A cross-study analysis of drug response prediction in cancer cell lines

Fangfang Xia, Jonathan Allen, Prasanna Balaprakash, Thomas Brettin, Cristina Garcia-Cardona, Austin Clyde, Judith Cohn, James Doroshow, Xiaotian Duan, Veronika Dubinkina, Yvonne Evrard, Ya Ju Fan, Jason Gans, Stewart He, Pinyi Lu, Sergei Maslov, Alexander Partin, Maulik Shukla, Eric Stahlberg, Justin M. Wozniak, Hyunseung Yoo, George Zaki, Yitan Zhu and Rick Stevens

Corresponding author: Fangfang Xia, 9700 S Cass Ave, Lemont, IL 60439 USA. E-mail: fangfang@anl.gov

Abstract

To enable personalized cancer treatment, machine learning models have been developed to predict drug response as a function of tumor and drug features. However, most algorithm development efforts have relied on cross-validation within a single study to assess model accuracy. While an essential first step, cross-validation within a biological data set typically provides an overly optimistic estimate of the prediction performance on independent test sets. To provide a more rigorous assessment of model generalizability between different studies, we use machine learning to analyze five publicly available cell line-based data sets: National Cancer Institute 60, ancer Therapeutics Response Portal (CTRP), Genomics of Drug Sensitivity in Cancer, Cancer Cell Line Encyclopedia and Genentech Cell Line Screening Initiative (gCSI). Based on observed

Fangfang Xia is a Computer Scientist in Data Science and Learning Division at Argonne National Laboratory.

Jonathan Allen is a Bioinformatics Scientist at Lawrence Livermore National Laboratory.

Prasanna Balaprakash is a Computer Scientist in Mathematics and Computer Science Division at Argonne National Laboratory.

Thomas Brettin is a Strategic Program Manager in Computing, Environment and Life Sciences at Argonne National Laboratory.

Cristina Garcia-Cardona is a Scientist at Los Alamos National Laboratory.

Austin Clyde is a Computational Scientist at Argonne National Laboratory and a PhD student in Computer Science at University of Chicago.

Judith Cohn is a Research Scientist at Los Alamos National Laboratory.

James Doroshow is the Director of Division of Cancer Treatment and Diagnosis at National Cancer Institute.

Xiaotian Duan is a PhD student in Computer Science at University of Chicago.

Veronika Dubinkina is a PhD student in Department of Bioengineering at University of Illinois at Urbana-Champaign.

Yvonne Evrard is an Operations and Program Manager at Frederick National Laboratory for Cancer Research.

Ya Ju Fan is a Computational Scientist in Center for Applied Scientific Computing at Lawrence Livermore National Laboratory.

Jason Gans is a Technical Staff Member at Los Alamos National Laboratory.

Stewart He is a Data Scientist at Lawrence Livermore National Laboratory.

Pinyi Lu is a Bioinformatics Analyst at Frederick National Laboratory for Cancer Research.

Sergei Maslov is a Professor and Bliss Faculty Scholar in Department of Bioengineering and Physics at the University of Illinois at Urbana-Champaign and holds joint appointment at Argonne National Laboratory.

Alexander Partin is a Computational Scientist in Data Science and Learning Division at Argonne National Laboratory.

Maulik Shukla is a Project Lead and Computer Scientist in Data Science and Learning Division at Argonne National Laboratory.

Eric Stahlberg is the Director of Biomedical Informatics and Data Science at the Frederick National Laboratory for Cancer Research.

Justin M. Wozniak is a Computer Scientist in Data Science and Learning Division at Argonne National Laboratory.

Hyunseung Yoo is a Software Engineer in Data Science and Learning Division at Argonne National Laboratory.

George Zaki is a Bioinformatics Manager at Frederick National Laboratory for Cancer Research.

Yitan Zhu a Computational Scientist in Data Science and Learning Division at Argonne National Laboratory.

Rick Stevens is an Associate Laboratory Director at Argonne National Laboratory and a Professor in Computer Science at University of Chicago.

Submitted: 19 April 2021; **Received (in revised form):** 26 July 2021

experimental variability across studies, we explore estimates of prediction upper bounds. We report performance results of a variety of machine learning models, with a multitasking deep neural network achieving the best cross-study generalizability. By multiple measures, models trained on CTRP yield the most accurate predictions on the remaining testing data, and gCSI is the most predictable among the cell line data sets included in this study. With these experiments and further simulations on partial data, two lessons emerge: (1) differences in viability assays can limit model generalizability across studies and (2) drug diversity, more than tumor diversity, is crucial for raising model generalizability in preclinical screening.

Key words: drug response prediction; deep learning; drug sensitivity; precision oncology

Introduction

Precision oncology aims at delivering treatments tailored to the specific characteristics of a patient's tumor. This goal is premised on the idea that, with more data and better computational models, we will be able to predict drug response with increasing accuracy. Indeed, two recent trends support this premise. First, high-throughput technologies have dramatically increased the amount of pharmacogenomic data. A multitude of omic types such as gene expression and mutation profiles can now be examined for potential drug response predictors. On the prediction target side, while clinical outcomes are still limited, screening results abound in preclinical models that mimic patient drug response with varying fidelity. Second, deep learning has emerged as a natural technique to capitalize on the data. Compared with traditional statistical models, the high capacity of neural networks enable them to better capture the complex interactions among molecular and drug features.

The confluence of these factors has ushered in a new generation of computational models for predicting drug response. In the following, we provide a brief summary of recent work in this field, with a focus on representative studies using cancer cell line data.

The National Cancer Institute 60 (NCI60) human tumor cell line database [1] was one of the earliest resources for anti-cancer drug screen. Its rich molecular characterization data have allowed researchers to compare the predictive power of different assays. For example, gene expression, protein and microRNA abundance have been shown to be effective feature types for predicting both single and paired drug response [2, 3]. In the last decade, new resources including Cancer Cell Line Encyclopedia (CCLE) [4, 5], Genomics of Drug Sensitivity in Cancer (GDSC) [6] and Cancer Therapeutics Response Portal (CTRP) [7, 8] have significantly expanded the number of screened cancer cell lines. Drug response prediction has moved beyond per-drug or per-cell line analysis [9, 10] to include integrative models that take both drug and cell line as input features. Community challenges have further inspired computational approaches [11, 12]. A wide range of new machine learning techniques have been explored, including recommendation systems [13], ranking methods [14, 15], generative models [10, 16], feature analysis [17], network modeling [18], ensemble models [19–23] and deep learning approaches [24–28], with some incorporating novel design ideas such as attention [29] and visual representation of genomic features [30]. A number of excellent review articles have recently been published on the topic of drug response prediction, with substantial overlap and special emphases on data integration [31], feature selection [32], experimental comparison [33], machine learning methods [34], systematic benchmarking [35], combination therapy [36], deep learning results [37] and meta-review [38].

Despite the tremendous progress in drug response prediction, significant challenges remain: (1) inconsistencies across studies in genomic and response profiling have long been documented [39–42]. The Genentech Cell Line Screening Initiative (gCSI) [43] was specifically designed to investigate the discordance between CCLE and GDSC by an independent party. While harmonization practices help [44–46], a significant level of data variability will likely remain in the foreseeable future due to the complexity in cell subtypes and experimental standardization. (2) The cross-study data inconsistencies suggest that a predictive model trained on one data source may not perform as well on another. Yet, most algorithm development efforts have relied on cross-validations within a single study, which likely overestimate the prediction performance. Even within a single study, validation R^2 (explained variance) rarely exceeds 0.8 when strict data partition is used [37], indicating difficulty in model generalization. (3) Without a single study that is sufficiently large, a natural next step is to pool multiple data sets to learn a joint model. Multiple efforts have started in this direction, although the gains to date from transfer learning or combined learning have been modest [47–49]. (4) It is also unclear how model generalization improves with increasing amounts of cell line or drug data. This information will be essential for future studies to prioritize screening experiments.

In this study, we seek to provide a rigorous assessment of the performance range for drug response prediction given the current publicly available data sets. Focusing on five cell line screening studies, we first estimate within- and cross-study prediction upper bounds based on observed response variability. After integrating different drug response metrics, we then apply machine learning to quantify how models trained on a given data source generalize to others. These experiments provide insights on the data sets that are more predictable or contain more predictive value. To understand the value of new experimental data, we further use simulations to study the relative importance of cell line versus drug diversity and how their addition may impact model generalizability.

Data integration

Pan-cancer screening studies have been comprehensively reviewed by Baptista *et al.* [37]. In this study, we focus on single-drug response prediction and include five public data sets: NCI60, CTRP, GDSC, CCLE and gCSI. The characteristics of the drug response portion of these studies are summarized in Table 1. These data sets are among the largest in response sample size of the nine data sources reviewed and, therefore, have been frequently used by machine learning studies. Together, they also capture a good representation of the diversity in viability assay for response profiling.

Data acquisition and selection

Of the five drug response data sets included in this study, NCI60 was downloaded directly from the NCI FTP and the remaining four were from PharmacoDB [44]. We use GDSC and CTRP to

Table 1. Characteristics of drug response data sets included in the cross-study analysis

Data set	Cells	Drugs	Dose response samples	Drug response groups[a]	Viability assay
NCI60	60	52,671	18 862 308	3,780,148	Sulforhodamine B stain
CTRP	887	544	6171 005	395 263	CellTiter Glo
CCLE	504	24	93 251	11 670	CellTiter Glo
GDSC	1075	249	1894 212	225 480	Syto60
gCSI	409	16	58 094	6455	CellTiter Glo

[a] We define a response group as the set of dose–response samples corresponding to a particular pair of drug and cell line.

denote the GDSC1000 and CTRPv2 data collections in PharmacoDB, respectively.

The different experimental designs of these studies resulted in differential coverage of cell lines and drugs. While NCI60 screened the largest compound library on a limited number of cell lines, CCLE and gCSI focused on a small library of established drugs with wider cell line panels. The remaining two studies, CTRP and GDSC, had more even coverage of hundreds of cell lines and drugs.

There is considerable overlap in cell lines and drugs covered by the five data sets. A naive partitioning of the samples by identifier would thus lead to leakage of training data into the test set by different aliases. Thanks to PharmacoDB's curation and integration effort, such alias relationships can be clearly identified through name mapping.

To create a working collection of samples, we first defined a cell line set and a drug set. Our cell line selection is the union of all cell lines covered by the five studies. For drugs, however, NCI60 screened a collection of over 50 000 drugs that would be too computationally expensive for iterative analysis. Therefore, we selected a subset of 1006 drugs from NCI60, favoring Food and Drugs Administration-approved drugs and those representative of diverse drug classes. From the remaining four studies, we included all the drugs. This resulted in a combined set of 1801 drugs.

Drug response integration

For cell line-based anticancer screening, drug sensitivity is quantified by dose–response values that measure the ratio of surviving treated to untreated cells after exposure to a drug treatment at a given concentration. The dose–response values can be further summarized, for a particular cell–drug pair, to form dose-independent metrics of drug response.

All five studies provided dose–response data, but with different ranges. To facilitate comparison across studies, dose–response values were linearly rescaled to share a common range from −100 to 100.

Dose-independent metrics were more complicated. For example, whereas GDSC used IC_{50} (dose of 50% inhibition of cell viability), NCI60 used GI_{50} (dose at which 50% of the maximum growth inhibition is observed). Given the lack of consensus, we adopted PharmacoDB's practice to recompute the aggregated dose-independent metrics for each study from the individual dose–response values, removing inconsistencies in calculation and curve fitting [44].

Among these summary statistics, AAC is the area above the dose–response curve for the tested drug concentrations, and DSS (drug sensitivity score) is a more robust metric normalized against dose range [50]. The definitions of these two metrics were directly borrowed from PharmacoDB, but they had to be recomputed in this study because PharmacoDB does not cover NCI60. We used the same Hill Slope model with three bounded parameters [44] to fit the dose-response curves. We also added another dose-independent metric, area under drug response curve (AUC), to capture the area under the curve for a fixed concentration range ($[10^{-10}, 10^{-4}]\mu M$). AUC can be viewed as the average growth and can be compared across studies. See the Methods section for the detailed definitions of these metrics.

Results

Response prediction upper bounds

Data sets from different studies often exhibit biases. Apart from differences in study design (e.g. choice of cell lines and drugs, sample size), there are two main types of biases associated with experimentally measured data: bias in molecular characterization and bias in response assay. Viewed in the context of a machine learning problem, these correspond to biases in input features and output labels. They form the main challenge to joint learning and transfer learning efforts over multiple data sources.

The first type of bias can be alleviated with feature preprocessing. Specifically, we used batch correction to homogenize gene expression profiles from different databases and followed a consistent protocol to generate drug features and fingerprints (see our previous work [49] and a summary in the Methods section).

The 2nd type of bias is the focus of this paper. That is, despite the integration efforts, significant differences in measured drug response remain. Figure 1 illustrates this heterogeneity with a common example: the LOXIMVI melanoma cell line treated with Paclitaxel. This same combination is included by multiple studies, but it exhibits distinct response curves across replicates as well as studies. Given the degree of variability in drug response measurements, it is natural to ask what's the best prediction models could do with infinite data. We explored this question based on within- and cross-study replicates.

Within-study response variability among replicates

Three of the studies (NCI60, CTRP, GDSC) contained replicate samples where the same pair of drug and cell line were assayed multiple times. These data allowed us to estimate within-study response variability as summarized in Table 2. The fact that these studies used three different viability assays was also an opportunity to sample the associations between variability level and assay technology.

Overall, NCI60 had the lowest SD in dose response at 14.5 for an average replicate group after rescaling response values linearly in all three studies to a common range. To establish a ceiling for the performance of machine learning models, we computed how well the mean dose response of a replicate group
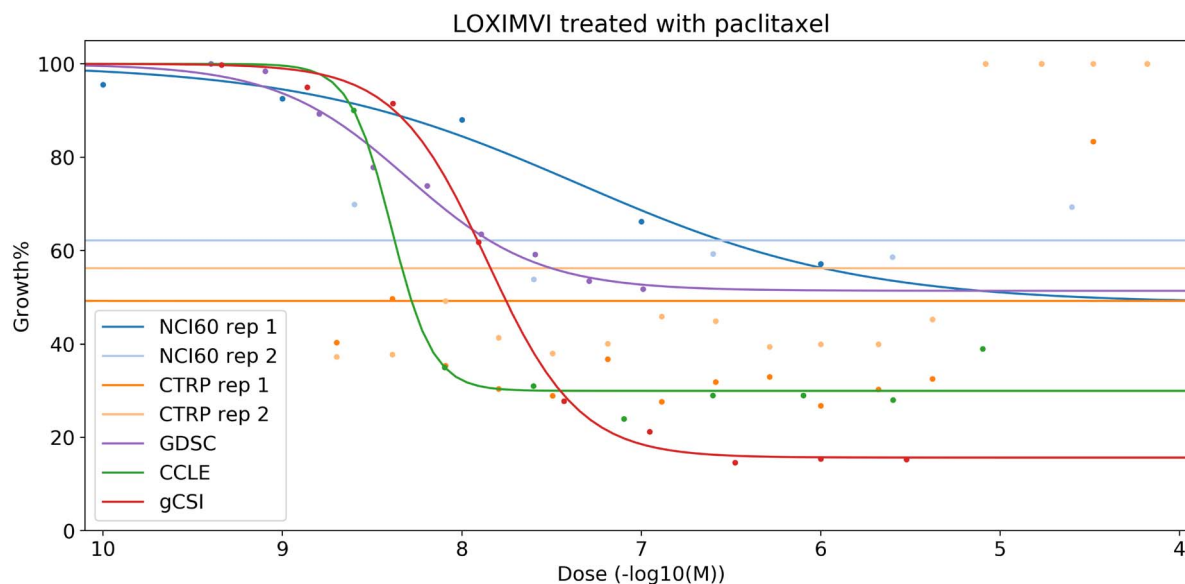
Figure 1. Fitted dose–response curves from multiple studies. This example shows the dose response of the LOXIMVI melanoma cell line treated with paclitaxel. Curves have been consistently fitted across the studies. Experimental measurements from multiple sources and replicates are not in complete agreement.

Table 2. Dose–response variability among replicates in the same study

Study	Samples with replicates	Replicates per group	Mean response SD in group	R^2 Explaining response with group mean	R^2 for Samples with Replicates
NCI60	41.56%	2.62	14.5	0.959	0.931
CTRP	4.09%	2.05	18.8	0.996	0.862
GDSC	2.62%	2.00	21.9	0.996	0.810

would predict individual response values. This resulted in an R^2 score of 0.959 for NCI60. The equivalent values for CTRP and GDSC were much higher, but only for the fact that they had much lower fractions of samples with replicates. When we confined the analysis to just nonsingleton replicate groups, the R^2 scores dropped significantly with GDSC being the lowest at 0.81. While it's unclear how well these replicate groups represent the entire study, we should expect machine learning models to not exceed the R^2 score in cross-validation. This is not a statement on the capability of the machine learning methods, but the noise level in the ground truth data that they work with.

Cross-study response variability

For comparing drug response across different studies, dose-dependent metrics are less useful, because each study generally has its own experimental concentrations. Instead, we opted for the three aforementioned dose-independent response metrics: AUC, AAC and DSS. A key difference between AUC and AAC is that AUC is based on a fixed dose range, whereas AAC is calculated for the drug-specific dose window screened in a particular study. DSS is another metric that integrates both drug efficacy and potency, with evidence of more differentiating power in understanding drug functional groups [50].

Using these three metrics, we analyzed cross-study response variability by focusing on the common subset of cell line and drug pairs that appear in multiple studies.

Table 3 summarizes two representative cases. In the 1st scenario, we used response values from CTRP to directly predict the corresponding metrics in CCLE. In this case, the source and target studies shared a common viability assay. Yet, the raw R^2 on all three metrics were only around 0.6, with AAC being the highest. In the 2nd scenario, where the target study GDSC used a different viability assay, the response projections were markedly worse: the best R^2 score was slightly above 0.3 achieved on AUC, and the other two scores were close to zero.

In the above analysis, the response values from one source were straightly used as the predicted response in another source. This is applicable to cases where the response information in the target study is unknown. When the target distribution is known, a mapping function can be used to fit the response values, as a way to bridge the measurement differences between the studies. Figure 2 illustrates this with simple linear regression. This improved the best R^2 scores (see the *Fit* columns in Table 3) to 0.68 for the mapping from CTRP to CCLE (same assay) and 0.41 going from CTRP to GDSC (different assays). Again, these numbers only gave rough estimates based on partial data; they nevertheless calibrated what we could expect from machine learning models performing cross-study prediction.

So far we have shown that there is greater variability between CTRP and GDSC than between CTRP and CCLE, but it's unclear whether this can be attributed to the difference in response assay alone. Another difference in drug diversity between the target studies is at play, with GDSC having far more drugs (249)

Table 3. Dose-independent response variability in identical cell–drug pairs across studies

Source study	Target study	Overlapping cell–drug groups	R^2 on AUC[a]		R^2 on AAC		R^2 on DSS[b]	
			Raw[c]	Fit[d]	Raw	Fit	Raw	Fit
CTRP	CCLE	2339	0.594	0.683	0.641	0.681	0.635	0.670
CTRP	GDSC	17 259	0.302	0.409	0.019	0.197	0.006	0.194

[a]AUC is not a direct complement of AAC. They are defined on different concentration ranges: AUC for a fixed range ($[10^{-10}, 10^{-4}]\mu M$) and AAC for study-specific, tested concentration ranges.[b] DSS is a normalized variation of AAC defined by Yadav *et al.* [50].[c] Raw R^2 scores (explained variance) were derived by directly comparing the corresponding drug response values, for a given metric, on the common experiments with cell-drug pairs shared by the two studies (see scatter plots in Figure 2).[d] To account for study bias in experimental response measurement, linear regression was used to fit the response values on the same cell–drug pair between source and target studies (see regression lines in Figure 2).
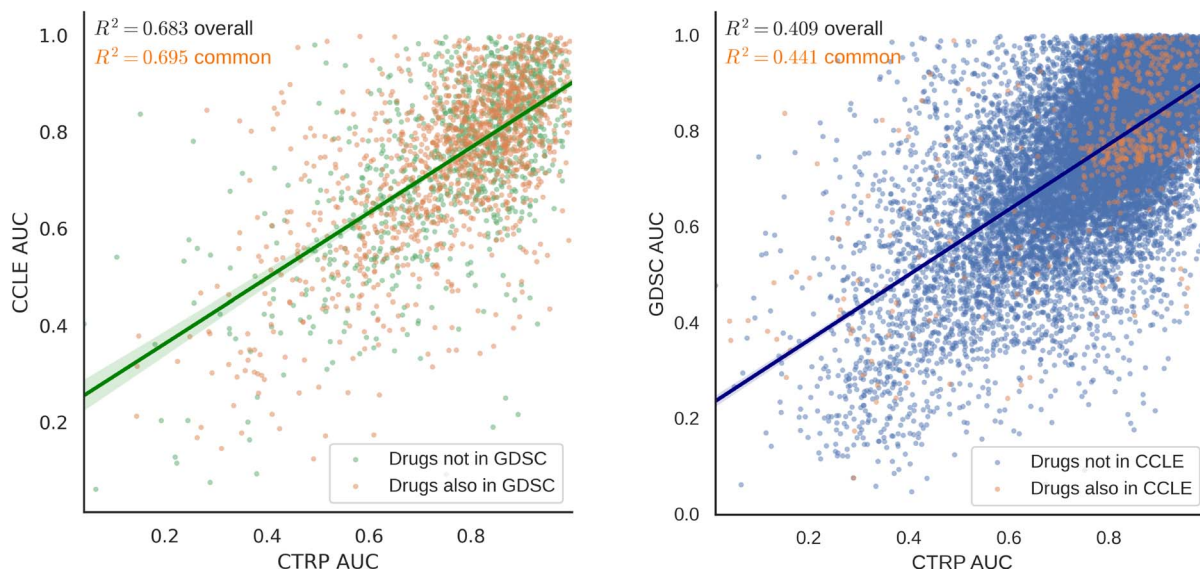
Figure 2. Estimating cross-study response variability based on overlapping experimental data. When the same combination of drug and cell line appears in multiple studies, we can use the reported differences to estimate cross-study response variability. Here we map the AUC values from CTRP to CCLE (left) and CTRP (right) in the scatter plots with linear regression fit. Orange dots represent experiments involving common drugs shared by CTRP, CCLE and GDSC, reducing sampling bias in different studies. In both plots, R^2 scores are reported separately for the overall fit and the subset with common drugs. Overall, there is greater agreement between CTRP and CCLE (same assay) than between CTRP and GDSC (different assay).

than CCLE (24). To tease apart these two factors, we narrowed down to a subset of seven drugs shared by all three studies: Paclitaxel, Lapatinib, AZD0530, AZD6244, Nilotinib, Erlotinib and Sorafenib. Scatter plots of response data points involving this common drug set are highlighted in orange in Figure 2. Linear regression fit on the subset resulted in slightly higher R^2 scores: 0.69 between CTRP and CCLE and 0.44 between CTRP and GDSC, both on the dose-independent AUC metric. Their unchanging relationship suggests that the observed cross-study response variability is more predominantly due to viability assay than sampling difference.

This comparison between CellTiter Glo (CTRP, CCLE) and Syto60 (GDSC) does not necessarily generalize to other pairs of viability assays. NCI60 is another study that used a different viability assay. When we swapped GDSC for NCI60 in the above analysis, the overall R^2 score from CTRP to NCI60 is 0.67, only slightly lower than that to CCLE. The R^2 score on the common set among CTRP, NCI60 and CCLE is lower at 0.60, suggesting that sampled drug diversity also plays a role in the estimate of cross-study response variability.

Cross-study prediction of drug response

We applied machine learning models to predict dose response. When evaluated in the context of a single data set, the prediction performance is dependent on two general factors: the input features and the model complexity. The models we reviewed in the introduction section have offered wide-ranging configurations on these two factors. Most of these models, however, stopped at evaluation by cross-validation within a single study. Given the cross-study variability observed, such evaluation likely overestimated the utility of drug response models for practical scenarios involving more than one data sets.

To test how well models trained on one data set could generalize to another, we went beyond study boundaries and tested all combinations of source and target sets. This introduced a 3rd factor impacting model performance, i.e. the distribution shift between the source and target data sets. To provide a rigorous assessment of cross-study prediction performance, we applied three machine learning models of increasing complexity. The 1st two, Random Forest [51] and LightGBM [52], were used as baseline models. The 3rd one, designed in-house, represented our best effort at a deep learning solution.

Baseline machine learning performance

The first baseline performance using Random Forest is shown in Table 4. Each cell in the matrix represents a learning and evaluation experiment involving a source study for training and a target study for testing. The prediction accuracy for dose response was

assessed using both R^2 score and mean absolute error (MAE). Again, the range for drug response values was unified across studies to be from −100 to 100.

As expected, the diagonal cells have relatively high scores for they represent cross validation done on the same data set. We are more interested in the nondiagonal values because they are less likely to be an artifact of overfitting. The nondiagonal cells of the matrix are color coded as follows: green for $R^2 > 0.1$, red for $R^2 < −0.1$ and yellow otherwise. As we mentioned in the data setup, since each cross-study validation experiment involved training multiple models and filling out a matrix of inference results, we limited ourselves to a subset of drugs from NCI60. For the remaining five studies, all cell lines and drugs were included. As for input features, we used cell line gene expression profiles, drug chemical descriptors and molecular fingerprints. For details on feature processing, see the Methods section.

Random Forest models trained on CTRP, NCI60 and GDSC achieved moderate cross-study prediction accuracy. CTRP-trained models performed the best, scoring the highest cross-study R^2 of 0.45 when CCLE was used as the testing set. CCLE-trained models had less generalization value and the ones trained on gCSI did not generalize at all. This was not surprising as gCSI had the smallest number of drugs and thus prone to overfitting.

LightGBM is a gradient boosting version of tree-based learning algorithm and is generally considered superior to Random Forest. Here, the LightGBM models outperformed Random Forest for most of the cells (Table 5). However, in the Random Forest matrix, the diagonal values were generally comparable to other cells, suggesting there was little overfitting (with the exception the gCSI row). In contrast, each diagonal cell in the LightGBM matrix was better than other cells in their row or column. This was consistent with the view that cross-validation within a study was an easier task than cross-study generalization. Overall, the average improvement in R^2 for corresponding cells between Random Forest and LightGBM models was 0.35 for the diagonal values and 0.19 for the cross-study comparisons.

Deep learning performance

Deep learning models generally performed on par or slightly better than LightGBM. We experimented with a variety of neural network model architectures, and our best prediction accuracy was achieved by a multitasking model that we called UnoMT. Figure 4 shows an example configuration of the UnoMT model where, in addition to the central task of predicting drug response, the model also tried to perform a variety of auxiliary classification and regression tasks (see Methods for details). These multitasking options allowed the model to use additional labeled data to improve its understanding of cell line and drug properties.

The best performance achieved by deep learning is shown in Table 6 with three additional prediction tasks turned on: cell line category (normal versus tumor), site and type. On average, the cross-study R^2 improved 0.11 over LightGBM models, and the model did nearly perfectly on the additional tasks such as cancer type prediction (not shown). On models trained on CCLE data, deep learning offered the most pronounced improvement. While the within-study R^2 was nearly identical to that of LightGBM, the models were able to, for the first time, generalize to NCI60, CTRP and GDSC to a moderate degree. The improved cells in Table 6 compared with both Tables 4 and 5 are highlighted in bold. Stacking LightGBM and deep learning models resulted in marginal improvement in cross-validation accuracy but did not improve cross-study generalizability.

Predictive and predictable cell line data sets

Used as training sets, each of the five cell line studies was ranked by how predictive they were of the remaining four cell line studies. Used as testing sets, each of the five cell lines studies were ranked by how predictable they were, using the other four cell line studies as training sets. By multiple measures (average, median and minimum R^2), machine learning algorithms trained on CTRP yield the most accurate predictions on the remaining testing data sets. This ranking was consistent across the results from all three machine learning models. By average and median R^2 on deep learning results, the gCSI data set was the most predictable cell line dataset, and CCLE was a close second.

How model generalizability improves with more data

Experimental screening data are expensive to acquire. It is thus critical to understand, from a machine learning perspective, how additional cell line or drug data impact model generalizability. Data scaling within a single study have been previous explored [53]. In our cross-study machine learning predictions, we observed that the models with poor generalizability tended to have been trained on a small number of drugs (CCLE and gCSI). To study the relative importance of cell line versus drug diversity, we conducted simulations on CTRP, the most predictive dataset of the five. We varied the fraction of randomly selected cell lines and kept all the drugs, and vice versa. The results are plotted in Figure 3.

Models trained on a small fraction of cell lines but all drugs could still quickly approach the full model performance, whereas models trained on limited drug data suffered from low accuracy and high uncertainty. In either case, it was far more difficult to predict response in a target dataset using a different viability assay (GDSC) than one with the same assay (CCLE). Inferred upper bounds (dotted lines) were loosely extrapolated from direct mapping results based on data intersection using the best dose-independent metric from Table 3.

When the numbers of samples were fixed, the model performance was not particularly sensitive to feature selection. Our previous work showed that deep learning models could still train well with a high dropout ratio of 0.45, suggesting redundancy in both cell line and drug features [49]. Here we also performed data-driven feature selection. Models trained with the top 1000 cell line features and top 1000 drug features, identified based on feature importance scores on a 1st round of training with LightGBM, had no loss in generalizability. Models trained with the top 100 cell line features and top 100 drug features reached 90% peak generalizability from CTRP to CCLE and 85% generalizability from CTRP to GDSC.

Methods

Feature selection and processing

Drug response is modeled as a function of cell line features and drug properties. We used three basic types of features in this study: RNAseq gene expression profiles, drug descriptors and drug fingerprints. Drug concentration information was also needed for dose–response prediction. The gene expression values were the batch-corrected RNA-seq measurements from the NCI60, GDSC and CCLE. Because CTRP and gCSI data sets did not provide RNA-seq data, these data sets were assigned gene expression values from CCLE based on cell line name matching. CCLE has molecular characterization data for more than 1000 cell lines [5], which cover the CTRP and gCSI cell line sets. Log

**Table 4.** Baseline cross-study validation results with Random Forest

Training Set	Testing Set				
	NCI60	CTRP	GDSC	CCLE	gCSI
NCI60	$R^2 = 0.45$ MAE = 30.4	$R^2 = 0.23$ MAE = 34.6	$R^2 = 0.15$ MAE = 37.3	$R^2 = 0.29$ MAE = 34.3	$R^2 = 0.14$ MAE = 54.0
CTRP	$R^2 = 0.41$ MAE = 31.7	$R^2 = 0.30$ MAE = 35.0	$R^2 = 0.15$ MAE = 37.4	$R^2 = 0.45$ MAE = 29.0	$R^2 = 0.17$ MAE = 39.6
GDSC	$R^2 = 0.33$ MAE = 36.0	$R^2 = 0.14$ MAE = 41.5	$R^2 = 0.13$ MAE = 40.4	$R^2 = 0.17$ MAE = 42.4	$R^2 = 0.08$ MAE = 43.0
CCLE	$R^2 = 0.12$ MAE = 42.6	$R^2 = -0.03$ MAE = 48.9	$R^2 = -0.11$ MAE = 47.1	$R^2 = 0.17$ MAE = 42.4	$R^2 = 0.32$ MAE = 38.5
gCSI	$R^2 = -0.38$ MAE = 55.0	$R^2 = -0.51$ MAE = 59.0	$R^2 = -0.59$ MAE = 58.7	$R^2 = -0.09$ MAE = 48.6	$R^2 = 0.25$ MAE = 39.9

Table 5. Baseline cross-study validation results with Random Forest

Training Set	Testing Set				
	NCI60	CTRP	GDSC	CCLE	gCSI
NCI60	$R^2 = 0.74$ MAE = 21.1	$R^2 = 0.32$ MAE = 35.7	$R^2 = 0.20$ MAE = 36.6	$R^2 = 0.40$ MAE = 35.0	$R^2 = 0.42$ MAE = 34.6
CTRP	$R^2 = 0.42$ MAE = 31.1	$R^2 = 0.67$ MAE = 23.3	$R^2 = 0.25$ MAE = 35.0	$R^2 = 0.56$ MAE = 30.0	$R^2 = 0.54$ MAE = 30.8
GDSC	$R^2 = 0.30$ MAE = 35.5	$R^2 = 0.19$ MAE = 38.0	$R^2 = 0.55$ MAE = 27.3	$R^2 = 0.41$ MAE = 36.0	$R^2 = 0.58$ MAE = 29.3
CCLE	$R^2 = 0.05$ MAE = 43.2	$R^2 = -0.01$ MAE = 44.5	$R^2 = -0.03$ MAE = 43.3	$R^2 = 0.60$ MAE = 28.7	$R^2 = 0.26$ MAE = 38.5
gCSI	$R^2 = 0.19$ MAE = 41.1	$R^2 = -0.02$ MAE = 45.2	$R^2 = -0.03$ MAE = 45.1	$R^2 = 0.22$ MAE = 41.8	$R^2 = 0.50$ MAE = 36.7

Table 6. Baseline cross-study validation results with Random Forest

Training Set	Testing Set				
	NCI60	CTRP	GDSC	CCLE	gCSI
NCI60	$R^2 = 0.81$ MAE = 17.1	$\mathbf{R^2 = 0.38}$ MAE = 32.2	$\mathbf{R^2 = 0.24}$ MAE = 35.3	$\mathbf{R^2 = 0.48}$ MAE = 33.4	$\mathbf{R^2 = 0.46}$ MAE = 33.4
CTRP	$\mathbf{R^2 = 0.44}$ MAE = 29.8	$R^2 = 0.68$ MAE = 22.7	$R^2 = 0.23$ MAE = 34.4	$\mathbf{R^2 = 0.61}$ MAE = 28.3	$\mathbf{R^2 = 0.60}$ MAE = 28.5
GDSC	$\mathbf{R^2 = 0.32}$ MAE = 34.0	$\mathbf{R^2 = 0.25}$ MAE = 36.7	$R^2 = 0.53$ MAE = 27.2	$\mathbf{R^2 = 0.50}$ MAE = 32.6	$\mathbf{R^2 = 0.60}$ MAE = 29.2
CCLE	$\mathbf{R^2 = 0.27}$ MAE = 36.9	$\mathbf{R^2 = 0.20}$ MAE = 39.2	$\mathbf{R^2 = 0.11}$ MAE = 38.9	$R^2 = 0.68$ MAE = 25.4	$\mathbf{R^2 = 0.39}$ MAE = 34.2
gCSI	$R^2 = 0.00$ MAE = 44.9	$\mathbf{R^2 = 0.11}$ MAE = 43.1	$R^2 = 0.05$ MAE = 42.8	$\mathbf{R^2 = 0.33}$ MAE = 40.6	$R^2 = 0.80$ MAE = 19.2
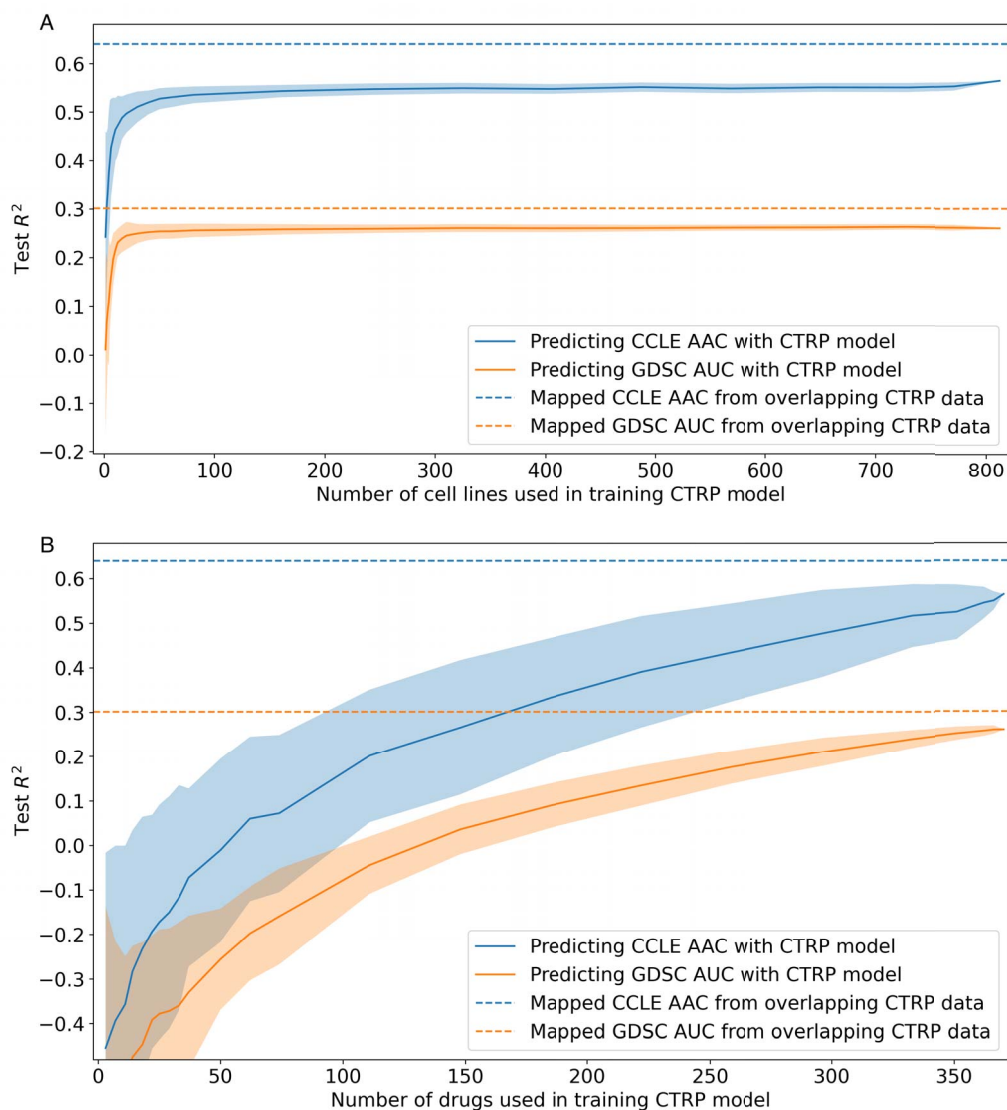
Figure 3. Impact of cell line or drug diversity on model generalizability. Models trained on partial CTRP data are tested on CCLE and GDSC. Shades indicate the SD of the cross-study R^2. (**A**). Models trained with all CTRP drugs and a fraction of cell lines. (**B**). Models trained with all CTRP cell lines and a fraction of drugs.

transformation was used to process FPKM-UQ normalized gene expression profiles [54]. Batch correction was performed using a simple linear transformation of the gene expression values such that the per-dataset mean and variance was the same for all data sets [49]. To reduce the training time, only the expression values of the LINCS1000 (Library of Integrated Network-based Cellular Signatures) landmark genes [55] were used as features. Our previous work on systematic featurization [49] found no superset of LINCS1000 (including the full set of over 17 000 genes or known oncogene sets) that clearly outperformed LINCS1000. Drug features included 3838 molecular descriptors and 1024 path fingerprints generated by the Dragon software package (version 7.0) [56]. Drug features/fingerprints were computed from 2D molecular structures downloaded from PubChem [57] and protonated using OpenBabel [58].

Drug response metrics

Data-driven prediction for drug response is based on the hypothesis that the interaction between tumor and drug treatment can be modeled as a function of three factors: the cancer genomic system, the compound chemical structure and the drug concentration. When the 1st two factors are given, it is further assumed that the rate-limiting step in the killing of cancer cells is the binding of the drug to a target receptor in the cells.

Intuitively, a higher concentration x would lead to a larger fraction of receptors bound to the drug molecules or a lower fraction y of surviving cancer cells. At the same time, a fraction E_∞ of the cancer cells are not susceptible to the drug regardless of the drug dose. This dose–response relationship can be worked out based on thermodynamic equilibrium of the bound drug–target complex, and we adopted the three-parameter Hill Slope equation below used in PharmacoDB [44].

$$y(x) = E_\infty + \frac{1 - E_\infty}{1 + \left(\frac{x}{EC_{50}}\right)^{HS}} \qquad (1)$$
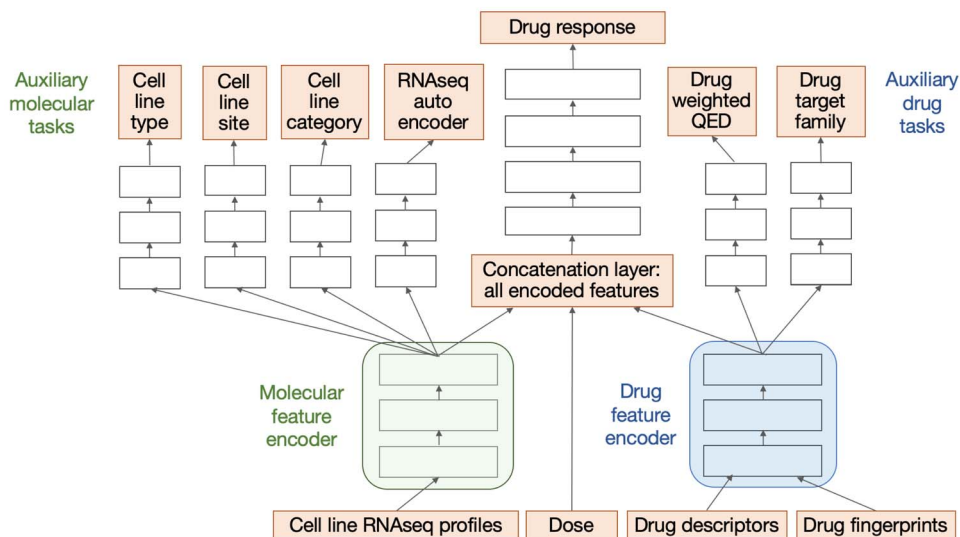
Figure 4. An example configuration of the multitasking drug response prediction network (UnoMT). The network predicts a number of cell line properties (tissue category, tumor site, cancer type, gene expression autoencoder) as well as drug properties (target family, drug-likeness score) in addition to drug response.

EC_{50} is the drug dose at which half of the target receptors are bound (half-maximal response). Both EC_{50} and E_∞ depend on cancer cell and drug properties. The Hill Slope coefficient *HS* quantifies the degree of interaction between binding sites.

With this model, we consistently fit the dose–response data from all the studies. Examples of the resulting dose–response curves are shown in Figure 1. This enabled the derivation of three dose-independent response metrics for comparing across experiments that used different dose levels:

- AUC:for $[10^{-10}, 10^{-4}]\mu M$, a fixed dose range.
- AAC: for the measured dose range in a study (same definition in PharmacoDB).
- DSS:same to DSS1 in PharmacoDB.

DSS is similar to AAC but is potentially more robust as it tries to calibrate the intended dose range of the drug. Specifically, it aggregates the response over the range where the drug response *R* exceeds an activity threshold A_{min} [50]:

$$DSS \sim \int_{R > A_{min}} R(x)dx \qquad (2)$$

Evaluation metrics

In this study, we chose two commonly used metrics to evaluate the performance of machine learning models. MAE measures the mean absolute difference between the observed responses y_i and the predicted responses $\hat{y}_i$. R^2, also known as coefficient of determination, measures the explained variance as a proportion of the total variance.

$$MAE = \sum_i |y_i - \hat{y}_i| \qquad (3)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \qquad (4)$$

R^2 was additionally used to evaluate the response variability levels within and across studies. This was done by examining the observed response values on the overlapping drug and cell

line groups from different experimental replicates or studies. For example, a high R^2 score would suggest good agreement among the multiple measurements and thus low variability in experimental data.

Machine learning

Three different machine learning algorithms were evaluated in this study: Random Forest, LightGBM and deep neural networks. The cell line prediction error for all methods was assessed using the MAE and Scikit-learn [59] definition of R^2 value, which is equal to the fractional improvement in mean squared error (MSE) of the method compared with the MSE of the prediction method that outputs the average response of the test set, independent of dose, gene expression and drug features. In the diagonal cells in the matrices (Tables 4–6), mean values of 5-fold cross-validation partitioned on cell lines are reported. For cross-study analysis, the entire selected set of source study data were used to train the model, and the nondiagonal cells report test metrics on the whole-target data set. The Random Forest models were trained using the default Scikit-learn implementation (version 0.22). The LightGBM models were trained using the Scikit-learn implementation with the number of trees set to be proportional to the number of drugs included in the training set.

Deep learning

The reported deep neural network is based on a unified model architecture, termed Uno, for single and paired drug treatment. This architecture was extended from a previously developed neural network called 'Combo' [3] to simultaneously optimize for feature encoding and response prediction. Cell line and drug features first pass through their separate feature encoder subnetworks before being concatenated with drug concentration to predict dose response through a 2nd level of subnetwork. In the multitasking configuration of the network (UnoMT), the output layers of molecular and drug feature encoders are further connected with additional subnetworks to predict auxiliary properties. All subnetworks have residual connections between

neighboring layers [60]. The multitasking targets include drug-likeness score (QED [61]) and target family for drugs, and tissue category (normal versus tumor), type, site and gene expression profile reconstruction for molecular samples. Not all labels were available for all samples. In particular, tissue category and site applied only to the Cancer Genome Atlas (TCGA [62]) samples but not the cell line studies. We included them, nonetheless, to boost the model's understanding of gene expression profiles (data downloaded from the NCI's Genomic Data Commons [63]). The drug target information was obtained through ID mapping and SMILES string search in ChEMBL [64], BindingDB [65] and PubChem databases. The binding targets curated by these sites were then mapped to the PANTHER gene ontology [66]. In total, 326 out of the 1801 combined compounds had known targets for the exact SMILES match. A Python data generator was used to join the cell line and drug features, the response value and other labels on the fly. For multitasking learning, the multiple partially labeled data collections were trained jointly in a round-robin fashion for each target, similar to how generative adversarial networks (GANs) [67] take turns to optimize loss functions for generators and discriminators.

Discussion

In this study, we sought to understand the performance of drug response prediction models in a broader context. We hypothesized that the observed R^2 score barrier of 0.8 [37] might partly be a result of variability in drug response assay. Indeed, we found that the measured dose–response values differed considerably among replicates in a single study. In the case of GDSC, this variability in terms of SD was more than 10% of the entire drug response range. Practically, this meant that, if we used one response value as prediction for another from the same cell–drug–dose combination, we would only obtain an average R^2 of 0.81. The cause for this variability is not well understood, but experimental protocol likely plays a big role, as evidenced by the lower variability observed among NCI60 replicates. Standardization in experimental design will be key to maximizing the value of screening data for joint learning.

When we moved beyond single studies to compare drug response values across studies, the variability increased. This phenomenon has been discussed by numerous previous studies [39–42] from a statistical consistency perspective. In this study, we approached it from a machine learning angle. Using the best available integrative metrics, we compared dose-independent response values across different studies. We arrived at rough upper bounds for how well models trained on one data source could be expected to perform in another. These numbers depended on whether the source and target studies used the same cell viability assay. In the case of identical assay, the R^2 ceiling was estimated to be 0.65 between CTRP and CCLE. In the case of different assays, it was markedly lower, at 0.31, between CTRP and GDSC.

These estimates put the recent progress in machine learning-based drug response prediction into perspective. We suggest that cross-study evaluation be used as an additional tool for benchmarking model performance; without it, it's difficult to know how much of the model improvement is generalizable. We illustrated this point with systematic characterization of cross-study performance using three different machine learning approaches. For example, going from a simple Random Forest model to LightGBM trained on CTRP, accuracy improved over 220% judging by cross validation R^2. However, the improvement on model generalization to CCLE was only 24%. In some cases,

extrapolation error actually increased as within-study performance improved. For an opposite example, the deep learning models made only marginal improvement over LightGBM in within-study performance, but the cross-study improvement, averaged 0.11 in R^2, was much more appreciable. This may be somewhat counterintuitive since neural networks are known to be overparameterized and prone to overfitting. However, as we have demonstrated with a multitasking model, the high capacity of deep learning models could be put to work with additional labeled data in auxiliary tasks.

Drug screening experiments are costly. How should we prioritize the acquisition of new data? A recent study showed with leave-one-out experiments that drug response models had much higher error when extrapolating to new drugs than new cell lines [25]. This is consistent with our finding. The models that did not generalize well tended to have been trained on data sets with fewer drugs (CCLE and gCSI). Further, when we withheld samples from training drug response models, we found that the loss of drug data was significantly more crippling than that of cell lines.

In addition to increasing the number of screened drugs, a careful selection based on mechanistic understanding or early experimental data may also help. As we have seen in 4, 5, 6, models trained on CTRP generalize well and this was not limited to studies with the same viability assay. A good example is the NCI60 column: as the target study, NCI60 used a different viability assay from both CTRP and GDSC, yet CTRP's prediction accuracy was notably higher than GDSC's. This may have to do with CTRP's design of an *Informer Set* of 481 compounds that target over 250 diverse proteins, covering a wide range of cell processes linked to cancer cell line growth [8]. Some probe molecules had no known protein targets but were selected for their ability to elicit distinct changes in gene expression profiles.

Taken together, this suggests that it would be beneficial for future screening studies to prioritize drug diversity. Given the vast design space of potentially active chemical compounds, estimated to be in the order of 10^{60} [68], intelligent methods that can reason about molecule classes are needed.

Conclusion

Precision oncology requires precision data. In this article, we reviewed five cancer cell line screening data sets, with a focus on drug response variabilities within and across studies. We demonstrated that these variabilities put constraints on the performance of machine learning models, in a way not obvious to traditional cross-validation within a single study. Through systematic analysis, we compared how models trained on one data set extrapolated to another, revealing a wide range in predictive power of both study data and machine learning methods. While deep learning results are promising, future success of drug response prediction will rely on the improvement of model generalizability. Experimental standardization and prioritization in drug diversity will also be key to maximizing the value of screening data for integrative learning.

Data availability

The integrated data files from this study are available in the Predictive Oncology Model & Data Clearinghouse hosted at the National Cancer Institute (https://modac.cancer.gov/assetDetails?dme_data_id=NCI-DME-MS01-8088592).

Key Points

- Cross-validation in a single study overestimates the accuracy of drug response prediction models, and differences in response assay can limit model generalizability across studies.
- Different machine learning models have varying performance in cross-study generalization, but they generally agree that models trained on CTRP are the most predictive and gCSI is the most predictable.
- Based on simulated experiments, drug diversity, more than tumor diversity, is crucial for raising model generalizability in preclinical screening.

References

1. Shoemaker RH. The NCI60 human tumour cell line anti-cancer drug screen. *Nat Rev Cancer* 2006;**6**(10):813–23.
2. Cortés-Ciriano I, van Westen GJP, Bouvier G, *et al*. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* 2016;**32**(1):85–95.
3. Xia F, Shukla M, Brettin T, *et al*. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics* 2018;**19**(18):486.
4. Barretina J, Caponigro G, Stransky N, *et al*. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**(7391):603–7.
5. Ghandi M, Huang FW, Jané-Valbuena J, *et al*. Next-generation characterization of the cancer cell line encyclopedia. *Nature* 2019;**569**(7757):503–8.
6. Yang W, Soares J, Greninger P, *et al*. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2012;**41**(D1):D955–61.
7. Basu A, Bodycombe NE, Cheah JH, *et al*. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;**154**(5):1151–61.
8. Seashore-Ludlow B, Rees MG, Cheah JH, *et al*. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* 2015;**5**(11):1210–23.
9. Ding MQ, Chen L, Cooper GF, *et al*. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol Cancer Res* 2018;**16**(2):269–78.
10. Rampášek L, Hidru D, Smirnov P, *et al*. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* 2019;**35**(19):3743–51.
11. Menden MP, Wang D, Mason MJ, *et al*. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun* 2019;**10**(1):1–17.
12. Douglass EF, Allaway RJ, Szalai B, *et al*. A community challenge for PANcancer drug mechanism of action inference from perturbational profile data. *bioRxiv* 2020. https://doi.org/10.1101/2020.12.21.423514.
13. Suphavilai C, Bertrand D, Nagarajan N. Predicting cancer drug response using a recommender system. *Bioinformatics* 2018;**34**(22):3907–14.
14. Gerdes H, Casado P, Dokal A, *et al*. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nat Commun* 2021;**12**(1):1–15.
15. Daoud S, Mdhaffar A, Jmaiel M, *et al*. Q-rank: reinforcement learning for recommending algorithms to predict drug sensitivity to cancer therapy. *IEEE J Biomed Health Inform* 2020;**24**(11):3154–61.
16. Kadurin A, Aliper A, Kazennov A, *et al*. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 2017;**8**(7):10883.
17. Huang C, Mezencev R, McDonald JF, *et al*. Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. *PLoS One* 2017;**12**(10):e0186906.
18. Dong W, Liu C, Zheng X, *et al*. Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. *BMC Bioinformatics* 2019;**20**(1):1–15.
19. Ran S, Liu X, Wei L, *et al*. Deep-resp-forest: a deep forest model to predict anti-cancer drug response. *Methods* 2019;**166**:91–102.
20. Rahman R, Dhruba SR, Ghosh S, *et al*. Functional random forest with applications in dose-response predictions. *Sci Rep*, 14 2019;**9**(1):1.
21. Bomane A, Gonçalves A, Ballester PJ. Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data. *Front Genet* 2019;**10**:1041.
22. Sidorov P, Naulaerts S, Ariey-Bonnet J, *et al*. Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front Chem* 2019;**7**:509.
23. Ran S, Liu X, Xiao G, *et al*. Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. *Brief Bioinform* 2020;**21**(3):996–1005.
24. Zeng X, Zhu S, Liu X, *et al*. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019;**35**(24):5191–8.
25. Li M, Wang Y, Zheng R, *et al*. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**18**:575–82.
26. Zhang F, Wang M, Xi J, *et al*. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci Rep* 2018;**8**(1):1–9.
27. Chang Y, Park H, Yang H-J, *et al*. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci Rep* 2018;**8**(1):1–11.

28. Liu P, Li H, Li S, *et al*. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics* 2019;**20**(1):1–14.

29. Oskooei A, Born J, Manica M, *et al*. PaccMann: prediction of anticancer compound sensitivity with multimodal attention-based neural networks. arXiv preprint arXiv:1811.06802. 2018.

30. Bazgir O, Zhang R, Dhruba SR, *et al*. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nat Commun* 2020;**11**(1):1–13.

31. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2018;**19**(2):325–40.

32. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev* 2019;**11**(1):31–9.

33. Pucher BM, Zeleznik OA, Thallinger GG. Comparison and evaluation of integrative methods for the analysis of multilevel omics data: a study based on simulated and experimental cancer data. *Brief Bioinform* 2019;**20**(2):671–81.

34. Adam G, Rampášek L, Safikhani Z, *et al*. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol* 2020;**4**(1):1–10.

35. Chen J, Zhang L. A survey and systematic assessment of computational methods for drug response prediction. *Brief Bioinform* 2021;**22**(1):232–46.

36. Wang Z, Li H, Guan Y. Machine learning for cancer drug combination. *Clin Pharmacol Therap* 2020;**107**(4):749–52.

37. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Brief Bioinform* 2021;**22**(1):360–79.

38. Paltun BG, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Brief Bioinform* 2021;**22**(1):346–59.

39. Haibe-Kains B, El-Hachem N, Birkbak NJ, *et al*. Inconsistency in large pharmacogenomic studies. *Nature* 2013;**504**(7480):389–93.

40. Mpindi JP, Yadav B, Östling P, *et al*. Consistency in drug response profiling. *Nature* 2016;**540**(7631):E5–6.

41. Safikhani Z, Smirnov P, Freeman M, *et al*. Revisiting inconsistency in large pharmacogenomic studies. *F1000Res* 2016;**5**.

42. Sadacca B, Hamy A-S, Laurent C, *et al*. New insight for pharmacogenomics studies from the transcriptional analysis of two large-scale cancer cell line panels. *Sci Rep* 2017;**7**(1):1–12.

43. Haverty PM, Lin E, Tan J, *et al*. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* 2016;**533**(7603):333–7.

44. Smirnov P, Kofia V, Maru A, *et al*. Pharmacodb: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res* 2017;**46**(D1):D994–1002.

45. Rahman R, Dhruba SR, Matlock K, *et al*. Evaluating the consistency of large-scale pharmacogenomic studies. *Brief Bioinform* 2019;**20**(5):1734–53.

46. Gupta A, Gautam P, Wennerberg K, *et al*. A normalized drug response metric improves accuracy and consistency of anticancer drug sensitivity quantification in cell-based screening. *Commun Biol* 2020;**3**(1):1–12.

47. Dhruba SR, Rahman R, Matlock K, *et al*. Application of transfer learning for cancer drug sensitivity prediction. *BMC Bioinformatics* 2018;**19**(17):51–63.

48. Zhu Y, Brettin T, Evrard YA, *et al*. Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci Rep* 2020;**10**(1):1–11.

49. Clyde A, Brettin T, Partin A, *et al*. A systematic approach to featurization for cancer drug sensitivity predictions with deep learning. arXiv preprint arXiv:2005.00095. 2020.

50. Yadav B, Pemovska T, Szwajda A, *et al*. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci Rep* 2014;**4**:5193.

51. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.

52. Ke G, Meng Q, Finley T, *et al*. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017;**30**:3146–54.

53. Partin A, Brettin T, Evrard YA, *et al*. Learning curves for drug response prediction in cancer cell lines. *BMC bioinformatics* 2020;**22**:1–8.

54. Shahriyari L. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Brief Bioinform* 2019;**20**(3):985–94.

55. Koleti A, Terryn R, Stathias V, *et al*. Data portal for the library of integrated network-based cellular signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res* 2018;**46**(D1):D558–66.

56. Kode srl. *Dragon (software for molecular descriptor calculation)* version 7.0.8, 2017. https://chm.kode-solutions.net.

57. Kim S, Chen J, Cheng T, *et al*. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**(D1):D1102–9.

58. O'Boyle NM, Banck M, James CA, *et al*. Open babel: an open chemical toolbox. *J Chem* 2011;**3**(1):1–14.

59. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.

60. He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA: IEEE, 2016, 770–8.

61. Bickerton GR, Paolini GV, Besnard J, *et al*. Quantifying the chemical beauty of drugs. *Nat Chem* 2012;**4**(2):90.

62. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol* 2015;**19**(1A):A68.

63. Jensen MA, Ferretti V, Grossman RL, *et al*. The NCI genomic data commons as an engine for precision medicine. *Blood* 2017;**130**(4):453–9.

64. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**(D1):D1100–7.

65. Gilson MK, Liu T, Baitaluk M, *et al*. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 2016;**44**(D1):D1045–53.

66. Mi H, Muruganujan A, Casagrande JT, *et al*. Large-scale gene function analysis with the panther classification system. *Nat Protoc* 2013;**8**(8):1551–66.

67. Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. *Adv Neural Inf Process Syst* 2014;**27**:2672–80.

68. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* 1996;**16**(1):3–50.