et al., 2019; Zhao et al., 2019). Furthermore, some studies that use more specialised layers for their drug branches, which will be discussed below, still use MLPs for their cell line branches (Jiang et al., 2022; Liu et al., 2022).

2.4.2. Convolutional layers

Convolutional layers have a couple of key properties that distinguish them from dense layers. Firstly, the weights are applied over an area of the input, therefore they combine information about parts of the input that are close in input space. Secondly, for a given channel, the same weights are applied over the full input, thus, they can infer motifs (Khan et al., 2020). These properties have helped both $1D$ and $2D$ convolutional neural networks achieve great success in multiple applications that can take advantage of these properties (He et al., 2016; Khan et al., 2020; Kiranyaz et al., 2021). For example, for image data repeating patterns such as edges can be identified with convolutional layers (Zeiler & Fergus, 2014).

$1D$ convolutional layers have been used for multiple DRP studies (Liu et al., 2019; Nguyen et al., 2021; Liu et al., 2020; Joo et al., 2019; Chang et al., 2018; Branson et al., 2023). Most notably for this paper tCNNS uses $1D$ convolutional layers in both the drug and omics branches.

Theoretically, convolutional layers can exploit drug data in the ways described above. They could pick out repeating patterns and capture information about local interactions of atoms within a drug.

It is less clear that the above properties of convolutional layers can be fully used to take advantage of omics data. This is because in DRP models the omics data is not normally in a particular order so they do not necessarily contain local spatial patterns or imply interactions between local features.

2.4.3. Attention

Generally, attention in deep learning is a way of assigning importance values to different parts of an input based on its context. Here we will look at attention and multi-head attention, as defined in (Vaswani et al., 2017). Attention is applied to three matrices called the query (Q), key (K) and value (V) that are derived from the input to the attention module.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{d_k})V.$$

Multi-head attention is given by the concatenation of multiple repeats of attention, where each repeat is called a *head*. Thus,

$$MultiHead(Q, K, V) = Concat(head_i, ...head_h)W^O.$$

Where

$$head_i = Attention(QW_i^Q, KW_i^k, VW_i^v).$$

Here W_i^Q, W_i^k, W_i^v and W^O are trainable weight matrices. In Multi-head self-attention, Q, K and V are all set as the same input. Self-attention assigns a weight for a given part of an input sequence, relative to the context in that input. This allows the network to give more weight to parts of inputs that are related. For example, for an input sentence with 'she' in it, the weights for 'she' are typically stronger for all references to 'she' in the input sentence such as 'girl' (Vig, 2019). Self-attention can also successfully capture long-range interactions (Vaswani et al., 2017).

The above properties have helped the neural networks that use attention achieve impressive performance for data types that can take advantage of these properties, such as language and images (Vaswani et al., 2017; Devlin et al., 2018; Liu et al., 2021).

In DRP, attention-based models have been used for the encoding of both omics and drug data (Oskooei et al., 2018; Manica et al., 2019; Jiang et al., 2022; Tao et al., 2020). In theory, the attention mechanism will weight parts of the omics and drug inputs that are the most important for DRP. Attention also allows for the interrogation of the model to see if the inputs the model gave the most weight match up with biochemical mechanisms.

2.4.4. Graph neural networks

Consider the representation of a graph, $G = (V, E)$, where V are the nodes of the graph and E are the edges of the graph. G can be represented using two matrices X and A. Here X is the feature matrix that describes the features of each node. A is the adjacency matrix that gives the connections between the nodes (graph edges).

There are multiple types of graph neural network (GNN) layers that propagate the information through G, using the local neighbourhoods of nodes. These include graph convolutional network (GCN) (Kipf & Welling, 2017), graph attention network (GAT) (Velickovic et al., 2017) and graph isomorphism network (GIN) (Xu et al., 2019) layers.

We will look in more depth at a GIN layer as an example because it was reported to outperform GCN and GAT layers for DRP (Nguyen et al., 2021) and we evaluate the GIN variant of GraphDRP in this paper. The new value of a node X_j after being passed through a GIN layer, Y_i is given by

$$Y_i = MLP\left((1+\mu)X_i + \sum_{j \in N(i)} X_j\right).$$

Where μ is a learnable scalar and the sum runs over $N(i)$ the nodes that are adjacent to X_i and MLP is a multi layer perceptron.

Graphs don't typically contain useful information in the ordering of their nodes, it's the edges of the graph that matter. Thus, to respect this property the output of a layer should be the same for isomorphic graphs (Gaudelet et al., 2021). GNNs have this property and it can be seen in the equation for a GIN layer where only X_i and edges of X_i are used to update X_i. Thus, by using a node's edges and features GINs and GNNs can exploit graphical structures of the input data to create graph embeddings.

GNNs have shown state-of-the-art results when applied to data that can naturally be represented as graphs (Zhou et al., 2020). Chemicals can be represented as graphs, where nodes are atoms and edges are chemical bonds between the atoms. Therefore, many applications that use chemicals as inputs have taken advantage of GNNs (Gaudelet et al., 2021), this includes DRP (Nguyen et al., 2021; Liu et al., 2022; Zhu et al., 2022; Liu et al., 2020; Chu et al., 2022).

2.5. Datasets and models

2.5.1. DRP models

We re-created tCNNS (Liu et al., 2019), GraphDRP (Nguyen et al., 2021), and DeepTTA (Jiang et al., 2022) models using the code made available in these papers. tCNNS uses convolutional layers for both cell line and drug branches. DeepTTA uses a transformer for its drug branch and an MLP for its cell line branch. GraphDRP uses graph convolution layers for its drug branch and convolutional layers for its cell line branch. Here we use the GIN variant of GraphDRP as it is reported to give the best performance (Nguyen et al., 2021). Once the learned drug and cell line representations are found and concatenated, all models use an MLP for prediction of IC50 values. An 80%, 10%, 10% train, validation, and test split was used for all testing and the validation data was used to select the optimal number of epochs. We ran the models for three train test splits and for each split we ran the model for three different seeds. We also repeated this for each of the testing types: mixed set, cancer blind and drug blind. We evaluated the models on the test set and found three metrics commonly used in DRP. The mean squared error (MSE), coefficient of determination R^2, and Pearson correlation coefficient (Pear).

2.5.2. Cell line representations and drug response data

For genomics cell line profiles we used genomics data from the genomics of drug sensitivity in cancer database (GDSC) (Yang et al., 2012). Genomics profiles that included genetic mutation and copy number variations information were used to recreate tCNNS and GraphDRP. For transcriptomic profiles, we used RNA-seq data from GDSC. RNA-seq profiles were used to recreate DeepTTA. For drug response data we used IC50 values from GDSC2. Where IC50 is the standard measure of drug response and was used in the original papers for tCNNS DeepTTA and GraphDRP. We only kept cell lines that were in all three of the above datasets. This led to 904 cell lines being used for the following analysis.

The omics data in GDSC has already undergone standard preprocessing, where the transcriptomics data is preprocessed using the robust multi-array analysis algorithm (RMA) (Irizarry et al., 2003). This is the same data that was used in DeepTTA. Similarly, we directly used the binary genomics data from GDSC as was done in tCNNS and GraphDRP. Furthermore, the IC50 values provided by GDSC are natural logarithm transformed

2.5.3. Drug representations

SMILES (simplified molecular-input line-entry system) and molecular graphs were used as the drug representations with chemical properties. SMILES were used in the tCNNS and DeepTTA models, while molecular graphs were used in the GraphDRP model. In a SMILES representation, each molecule is represented as a string of characters. In molecular graph representations of molecules/drugs, each node represents an atom in the molecule and each edge represents a bond between the atoms We downloaded SMILES representations from PubChem (Kim et al., 2023).

2.5.4. Dataset size

904 cell lines and 181 drugs were used in the following analysis. Thus, there are a possible 163,624 drug cell line pairs. However, not all drug cell line pairs have IC50 values in GDSC2. Thus, removing drug cell line pairs without IC50 values left us with 147,713 drug cell line pairs which is the final size of our full dataset.

2.6. Null Hypothesis Benchmarks

We create three null hypothesis benchmarks, (1) drug average (2) cell line (CL) average and (3) marker benchmark, for use in cancer blind, drug blind and mixed set testing respectively. None of these benchmarks use any omics data or drug structures. We note that the drug and CL average benchmarks have recently been shown to have strong performance (Branson et al., 2023; Li et al., 2023) but have not yet been fully explored. The benchmarks are defined as follows:

Drug average benchmark: predictions for response values for a given drug in the test set were simply calculated as the mean of all response values for that drug in the training set. Thus, for the drug average benchmark the prediction for the drug, d, and any cell line, $c, \in C_{test}$, $\hat{y}_{dc_{test}}$ is given by

$$\hat{y}_{dc_{test}} = \frac{\sum_{c \in C_{train}} y_{dc}}{n_{dC_{train}}}.$$

The sum runs over all cell lines in the training set $c \in C_{train}$ where drug cell line pair (c, d) has an associated response value y_{dc}. $n_{dC_{train}}$ is the number of drug cell line pairs in the training set that include d. Note that due to missing response values $n_{dC_{train}}$ can be different for each drug.

CL average benchmark: the above was done with cell lines instead of drugs. Thus, the prediction for the cell line, c and any drug in the testing set $d \in D_{test}$ is given by

$$\hat{y}_{d_{test}c} = \frac{\sum_{d \in D_{train}} y_{dc}}{n_{cD_{train}}}.$$

Where $n_{D_{train}c}$ is the number of drug cell line pairs in the training set that include c.

Marker benchmark: we define a marker representation of either a cell line or drug to mean a vector that uniquely identifies that cell line or drug but doesn't have any biological or chemical properties. Here, we use a column vector to achieve this by one-hot encoding each drug or cell line. The marker benchmark took cell line and drug inputs as marker representations and fed them through an MLP.

We hypothesised each benchmarks were effective for similar reasons:

- Drug average benchmark: there are drugs that are generally effective at killing most cell lines and other drugs that are ineffective for most cell lines.
- CL average benchmark: there are cell lines that are generally harder or easier to kill for most drugs.
- Marker benchmark: all drugs and cell lines are in the training set for mixed-set evaluation. Thus, a combination of the above properties can be learnt during training for inference.

Importantly, all of the null hypothesis benchmarks allowed us to see if adding omics data or drug structures improves performance and how much of model performance can be attributed to the above property of the data.

3. Results and Discussion

3.1. Mixed set testing

Table 1 shows the results of mixed set testing for the three published models and the marker model benchmark. It shows that none of the published models outperform the marker model, for all metrics. The marker model does not use omics data or chemical drug features. This means the performance of the model comes from the truth values in the training set. Therefore, the published models are unable to use omics or drug data to improve performance over what information can be gained from the truth values in the training set. This result shows the importance of the benchmark and that it should be used when evaluating future DRP models for mixed set testing.

A key reason the marker model is so successful is that, for mixed set testing, every drug and cell line in the test set is also in the training set while unique drug cell line pairs are only in the training or testing set. Therefore, the model has ample information on the distribution of IC50 values for any cell line or drug in the training set to learn from. Thus, rather than having to learn important cell lines or drug features, it can just use these distributions for predictions.

Comparing just the published models our results agree with the ones reported in the original DeepTTA and GraphDRP papers, where it was shown DeepTTA outperforms tCNNS (Jiang et al., 2022) and GraphDRP outperforms tCNNS respectively (Nguyen et al., 2021).

We repeated the above analysis for two further test train splits, which produced similar values to those in Table 1. Full results can be found in Appendix D.

3.2. Cancer blind testing

Table 2 shows the results of cancer blind testing for the published models' drug average benchmark and DeepTTA-DB. We created DeepTTA minus drug branch (DeepTTA-DB) to quantify the impact of chemical drug structures on cancer blind performance. For DeepTTA-DB, we replace DeepTTA's transformer drug branch with a one-hot encoded marker representation of the drugs. The reported performance difference seen between the stratified and non-stratified results (defined in section 2.3) shows that cell line stratification is an easier problem. This is expected and is due to the statistics of the response values specifically due to observation 1:

Observation 1. *Much of the variance in the response values between drugs can be explained by the average behaviour of the drugs.*

We note that Observation 1 is a direct implication of the hypothesis made in section 2.6, that there are drugs that are generally effective or ineffective. The drug average benchmark clearly demonstrates Observation 1 as it is directly built to reflect this property. Thus, when we test across all of the drugs, Observation 1 boosts the performance of the models. In contrast with drug stratification where performance is evaluated for a given drug, the variance cannot be explained by Observation 1 as we are looking across cell lines for that drug. This is again shown by the drug average benchmark that predicts the same value, for a given drug across all cell lines, and thus, gives an $R^2 \sim 0$ and has an undefined Pearson correlation coefficient. This is also why the models have a much better performance differential compared with the benchmark for drug stratification than for cell line stratification. This shows the importance of considering how results are stratified, whereas DRP studies typically only report cell line stratified metrics and to our knowledge do not consider drug stratification. Furthermore, it shows that deep learning models can have the most impact in a drug stratified use case.

Table 2 also shows the drug average benchmark outperforms both of the models that use genomic cell line profiles, tCNNS and GraphDRP, in terms of all metrics bar drug stratified Pearson correlation. Thus, these models are not generally learning useful features above those that can be derived from Observation 1. On the other hand, the two models that use transcriptomic cell line profiles, DeepTTA and DeepTTA-DB, comfortably outperform the benchmark for all metrics. This suggests that transcriptomic profiles play a key role in cancer blind testing. To further explore this result we constructed tCNNS_Tran and GraphDRP_Tran, where we replaced the genomic cell line profiles of tCNNS and GraphDRP with transcriptomic profiles with all genes as features. The results for these models are shown in Table 3. The table shows that tCNNS_Tran and GraphDRP_Tran both outperform the benchmark, further showing that the strong performance is caused by the transcriptomic cell line profiles. There is also a strong theoretical justification for this result. Many of the processes in cancer can not be explained by genomic profiles (mutation and copy number variation data) alone (Lin & Yang, 2019). In contrast, transcriptomics operates at the level of post-transcription so contains vital information about biochemical pathways, important for understating diseases (Khodadadian et al., 2020). Therefore, transcriptomics describes parts of the biology of cancer that genomics can not.

Table 2 also shows that DeepTTA-DB and DeepTTA perform equally within the margin of uncertainty. Therefore, the performance is due to the omics branch, further confirming that the transcriptomic cell line profiles cause the performance improvement seen over the drug average benchmark. We further explore this result by removing the drug branch from tCNNS_Tran and GraphDRP_Tran to create tCNNS_Tran-DB and GraphDRP_Tran-DB. As with DeepTTA-DB, tCNNS_Tran-DB and GraphDRP_Tran-DB

Table 1. Metrics for mixed set testing for three models from the literature: tCNNS, DeepTTA and GraphDRP and a benchmark (marker model) that does not use omics data or drug features. The table is in order from highest to lowest MSE. The results are for one test train split and the metrics are the average of three seeds for this split. The uncertainties are the standard deviation across these seeds. The table shows that for mixed set testing, the published models do not outperform the benchmark for any metric.

Method	Drug branch	Omics branch	MSE	R2	Pear
tCNNS	SMILES	Genomics	1.256 ± 0.004	0.8283 ± 0.0005	0.9104 ± 0.0003
GraphDRP	Molecular graph	Genomics	1.06 ± 0.04	0.856 ± 0.005	0.930 ± 0.001
DeepTTA	SMILES	Transcriptomics	0.98 ± 0.01	0.866 ± 0.002	0.931 ± 0.001
Marker Model	Marker	Marker	**0.88 ± 0.01**	**0.880 ± 0.001**	**0.9379 ± 0.0007**

Table 2. Metrics for cell line and drug stratified cancer blind testing, for three models from the literature: tCNNS, DeepTTA, GraphDRP, DeepTTA-DB (a model that we create), and our null hypothesis drug average benchmark.

Method	MSE cl strat	Pear cl strat	R2 cl strat	Pear drug strat	R2 drug strat
tCNNS	2.41 ± 0.04	0.870 ± 0.003	0.61 ± 0.01	0.16 ± 0.02	-0.21 ± 0.01
GraphDRP	2.41 ± 0.08	0.872 ± 0.004	0.60 ± 0.01	0.220 ± 0.009	-0.26 ± 0.06
Drug Average	2.190	0.884	0.640	N/A	-0.012
DeepTTA-DB	1.575 ± 0.009	0.902 ± 0.001	**0.743 ± 0.004**	**0.526 ± 0.009**	**0.257 ± 0.004**
DeepTTA	**1.57 ± 0.02**	**0.904 ± 0.002**	**0.743 ± 0.005**	0.523 ± 0.006	0.253 ± 0.008

are tCNNS_Tran and GraphDRP_Tran with a one-hot encoded marker representation of the drugs used in place of the respective drug branches. The results for these models are shown in Table 3. The table shows that removing the drug branch from both models does not impact performance. This supports the above discussion, that the performance improvements over the benchmark are due to the transcriptomic cell line profiles and the models do not learn useful representations from the chemical drug structures.

We repeated the above analysis for two further test train splits. The values of all metrics (see Appendix E) are similar to those in Table 2 and 3, further supporting the conclusions arrived at in the above discussion.

3.3. Drug blind testing

For drug blind testing we found a greater variation between different train test splits than for the other testing types. Therefore, in this section, we look in more detail at each test train split.

Table 4 gives the results for non-stratified drug blind testing for all three test train splits. The results are shown for the three published models and the cell line (CL) average benchmark. The metrics that are better than, or have a bound that is better than, the CL average benchmark are in bold. For example, Table 4 shows that for the first split, tCNNS outperforms the benchmark in terms of R^2 and Spearman correlation. However, the benchmark is within the range of MSE and Pearson correlation values given by the uncertainties. In general, Table 4 shows that while there are large differences between the test train splits, for each split there are metrics that outperform the CL average benchmark, for tCNNS and GraphDRP. Therefore, the omics and drug features do provide a benefit for non-stratified drug blind testing.

The uncertainties in Table 4 are much larger than the uncertainties for cancer blind and mixed set testing This shows the instability in training the published models for drug blind testing. It is important to note that these uncertainties are from the same models being re-trained for different seeds with the same data. Thus, the uncertainties are due to the stochastic nature of model training, not a change in the data.

The fact that the published models outperform the CL benchmark for some splits suggests there is useful information in the input data for drug blind testing. However, more needs to be done to improve the models' drug blind testing abilities so they outperform the benchmark consistently. One way of doing this would be to improve the stability of the models. This could be done methodologically by improving the architecture or training of the models. For example, by decreasing the complexity of the models. Collecting more training data would also help with improving stability as well as performance (Branson et al., 2023; Partin et al., 2021). Another way that could improve performance is using transfer learning with the vast amount of unlabelled chemical structures of drugs that are publicly available to pre-train the model's drug branch. Furthermore, as advances in deep learning are made it seems very likely that DRP models could be improved, so they extract more useful in-

Table 3. Cancer blind cell line stratified metrics for tCNNS_Tran and GraphDRP_Tran (tCNNS and GraphDRP but by replacing the genomics cell line profiles with transcriptomics profiles), and for CNNS_Tran-DB and GraphDRP_Tran-DB, removing the drug branch from the respective models

Method	MSE	R2	Pear
tCNNS	2.41 ± 0.04	0.61 ± 0.01	0.870 ± 0.003
GraphDRP	2.41 ± 0.08	0.60 ± 0.01	0.872 ± 0.004
Drug Average	2.190	0.640	0.884
tCNNS_Tran	1.83 ± 0.05	0.698 ± 0.008	0.8865 ± 0.0005
tCNNS_Tran-DB	1.78 ± 0.04	0.702 ± 0.007	0.8919 ± 0.0009
GraphDRP_Tran	1.77 ± 0.02	0.714 ± 0.004	0.898 ± 0.002
GraphDRP_Tran-DB	**1.56 ± 0.01**	**0.743 ± 0.003**	**0.9062 ± 0.0008**

Table 4. Metrics for non-stratified drug blind testing for three train test splits. The metrics in **bold** are better than or have bounds better than, the CL average benchmark.

Split	Method	MSE	R2	Pear
1	GraphDRP	11.0 ± 3	-0.5 ± 0.4	-0.0 ± 0.2
	DeepTTA	7.5 ± 0.5	-0.04 ± 0.07	**0.28 ± 0.07**
	tCNNs	**7.0 ± 1**	**0.0 ± 0.2**	**0.3 ± 0.1**
	CL Average	6.493	0.093	0.318
2	GraphDRP	6.0 ± 1	-0.1 ± 0.2	**0.3 ± 0.2**
	DeepTTA	5.6 ± 0.2	-0.02 ± 0.04	**0.42 ± 0.03**
	tCNNs	6.4 ± 0.8	-0.2 ± 0.2	**0.3 ± 0.1**
	CL Average	4.999	0.090	0.348
3	GraphDRP	11.2 ± 0.8	-0.27 ± 0.09	0.06 ± 0.08
	DeepTTA	**7.7 ± 0.6**	**0.13 ± 0.07**	**0.47 ± 0.01**
	tCNNs	**6.7 ± 0.2**	**0.24 ± 0.02**	**0.52 ± 0.04**
	CL Average	8.113	0.082	0.288

formation from the input data more reliably, thus improving performance.

4. Conclusion

In this paper, we have recreated three drug response prediction (DRP) models, which all have different architectures, with reported state-of-the-art performance. These published models used omics data to represent cell lines and SMILES strings or molecular graphs to represent drugs. We also created null hypothesis benchmarks that did not use omics data or chemical drug features to compare with the published models. These benchmarks show strong performance relative to the models, suggesting that they should be used when evaluating future DRP models.

For cancer blind testing we found that the published models that used genomics cell line profiles were generally outperformed by the null hypothesis benchmarks. On the other hand, the models that used transcriptomics outperformed the benchmark, suggesting that performance is due to the transcriptomics cell line profiles. We confirmed this by retaining the models that used genomics with transcriptomics cell line profiles instead. We also found that for all of the models, none of the performance comes from their chemical drug structures. Furthermore, the input datatype used has more effect on performance than the architecture. We also showed that how the results are stratified significantly impacted performance values and performance differentials to the benchmark. We showed that this is because much of the variance in the response values between drugs can be explained by the average behaviour of the drugs. Therefore, when stratified by cell line the benchmark reflects the performance that this causes and it boosts the performance of the models. However, this is not the case when evaluating using drug stratification as the variation in efficacy of a given drug, across a set of cell lines, clearly can not be explained by its average behaviour.

For mixed set testing, we showed that none of the models were able to outperform the benchmark. This suggests that useful information from omics cell line profiles and

Table 5. Metrics for stratified drug blind testing.

Split	Method	Drug branch	Omics branch	MSE	R2	Pear
1	GraphDRP	Molecular graph	Genomics	10.0 ± 2	**-4.0 ± 2**	0.56 ± 0.02
	DeepTTA	SMILES	Transcriptomics	7.2 ± 0.5	**-2.3 ± 0.3**	0.55 ± 0.03
	tCNNS	SMILES	Genomics	**6.4 ± 0.9**	**-2.1 ± 0.4**	0.56 ± 0.03
	CL Average	None	None	6.228	-2.154	0.612
2	GraphDRP	Molecular graph	Genomics	6.0 ± 1	-2.5 ± 0.5	**0.5 ± 0.2**
	tCNNS	SMILES	Genomics	6.0 ± 0.3	-2.4 ± 0.2	0.569 ± 0.002
	DeepTTA	SMILES	Transcriptomics	5.4 ± 0.2	-2.4 ± 0.1	0.584 ± 0.009
	CL Average	None	None	4.988	-2.18	0.627
3	GraphDRP	Molecular graph	Genomics	10.0 ± 1	-4.3 ± 0.8	0.60 ± 0.04
	DeepTTA	SMILES	Transcriptomics	**8.0 ± 1**	**-3.2 ± 0.5**	0.61 ± 0.01
	CL Average	None	None	7.633	-3.279	0.659
	tCNNS	SMILES	Genomics	**7.0 ± 0.5**	**-2.3 ± 0.3**	0.62 ± 0.01

chemical drug features can not be leveraged for mixed set testing. Instead, the performance is due to patterns in the truth values.

This paper also opens up interesting future directions for DRP. For cancer blind testing our results suggest that improving the cell line branch would be a very effective way to increase performance. Furthermore, work could be done looking at how drug features can be integrated to improve cancer blind performance, e.g. by using fusion methods other than concatenation. For drug blind testing our results suggest that improving the stability of model training is an important future direction, as well as general model performance improvements. This could, for example, be achieved using transfer learning with the vast amount of unlabelled chemical structures of drugs that are publicly available.

Software

The code is made available at https://github.com/Nik-BB/Understanding_DRP_models

Acknowledgments

This work was supported by PhD studentship funding from the Wellcome Trust 218584/Z/19/Z. This work also utilised computational resources from the Andrena cluster in Queen Mary's Apocrita HPC facility supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045.

References

Baptista, D., Ferreira, P. G., and Rocha, M. Deep learning for drug response prediction in cancer. *Briefings in bioinformatics*, 22(1):360–379, 2021.

Branson, N., Cutillas, P. R., and Bessant, C. Comparison of multiple modalities for drug response prediction with learning curves using neural networks and xgboost. *Bioinformatics Advances*, pp. vbad190, 2023.

Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T. S., Jung, J., and Shin, J.-M. Cancer drug response profile scan (cdrscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports*, 8(1):1–11, 2018.

Chu, T., Nguyen, T. T., Hai, B. D., Nguyen, Q. H., and Nguyen, T. Graph transformer for drug response prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1065–1072, 2022.

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

Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., Hayter, J. B., Vickers, R., Roberts, C., Tang, J., et al. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics*, 22(6):bbab159, 2021.

Gerdes, H., Casado, P., Dokal, A., Hijazi, M., Akhtar, N., Osuntola, R., Rajeeve, V., Fitzgibbon, J., Travers, J., Britton, D., et al. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nature communications*, 12(1):1–15, 2021.

He, D., Liu, Q., Wu, Y., and Xie, L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence*, 4(10):879–892, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2): 249–264, 2003.

Jiang, L., Jiang, C., Yu, X., Fu, R., Jin, S., and Liu, X. Deeptta: a transformer-based model for predicting cancer drug response. *Briefings in bioinformatics*, 23(3): bbac100, 2022.

Joo, M., Park, A., Kim, K., Son, W.-J., Lee, H. S., Lim, G., Lee, J., Lee, D. H., An, J., Kim, J. H., et al. A deep learning model for cell growth inhibition ic50 prediction and its application for gastric cancer patients. *International journal of molecular sciences*, 20(24):6276, 2019.

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516, 2020.

Khodadadian, A., Darzi, S., Haghi-Daredeh, S., Sadat Eshaghi, F., Babakhanzadeh, E., Mirabutalebi, S. H., and Nazari, M. Genomics and transcriptomics: the powerful technologies in precision medicine. *International Journal of General Medicine*, pp. 627–640, 2020.

Kim, S., Bae, S., Piao, Y., and Jo, K. Graph convolutional network for drug response prediction using gene expression data. *Mathematics*, 9(7):772, 2021.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1): D1373–D1380, 2023.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y., Hostallero, D. E., and Emad, A. Interpretable deep learning architectures for improving drug response prediction performance: myth or reality? *Bioinformatics*, 39 (6):btad390, 2023.

Lin, V. T. and Yang, E. S. The pros and cons of incorporating transcriptomics in the age of precision oncology. *JNCI: Journal of the National Cancer Institute*, 111(10):1016–1022, 2019.

Liu, P., Li, H., Li, S., and Leung, K.-S. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC bioinformatics*, 20(1): 1–14, 2019.

Liu, Q., Hu, Z., Jiang, R., and Zhou, M. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement_2):i911–i918, 2020.

Liu, X., Song, C., Huang, F., Fu, H., Xiao, W., and Zhang, W. Graphcdr: a graph neural network method with contrastive learning for cancer drug response prediction. *Briefings in Bioinformatics*, 23(1):bbab457, 2022.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Manica, M., Oskooei, A., Born, J., Subramanian, V., Sáez-Rodríguez, J., and Rodríguez Martínez, M. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*, 16(12):4797–4806, 2019.

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., and Saez-Rodriguez, J. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8 (4):e61318, 2013.

Nguyen, T., Nguyen, G. T., Nguyen, T., and Le, D.-H. Graph convolutional networks for drug response prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(1):146–154, 2021.

Oskooei, A., Born, J., Manica, M., Subramanian, V., Sáez-Rodríguez, J., and Martínez, M. R. Paccmann: prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. *arXiv preprint arXiv:1811.06802*, 2018.

Partin, A., Brettin, T., Evrard, Y. A., Zhu, Y., Yoo, H., Xia, F., Jiang, S., Clyde, A., Shukla, M., Fonstein, M., et al. Learning curves for drug response prediction in cancer cell lines. *BMC bioinformatics*, 22:1–18, 2021.

Partin, A., Brettin, T. S., Zhu, Y., Narykov, O., Clyde, A., Overbeek, J., and Stevens, R. L. Deep learning methods for drug response prediction in cancer: predominant and emerging trends. *Frontiers in Medicine*, 10:1086097, 2023.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13 (2):206–219, 2019.

Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., Moss, T. J., Piha-Paul, S., Zhou, H., Kardala, E., et al. A deep learning framework for predicting response to therapy in cancer. *Cell reports*, 29(11):3367–3373, 2019.

Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., Jones, V., Bodycombe, N. E., Soule, C. K., Gould, J., et al. Harnessing connectivity in a large-scale small-molecule sensitivity datasetharnessing connectivity in a sensitivity dataset. *Cancer discovery*, 5 (11):1210–1223, 2015.

Sharifi-Noghabi, H., Harjandi, P. A., Zolotareva, O., Collins, C. C., and Ester, M. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nature Machine Intelligence*, 3 (11):962–972, 2021a.

Sharifi-Noghabi, H., Jahangiri-Tazehkand, S., Smirnov, P., Hon, C., Mammoliti, A., Nair, S. K., Mer, A. S., Ester, M., and Haibe-Kains, B. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Briefings in Bioinformatics*, 22(6):bbab294, 2021b.

Shubham, K., Jayagopal, A., Danish, S. M., AP, P., and Rajan, V. Wiser: Weak supervision and supervised representation learning to improve drug response prediction in cancer. *arXiv preprint arXiv:2405.04078*, 2024.

Tao, Y., Ren, S., Ding, M. Q., Schwartz, R., and Lu, X. Predicting drug sensitivity of cancer cell lines via collaborative filtering with contextual attention. In *Machine Learning for Healthcare Conference*, pp. 660–684. PMLR, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al. Graph attention networks. *stat*, 1050 (20):10–48550, 2017.

Vig, J. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*, 2019.

Xia, F., Allen, J., Balaprakash, P., Brettin, T., Garcia-Cardona, C., Clyde, A., Cohn, J., Doroshow, J., Duan, X., Dubinkina, V., et al. A cross-study analysis of drug response prediction in cancer cell lines. *Briefings in bioinformatics*, 23(1):bbab356, 2022.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1): D955–D961, 2012.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

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

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

Zhu, Y., Ouyang, Z., Chen, W., Feng, R., Chen, D. Z., Cao, J., and Wu, J. Tgsa: protein–protein association-based twin graph neural networks for drug response prediction with similarity augmentation. *Bioinformatics*, 38(2):461–468, 2022.

A. Calculating stratified metrics

For a cell line stratified metric MC_{strat},

$$MC_{strat} = \frac{\sum_{c \in C} M_c}{N_c}. \tag{1}$$

Where the sum runs over, C all cell lines in the test set and N_c is the number of cell lines in the test set. M_c is the metric M for cell line c such that

$$M_c = f(y_c, \hat{y_c}).$$

Where y_c and $\hat{y_c}$ are the truth values and predicted values for the cell line drug pairs that include c respectively. f is a function that gives metric M for example it could be the mean squared error. In contrast drug-stratified metric is instead given by MD_{strat},

$$MD_{strat} = \frac{\sum_{d \in D} M_d}{N_d} \tag{2}$$

Where the sum runs over, D all drugs and N_d is the number of drugs. M_d is the metric M for drug d such that

$$M_d = f(y_d, \hat{y_d}).$$

Where y_d and $\hat{y_d}$ are the truth values and predicted values for all drug cell lines pairs that include d , and are in the test set.

Similarly, drug stratified drug blind testing is defined by equation 2 but with the sum only running over drugs in the testing set, and N_d giving the number of drugs in the testing set. Furthermore, cell line stratified drug blind testing is defined by equation 1 but with the sum running over all cell lines, and where y_c, $\hat{y_c}$ are only the truth and predicted values for the drug cell line pairs in the test set that contain c.

A non-stratified M is simply given by

$$M = f(y, \hat{y}).$$

Where y and $\hat{y}$ are the truth values and predicted values for the cell line drug pairs in the test set respectively.

B. Metrics reported

When evaluating models that predicted continuous drug response values we calculate Mean squared error (MSE), Pearson correlation coefficient (Pear) and the Coefficient of determination (R2). We report Pear and R2 for both cell line (cl) stratification and drug stratification as well as non-stratification, as described above.

When evaluating models that predicted binary drug response values we calculate the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR).

C. Marker model hyperparameters

3-layers dense were used for the MLP with 4096 neurons per layer, with ReLU activation, followed by an output dense layer with one node. These hyperparameters were selected using the validation set.

D. Mixed set testing

This section shows the tables for mixed set testing for two additional train test splits.

Table 6. Metrics for mixed set testing for the second test train split.

Method	Drug branch	Omics branch	MSE	R2	Pear
tCNNS	SMILES	Genomics	1.25 ± 0.01	0.830 ± 0.001	0.9114 ± 0.0007
GraphDRP	Molecular graph	Genomics	1.08 ± 0.02	0.854 ± 0.003	0.9312 ± 0.0006
DeepTTA	SMILES	Transcriptomics	0.973 ± 0.009	0.868 ± 0.001	0.9317 ± 0.0008
Marker Model	Marker	Marker	0.890 ± 0.001	0.8789 ± 0.0002	0.9375 ± 0.0001

Table 7. Metrics for mixed set testing for the third test train split.

Method	Drug branch	Omics branch	MSE	R2	Pear
tCNNS	SMILES	Genomics	1.31 ± 0.01	0.819 ± 0.002	0.9060 ± 0.0007
GraphDRP	Molecular graph	Genomics	1.047 ± 0.009	0.856 ± 0.001	0.93 ± 0.0008
DeepTTA	SMILES	Transcriptomics	1.00 ± 0.01	0.863 ± 0.001	0.9290 ± 0.0007
Marker Model	Marker	Marker	0.904 ± 0.006	0.8757 ± 0.0009	0.9359 ± 0.0005

E. Cancer blind testing

This section shows the tables cancer blind testing for additional train test splits for. Table 8 and 9 show two additional train test splits for tCNNS, GraphDRP, DeepTTA-DB and DeepTTA.

Table 11 shows the results for tCNNS and GraphDRP but by replacing the genomics cell line profiles with transcriptomics profiles, (tCNNS_Tran an GraphDRP_Tran). It also shows these results for tCNNS_Tran-DB and GraphDRP_Tran-DB, removing the drug branch from the respective models. Where these tables show that adding transcriptomics cell line profiles causes the models to outperform the drug average benchmark. Furthermore, removing the drug features does not decrease performance.

Table 8. Metrics for cancer blind testing for the second test train split.

Method	Drug branch	Omics branch	MSE	Pear cl strat	R2 cl strat	Pear drug strat	R2 drug strat
tCNNS	SMILES	Genomics	2.4 ± 0.005	0.864 ± 0.002	0.620 ± 0.004	0.21 ± 0.04	-0.12 ± 0.03
GraphDRP	Molecular graph	Genomics	2.45 ± 0.05	0.867 ± 0.002	0.61 ± 0.01	0.24 ± 0.01	-0.19 ± 0.05
Drug Average	None	None	2.303	0.877	0.634	N/A	-0.010
DeepTTA-DB	Marker	Transcriptomics	**1.71 ± 0.02**	0.898 ± 0.001	**0.725 ± 0.004**	0.507 ± 0.007	0.23 ± 0.01
DeepTTA	SMILES	Transcriptomics	1.72 ± 0.02	**0.8991 ± 0.0008**	0.720 ± 0.003	**0.514 ± 0.006**	**0.236 ± 0.006**

Table 9. Metrics for cancer blind testing for the third test train split.

Method	Drug branch	Omics branch	MSE	Pear cl strat	R2 cl strat	Pear drug strat	R2 drug strat
tCNNS	SMILES	Genomics	2.7 ± 0.1	0.865 ± 0.003	0.57 ± 0.02	0.15 ± 0.02	-0.17 ± 0.03
GraphDRP	Molecular graph	Genomics	2.7 ± 0.2	0.872 ± 0.002	0.56 ± 0.03	0.22 ± 0.02	-0.17 ± 0.09
Drug Average	None	None	2.546	0.877	0.602	N/A	-0.018
DeepTTA-DB	Marker	Transcriptomics	**1.679 ± 0.009**	**0.8982 ± 0.0007**	**0.7303 ± 0.0009**	**0.573 ± 0.005**	**0.312** ± 0.004
DeepTTA	SMILES	Transcriptomics	1.70 ± 0.01	0.897 ± 0.001	0.724 ± 0.003	**0.573 ± 0.004**	0.305 ± 0.008

Table 10. Cancer blind testing with no stratification (apart from last col MSE drug strat, which is only marginally different from the non-stratfied metric due to a different number of cell lines being evaluated for each drug caused by missing truth values). For three train test splits

Train test split	Method	MSE	R2	Pear	MSE drug strat
Split 1	tCNNs	2.42 ± 0.04	0.660 ± 0.006	0.819 ± 0.004	2.44 ± 0.03
	GraphDRP	2.42 ± 0.08	0.66 ± 0.01	0.827 ± 0.003	2.43 ± 0.08
	Drug Average	2.191	0.692	0.832	2.189
	DeepTTA-DB	**1.57 ± 0.01**	**0.779 ± 0.002**	0.883 ± 0.001	**1.57 ± 0.01**
	DeepTTA	**1.57 ± 0.02**	**0.779 ± 0.003**	0.883 ± 0.002	**1.57 ± 0.02**
Split 2	tCNNs	2.412 ± 0.009	0.664 ± 0.001	0.818 ± 0.002	2.45 ± 0.03
	GraphDRP	2.47 ± 0.05	0.657 ± 0.007	0.821 ± 0.002	2.48 ± 0.04
	Drug Average	2.312	0.678	0.824	2.334
	DeepTTA-DB	**1.72 ± 0.02**	**0.760 ± 0.003**	0.873 ± 0.001	**1.73 ± 0.03**
	DeepTTA	1.73 ± 0.02	0.759 ± 0.002	**0.874 ± 0.001**	1.74 ± 0.02
Split 3	tCNNs	2.7 ± 0.1	0.64 ± 0.01	0.801 ± 0.006	2.69 ± 0.07
	GraphDRP	2.7 ± 0.1	0.64 ± 0.02	0.815 ± 0.003	2.7 ± 0.1
	Drug Average	2.523	0.658	0.812	2.505
	DeepTTA-DB	**1.674 ± 0.009**	**0.773 ± 0.001**	**0.880 ± 0.001**	**1.676 ± 0.008**
	DeepTTA	1.70 ± 0.01	0.770 ± 0.002	0.8790 ± 0.0008	1.70 ± 0.01

Table 11. Cancer blind cell line stratified metrics for tCNNS and GraphDRP but by replacing the genomics cell line profiles with transcriptomics profiles, (tCNNS_Tran an GraphDRP_Tran) and for CNNS_Tran-DB and GraphDRP_Tran-DB, removing the drug branch from the respective models

Train test split	Method	MSE	R2	Pear
Split 1	tCNNS	2.41 ± 0.04	0.61 ± 0.01	0.870 ± 0.003
	GraphDRP	2.41 ± 0.08	0.60 ± 0.01	0.872 ± 0.004
	Drug Average	2.190	0.640	0.884
	tCNNS_Tran	1.83 ± 0.05	0.698 ± 0.008	0.8865 ± 0.0005
	tCNNS_Tran-DB	1.78 ± 0.04	0.702 ± 0.007	0.8919 ± 0.0009
	GraphDRP_Tran	1.77 ± 0.02	0.714 ± 0.004	0.898 ± 0.002
	GraphDRP_Tran-DB	1.56 ± 0.01	0.743 ± 0.003	0.9062 ± 0.0008
Split 2	tCNNS	2.4 ± 0.005	0.620 ± 0.004	0.864 ± 0.002
	GraphDRP	2.45 ± 0.05	0.61 ± 0.01	0.867 ± 0.002
	Drug Average	2.303	0.634	0.877
	tCNNS_Tran	2.05 ± 0.05	0.67 ± 0.01	0.875 ± 0.003
	tCNNS_Tran-DB	1.94 ± 0.04	0.685 ± 0.007	0.884 ± 0.001
	GraphDRP_Tran	1.87 ± 0.06	0.70 ± 0.01	0.898 ± 0.001
	GraphDRP_Tran-DB	1.71 ± 0.03	0.723 ± 0.004	0.9001 ± 0.0007
Split 3	tCNNS	2.7 ± 0.1	0.57 ± 0.02	0.865 ± 0.003
	GraphDRP	2.7 ± 0.2	0.56 ± 0.03	0.872 ± 0.002
	Drug Average	2.546	0.602	0.877
	tCNNS_Tran	2.07 ± 0.06	0.67 ± 0.01	0.873 ± 0.002
	tCNNS_Tran-DB	2.00 ± 0.02	0.676 ± 0.005	0.883 ± 0.001
	GraphDRP_Tran	1.93 ± 0.07	0.69 ± 0.01	0.892 ± 0.003
	GraphDRP_Tran-DB	1.75 ± 0.01	0.718 ± 0.003	0.8968 ± 0.0006

F. Drug blind testing

This section shows the tables for non stratified drug blind testing for all three train test splits.

Table 12. Metrics for drug stratified drug blind testing for three train test splits. The metrics in **bold** are better than, or have bounds better than, the CL average benchmark. Or better than zero for the case of Pear where the CL average benchmark is undefined.

Train test split	Method	MSE	R2	Pear
Split 1	GraphDRP	10.0 ± 3	-4.0 ± 2	0.552 ± 0.007
	DeepTTA	7.2 ± 0.5	**-2.3 ± 0.3**	0.55 ± 0.03
	tCNNs	**7.0 ± 1**	**-2.3 ± 0.6**	0.54 ± 0.03
	CL Average	6.228	-2.154	0.612
Split 2	GraphDRP	6.0 ± 1	-2.4 ± 0.6	0.54 ± 0.01
	DeepTTA	5.4 ± 0.2	-2.4 ± 0.1	0.584 ± 0.009
	tCNNs	6.0 ± 1	**-2.6 ± 0.5**	0.57 ± 0.01
	CL Average	4.988	-2.180	0.627
Split 3	GraphDRP	10.5 ± 0.6	-4.9 ± 0.5	0.58 ± 0.04
	DeepTTA	**8.0 ± 1**	**-3.2 ± 0.5**	0.61 ± 0.01
	tCNNs	**6.3 ± 0.2**	**-1.9 ± 0.2**	0.621 ± 0.001
	CL Average	7.633	-3.279	0.659

Table 13. Metrics for cell line stratified drug blind testing for three train test splits. The metrics in **bold** are better than, or have bounds better than, the CL average benchmark.

Train test split	Method	MSE	R2	Pear
Split 1	GraphDRP	11.0 ± 3	**-0.7 ± 0.5**	-0.2 ± 0.2
	DeepTTA	7.5 ± 0.5	**-0.18 ± 0.09**	**0.1 ± 0.1**
	tCNNs	**7.0 ± 1**	**-0.1 ± 0.2**	**0.2 ± 0.2**
	CL Average	6.503	-0.023	N/A
Split 2	GraphDRP	**6.0 ± 1**	**-0.3 ± 0.3**	**0.1 ± 0.2**
	DeepTTA	5.6 ± 0.2	-0.19 ± 0.04	**0.30 ± 0.04**
	tCNNs	6.4 ± 0.8	-0.4 ± 0.2	**0.1 ± 0.1**
	CL Average	5.001	-0.057	N/A
Split 3	GraphDRP	11.2 ± 0.8	-0.4 ± 0.1	-0.1 ± 0.1
	DeepTTA	7.7 ± 0.6	**0.04 ± 0.08**	**0.407 ± 0.008**
	tCNNs	6.7 ± 0.2	**0.17 ± 0.03**	**0.47 ± 0.06**
	CL Average	8.125	-0.010	N/A