Drug Sensitivity Prediction From Cell Line-Based Pharmacogenomics Data: Guidelines for Developing Machine Learning Models

Hossein Sharifi-Noghabi[1,2,4], Soheil Jahangiri-Tazehkand[3,4,6], Casey Hon[4,6], Petr Smirnov[3,4,6], Anthony Mammoliti[3,4,6], Sisira Kadambat Nair[4,6], Arvind Singh Mer[3,4,6], Martin Ester[1,2,*], Benjamin Haibe-Kains[3,4,5,6*]

[1]School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada.
[2]Vancouver Prostate Center, Vancouver, British Columbia, Canada.
[3]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.
[4]Princess Margaret Cancer Centre, Toronto, Ontario, Canada.
[5]Ontario Institute for Cancer Research, Toronto, Ontario, Canada.
[6]University of Toronto, Toronto, Ontario, Canada.

*Corresponding authors: ester@sfu.ca; bhaibeka@uhnresearch.ca

ABSTRACT

The goal of precision oncology is to tailor treatment for patients individually using the genomic profile of their tumors. Pharmacogenomics datasets such as cancer cell lines are among the most valuable resources for drug sensitivity prediction, a crucial task of precision oncology. Machine learning methods have been employed to predict drug sensitivity based on the multiple omics data available for large panels of cancer cell lines. However, there are no comprehensive guidelines on how to properly train and validate such machine learning models for drug sensitivity prediction. In this paper, we introduce a set of guidelines for different aspects of training a predictor using cell line datasets. These guidelines provide extensive analysis of the generalization of drug sensitivity predictors, and challenge many current practices in the community including the choice of training dataset and measure of drug sensitivity. Application of these guidelines in future studies will enable the development of more robust preclinical biomarkers.

INTRODUCTION

Cancer is a complex genetic disease. Due to the heterogeneous nature of tumors the treatment of cancer is very challenging. Precision oncology aims to tailor the therapies according to the genomic profile of the tumor. Pharmacogenomics, a crucial component of precision oncology, promises utilizing the genomic landscape of each individual patient to find most effective treatment options [5–7]. However, it still has limited clinical utility [8] and the availability of clinical pharmacogenomics datasets is limited by a lack of public access and small size, both in terms of patient cohorts and investigated therapies for the few publicly available datasets. As a

result, resources such as cancer cell lines [1–4,9–11], patient-derived xenografts [12,13], or organoids [14] are being employed in pharmacogenomics to decipher drug sensitivity prediction. Although these preclinical resources do not fully recapitulate the inter- and intra-tumor heterogeneity of cancer, they act as proxies for patient tumors and provide larger dataset, usually screened with hundreds or thousands of drugs separately or in combination along with multi-omics characterization [15].

Due to the complexity of generating pharmacogenomics datasets, discrepancies can even exist across cell line datasets [11,16–22]. However, recent efforts such as the PharmacoDB project (pharmacodb.ca) [15], the ORCESTRA platform (orcestra.ca) [23], and CellMinerCDB [24] aimed at standardizing, and integrating different preclinical pharmacogenomics datasets to improve downstream machine learning modeling. The data-rich nature of preclinical pharmacogenomics datasets has paved the way for the development of machine learning approaches to predict drug sensitivity *in vitro* and *in vivo* [25–27]. These computational approaches range from simple linear regression models [28,29] Lasso [30], and Elastic Net [31] to Random Forest [32], kernel-based models [33–36], highly non-linear models based on Deep Neural Networks [37–44], and most recently, reinforcement learning [45], few-shot learning [46], and multi-task learning [47]. These methods often take gene expression as input and predict the area above/under the dose-response curve (AAC/AUC) or half-maximal inhibitory concentration (IC50), the concentration of the drug that reduces the viability by 50%.

While machine learning for pharmacogenomics is a promising direction [25], existing guidelines are based on a single pharmacogenomics dataset [48] or based on benchmarking different methods without considering technical differences between molecular profiles or drug screening assays across different datasets [26]. We believe that there is a need for comprehensive guidelines based on multiple uniformly processed datasets on how to properly train and evaluate drug sensitivity predictors. In this study, we conduct a systematic and comprehensive analysis based on RNA-seq data as the input and different measures of drug sensitivity such as AAC and IC50 as the output. We employ univariable modeling (using known biomarkers) and multivariable modeling (using state-of-the-art machine learning methods) to investigate generalization in drug sensitivity prediction. We consider two common machine learning paradigms: within-domain analysis and cross-domain analysis. In within-domain analysis, models are trained and tested on the same dataset via cross-validation which means train and test data are from the same distribution. In cross-domain analysis, models are trained and tested on different cell line datasets to investigate generalization capability. We also examine the effect of an analysis choice first proposed by [1], to separate the data for cell lines originating from hematopoietic cancers and solid tumours on the ability to learn predictors of drug sensitivity.

As a result of this study, we provide guidelines, which we refer to as PGx Guidelines (Figure 1), on the following questions:

1. Which dataset(s) and measure(s) of drug sensitivity are best for training predictors?

2. How much does the performance of pharmacogenomics methods change when moving from within-domain analysis to cross-domain analysis?
3. What is the impact of non-solid tumors on the performance of drug sensitivity predictors?

We argue that it is necessary to evaluate generalization of cell line-based predictors first on cell line datasets before employing them on PDX or patient data, and therefore we focus on cell line datasets in this study. We believe that the PGx Guidelines will lead to the development of more accurate and more generalizable machine learning models for drug sensitivity prediction from pharmacogenomics data and will contribute toward the goal of extending the benefits of precision oncology to a wider range of patients.

METHODS

Datasets

We employed the following pan-cancer datasets (Table 1):

- The Cancer Therapeutics Response Portal (CTRPv2) [1,2]
- The Genentech Cell Line Screening Initiative (gCSI) [10,11]
- The Genomics of Drug Sensitivity in Cancer (GDSCv2) [3,4]

Table 1 Characteristics of the studied datasets

	CTRPv2	GDSCv2	GDSCv1	gCSI
# Drugs	544	190	343	16
# Cell Lines	821	328	427	334
# Tissue Types	25	27	28	22
# Genes (# Protein Coding)	60662 (19957)	60662 (19957)	60662 (19957)	60662 (19957)
Gene Expression Assay	RNA-seq **	RNA-seq	RNA-seq	RNA-seq
Sensitivity Assay	CellTiter Glo	CellTiter Glo	Syto60	CellTiter Glo
Usage	Training	Test	Training	Test

* Obtained from the ORCESTRA platform. Numbers are dependent to the preprocessing method for the RNA-seq

** Gene expression for CTRPv2 were extracted from the CCLE dataset.

We obtained these datasets in the format of PharmacoSet (PSet) which is an R-based data structure that aids in reproducible research for drug sensitivity prediction. PSets are obtained via the ORCESTRA platform (orcestra.ca) [23]. The molecular profiles (RNA-seq) were preprocessed via Kallisto 0.46.1 [49] using GENCODE v33 transcriptome as the reference and the pharmacological profiles (AAC and IC50) were preprocessed and recomputed via PharmacoGx package [50]. In this paper, we focused on 11 drugs in common between these

datasets including: Bortezomib, Entinostat, Sirolimus, Docetaxel, Gemcitabine, Crizotinib, Lapatinib, Vorinostat, Erlotinib, Paclitaxel, and Pictilisib. These datasets have missing values for different samples and given a specific drug, the number of available cell lines for training/test can change (Table S1).

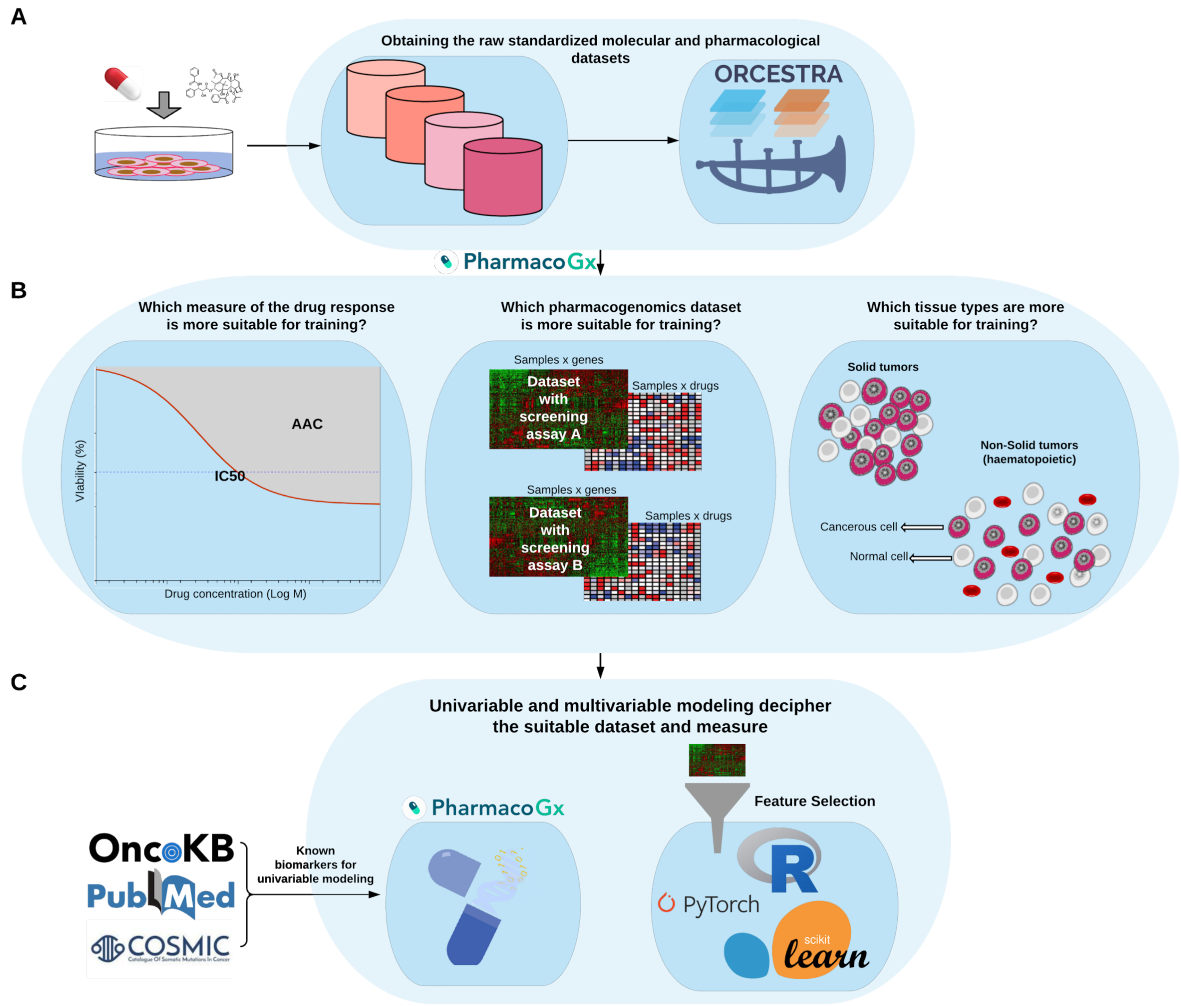

Figure 1. Schematic overview of the PGx Guidelines; A) The raw pharmacogenomics datasets are obtained from the ORCESTRA platform. B) The molecular (RNA-seq) and pharmacological profiles (Area Above dose-response Curve--AAC) are obtained for each dataset via the PharmacoGx package. Finally, C) univariable modeling based on known biomarkers from the literature is performed via the PharmacoGx package, and multivariable modeling using protein coding genes and feature selection is performed via different packages.

State-of-the-art in preclinical pharmacogenomics

We categorized the state-of-the-art predictors of drug sensitivity based on their input, output, and the pharmacogenomics datasets that they used for training and test (Figure 2). Gene expression was the most common input data type to predict drug sensitivity as it was determined to be the most effective data type in multiple studies [4,26,31,33]. However, some

studies based on multi-omics data also demonstrated that adding other omics data types can improve the prediction performance [26,40]. For drug sensitivity, IC50 was the most common measure used. The cross-domain training approach was more common compared to the within-domain approach. Moreover, the majority of these methods were trained on GDSCv1 gene expression data. We also observed that incorporating drug structure, such as the SMILES representation of the drug molecule, is an emerging trend in the field. Since the goal of this study is investigating generalization in gene expression cell line-based predictors, we did not provide detailed descriptions of drug structure, interaction, adverse reaction, the type of clinical or PDX datasets that existing methods have employed and illustrated all of them under broad categories of “drugs”, “patients”, and “PDX”, respectively.

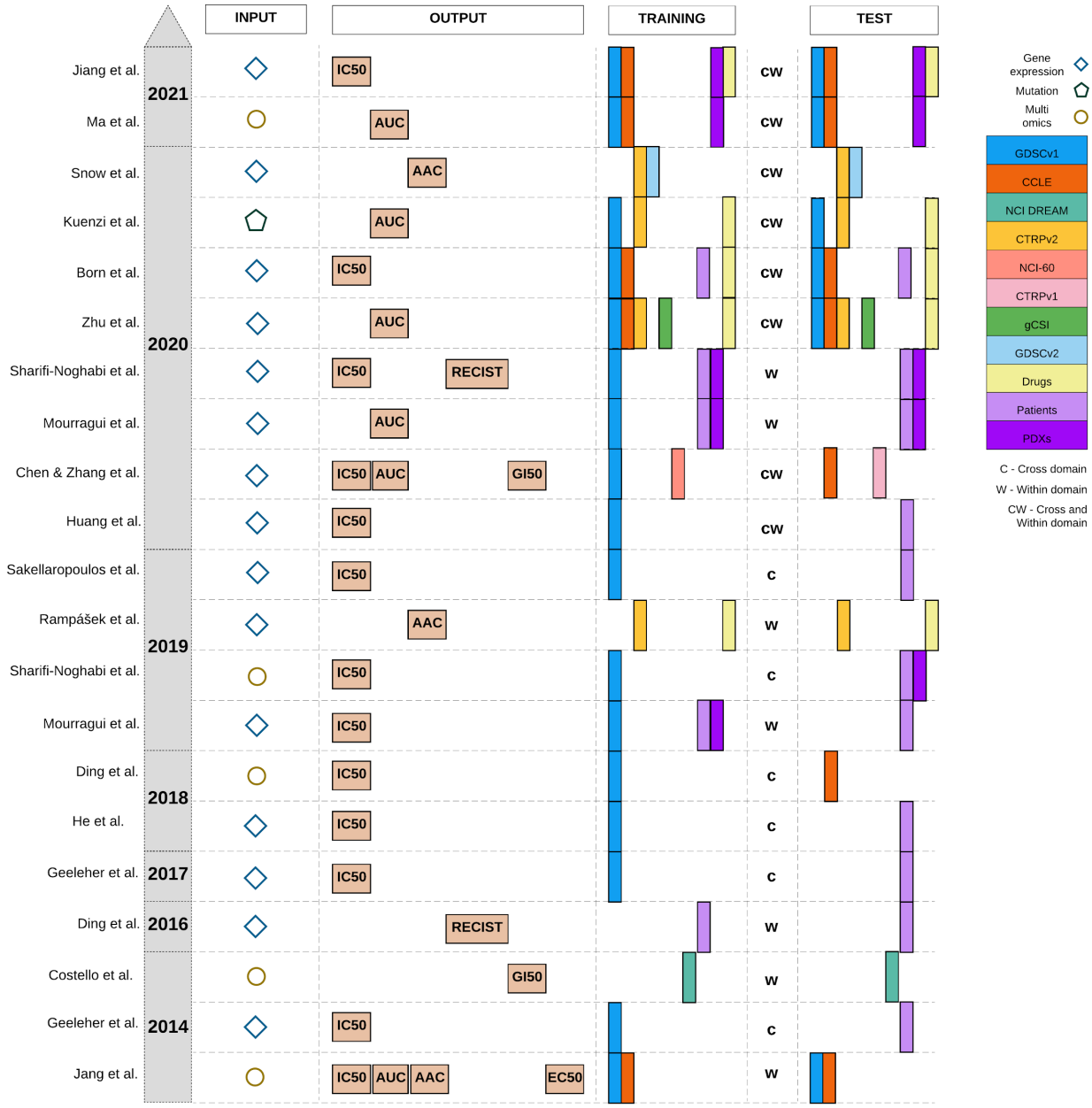

Figure 2. **Published studies for drug sensitivity prediction. Gene expression is the most common molecular profile and IC50 is the most common pharmacological profile, but AAC/AUC has become more common in recent studies. GDSCv1 (originally named CGP) is the most common training dataset and the use of drug information for training has been more frequent in recent years. The cross-domain training approach denoted by "c" was more common compared to the within-domain approach denoted by "w". When a method employs both of them, we denote it by "cw".**

Univariable and Multivariable Analysis

To study the impact of model complexity on the performance and generalization of drug sensitivity predictors, we investigate a wide range of modeling approaches ranging from

univariable modeling based on known biomarkers to highly complex and multivariable approaches such as Deep Neural Networks.

The univariable analysis consists of retrieving estimates for the association between the known biomarkers and AAC or IC50 as measures of drug sensitivity. with the known biomarkers. In this study, we focused on the 11 drugs in common between the studied datasets as molecular features. A total of 69 biomarkers were retrieved from literature and used in the analysis (Table S2). Since the pharmacological datasets store genes as Ensembl gene variants, the biomarkers were mapped to the variant identifier using the Uniprot Retrieve/ID mapping tool (Table S3).

To address the issues with IC50 values described above, we explored three preprocessing approaches. Log transformation of the values, truncating the values, and a combination of both. Log transformation stretches small values and truncating based on predefined concentration ranges addresses very large IC50 values. AAC values were left unchanged.

Then, we used the PharmacoGx *drugSensitivitySig* function to compute estimates of the association between biomarkers and drug sensitivity. This is modelled using a linear regression model: $Y = \beta_0 + \beta_i G_i + \beta_t T + \beta_b B$. *Y* denotes the drug sensitivity variable, G_i denotes the expression of gene *i*, *T* denotes the tissue type, and the βs are the regression coefficients [50]. β represents the strength of the gene-drug association, while scaled *Y* and *G* estimate the standardized coefficients of the linear model. Ydenotes the estimates for each drug-gene association.

The multivariable analysis consists of making predictions of the drug sensitivity measures (AAC or IC50) given the level of expression of input gene features. Unlike univariable analysis which considers each biomarker at a time, multivariable analysis considers all of the input genes together. The goal of multivariable modeling is to learn a mapping function $f_\theta(X)$parametrized by θthat maps the input gene expression matrix $X^{N\times M}$ to the drug sensitivity values $Y^{N\times 1}$,where N is the number of samples and M is the number of input features. We employed the mean squared error as the loss function to optimizeθas follows: $L_{MSE} = ||Y - \overline{Y}||_2^2 + \Omega(\theta)$,where $\Omega(.)$ denotes regularization. In PGx guidelines, we explored Ridge Regression where $\Omega(\theta) = \alpha||\theta||_2^2$, Elastic Net where $\Omega(\theta) = \lambda[(1-\alpha)||\theta||_2^2 + \alpha||\theta||_1]$, Random Forest which utilizes bootstrapping to split the data into subsets and fits a decision tree to each of them in order to make the final predictions by averaging the results, and Deep Neural Networks that utilizes $\Omega(\theta) = ||\theta||_2^2$, as well as dropout and early stopping for regularization and learnsθby layer-wise (fully-connected) transformation. More details on training and hyper-parameters of the final models corresponding to each method are provided in the supplementary material.

Within-Domain and Cross-Domain Analysis

To study the impact of data discrepancy on generalization of drug sensitivity predictors, we investigate two common approaches of within-domain and cross-domain. In within-domain analysis, the goal is to train and test models on the same dataset via cross-validation. The hypothesis is that if a model trained to predict sensitivity for a given drug cannot make accurate predictions for the same dataset (on the test splits), it is very unlikely that it generalizes to other datasets for the same drug. In cross-domain analysis, the goal is to train and test models on different datasets. The hypothesis is that models that demonstrate high performance in the within-domain should have better performance in cross-domain and models that perform poorly in within-domain should also perform poorly in cross-domain analysis. Intuitively, if there is enough predictive information in the training data for the given drug, the model should have a higher chance of generalizing to other datasets for the same drug (if the other datasets also have adequate predictive information).

Experimental Design

We designed our experiments to justify the choice of training dataset (the input) and the measure of drug sensitivity (the output) as well as studying generalization of different models in within-domain and cross-domain analysis.

PGx Guideline experimental questions: IC50 is the most common measure of the drug sensitivity for machine learning (Figure 2) however it suffers from known limitations. By definition, the IC50 does not exist for any experiment where the maximum inhibition of growth is not at least 50%. Furthermore, IC50 estimation is unstable when there is not at least 1 point measured on each plateau of the curve [51], and the common technique to overcome this limitation, by setting IC50's outside the measured range to the maximal tested concentration, effectively creates a right censored measurement and loses all differences in sensitivity between such experiments. Our hypothesis is models trained to predict AAC should generalize better than those trained to predict IC50.

To answer the first PGx Guideline question on the best measure of drug sensitivity, we investigated a wide range of modeling approaches ranging from simple univariable models based on known biomarkers of each drug to more complex multivariable models based on Ridge Regression, Elastic Net, Random Forest, and Deep Neural Networks. We also compared CTRPv2 and GDSCv1 to determine which dataset is a better training dataset to build drug sensitivity predictors. We picked GDSCv1 as the competitor because it is the most common training dataset (Figure 2). The issue with GDSCv1 utilizes a different drug screening assay compared to the other datasets. Our hypothesis is that CTRPv2 is a more generalizable training dataset because it utilizes the same assay as other datasets and also has a relatively large sample size.

To answer the second question on generalization performance, we utilized the state-of-the-art cell line datasets (see Datasets section) in within-domain and cross-domain analysis. For within-domain analysis, we used 10-fold nested cross validation on CTRPv2, the largest dataset

in our collection -- 9 folds for train and validation and the 10th fold for testing. For cross-domain analysis, we trained models on the CTRPv2 and tested them on the other cell line datasets (GDSCv2 and gCSI).
To answer the last question, we investigated the association between tissue type and model predictions when the model was trained with all available tissue types (solid and non-solid tissues) and when it was trained only on solid tissue types.

Evaluation: We employed the Pearson correlation coefficient as the performance metric for all of the analyses and experiments of the PGx Guidelines. In all of the analyses, the Baseline performance indicates the Pearson correlation of cell lines in common between train and test dataset of that particular analysis. To summarize the figures, we also reported average±standard deviation of the Pearson correlation over the 11 drugs in common (only in the main text). We also reported the same results using the Spearman correlation in the supplementary material.

Implementation details: To ensure reproducibility of this study, we provide a detailed description of preprocessing, training, and evaluation. For all of the multivariable experiments, gene expression input data was normalized via z-score transformation using the parameters of the training dataset. Furthermore, to correct for the impact of tissue type, the one-hot encoded representation of it was added as an input feature to the normalized expression data after removing non-solid tissue types (except when the goal was to study the impact of non-solid samples) and those tissue types that were not available in the training data (CTRPv2).
We implemented the univariable analysis in R via the PharmacoGx package (version 2.0.5) [50]. The multivariable analyses were implemented in Python using the scikit-learn package (version 0.23.2). All of the hyper-parameter tunings were performed via grid search in nested 10-fold cross validation. We repeated the within-domain experiments 10 times and fixed the random seed for the cross-domain experiments and performed it once.
We implemented the deep neural networks in the Pytorch framework (version 1.4 cpu only) and used 10-fold cross validation and 100 trials of random search to select the best hyper-parameter settings for each drug.
In all of the analyses, we employed previously reported values of the hyper-parameters as our initial sets for each method and tuned to select the best setting for each method. More details on the considered values and the selected ones are provided in the supplementary material (Table S4).

Research reproducibility: All the data, code, and results employed and obtained in this study are publicly available for research reproducibility.
Code and supplementary tables/data: https://github.com/bhklab/PGx_Guidelines
Data and models: https://zenodo.org/record/4642024#.YGCkbK9KiUl

RESULTS

Models trained to predict AAC outperforms those trained to predict IC50

Two common summary metrics have been used in the literature for summarizing dose-response experiments, the IC50 and the AAC. The IC50 is a measurement of potency, while the AAC can be interpreted as measuring an average of potency and maximal efficacy, or as a measure of the mean viability across the concentrations tested. While the IC50 is easily interpretable and is an absolute metric (unlike the AAC, which depends on the concentration range tested), the IC50 has some technical drawbacks which may make it difficult to use in training machine learning models. AAC/AUC is a normalized value between zero and one, but IC50 (the concentration) is not necessarily bounded and can be very small (close to zero) when samples are highly sensitive to a given drug or very large when they are highly resistant to a given drug. These issues make preprocessing of IC50 critical. Therefore, we investigated both methods to preprocess IC50, as one of the key measures of the drug sensitivity in previous studies, and then exploited univariable and multivariable analysis to compare these two metrics.

Univariable analysis reveals Log truncated IC50 outperforms Log IC50 or Truncated IC50 for estimating associations using known biomarkers: We employed univariable analysis of the known biomarkers on different preprocessing approaches for IC50 including, 1) estimating the associations using the raw IC50, 2) estimating the associations using the truncated IC50, 3) estimating the association using log transformed IC50, and 4) estimating the association using log truncated IC50 values. We chose Lapatinib and Erlotinib as the showcase because these drugs have ERBB2 and EGFR genes as strong known biomarkers but included this analysis for all known biomarkers of the studied drugs (Table S5). We observed that log truncated IC50 demonstrates the most predictive and consistent results across different datasets (Figure 3A). This indicates that among different approaches to preprocess IC50, it is better to log truncate the values. For simplicity, we refer to log truncated IC50 as IC50 for the rest of the paper.

Univariable analysis reveals that AAC outperforms IC50 for estimating associations using known biomarkers: We performed the same process of estimating associations for AAC and compared it to IC50. AAC captures the associations between known biomarkers and drugs more accurately compared to IC50. Moreover, AAC demonstrated more consistent results across different datasets compared to IC50 (Figure 3A). These results indicate that AAC is a better metric for univariable drug sensitivity prediction compared to IC50. To further validate these results, we also compared AAC and IC50 via multivariable analysis.

Multivariable within-domain analysis confirms that models trained to predict AAC outperforms those of IC50: We compared Ridge Regression, Elastic Net, and Random Forest when trained on protein coding genes to predict AAC and IC50 in a within-domain analysis using CTRPv2. We observed that AAC achieved significantly higher within-domain performance

(Figure 3B -- Ridge Regression achieved 0.24±0.17 in AAC vs. 0.23±0.14 in IC50; Elastic Net achieved 0.4±0.17 in AAC vs. 0.31±0.12 in IC50; Random Forest achieved 0.41±0.1 in AAC vs. 0.33±0.07 in IC50). These results reconfirm univariable results that AAC is a better metric for drug sensitivity prediction compared to IC50. We observed a similar pattern in the Spearman results (Figure S1). These experimental results also align with the within-domain results in the previous study [48].

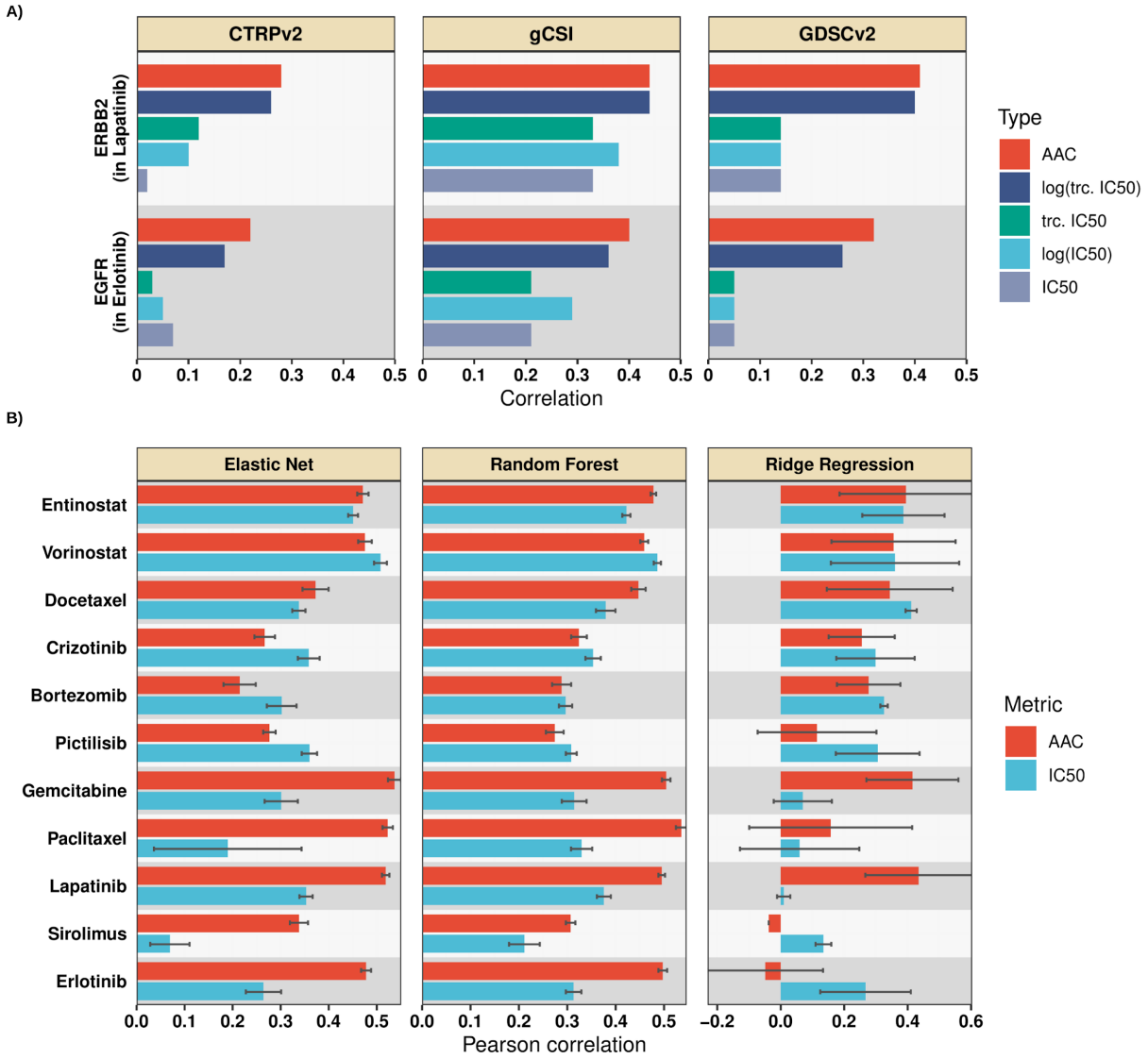

Figure 3. A) Comparison of AAC (Red) to different approaches of preprocessing IC50 via univariable analysis, tested on top biomarkers of Lapatinib (ERBB2) and Erlotinib (EGFR) in terms of estimated correlation. B) Mean Pearson correlation (over 10 runs) for different multivariable methods in within-domain analysis trained to predict AAC or IC50. Both univariable and multivariable within-domain analysis results indicated that AAC outperforms IC50 on average.

Cross-domain analysis decreases generalization performance

To study the generalization capabilities of drug sensitivity predictors, we analyzed them in a cross-domain setting where the models are trained and tested on different cell line datasets. To perform this analysis, first we determined the most suitable training dataset.

Multivariable analysis reveals that models trained on CTRPv2 outperform those of GDSCv1 in generalization: As mentioned before, GDSCv1 is the most common training dataset for machine learning in pharmacogenomics. However, this dataset utilized the Syto60 assay in contrast to other major pharmacogenomics datasets that utilized the CellTiter Glo assay. We believe the difference in the drug screening assay influences the generalization capability of models because the Syto60 assay generates nosier drug response estimates [11]. To validate this, we trained two Ridge Regression models to predict AAC using protein coding genes: one trained on CTRPv2 and the other trained on GDSCv1 and then tested both of them on gCSI. We removed GDSCv2 from this analysis because of the overlap of molecular profiles with GDSCv1. We selected CTRPv2 because it is larger compared to gCSI which makes it naturally a more viable choice for training.
We observed that the model trained on CTRPv2 demonstrated significantly better performance in terms of the Pearson correlation compared to the model which was trained on GDSCv1 (Figure 4). To be more specific, on average (over 11 drugs), the model trained on CTRPv2 achieved the Pearson correlation of 0.4±0.21 (0.39±0.18 for IC50--Table S6), while the model trained on GDSCv1 achieved 0.26±0.16 (0.16±0.17 for IC50--Table S6). This suggests that agreement between the drug screening assay plays a significant role in cross-domain generalization. We observed a similar pattern in the Spearman correlation (Figure S2)

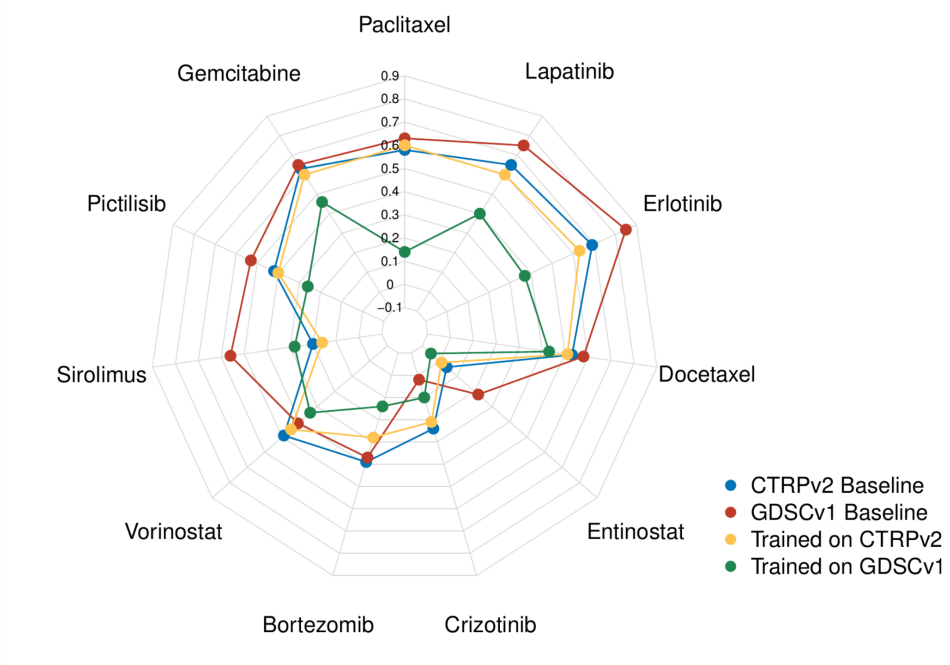

Figure 4. Comparison of CTRPv2 and GDSCv1 as training datasets for Ridge Regression models in terms of Pearson correlation in cross-domain analysis to predict AAC on gCSI. For the majority of the drugs, models trained on CTRPv2 (Yellow) achieved better performance compared to those trained on GDSCv1 (Green) even though GDSCv1 has higher baseline correlation (correlation of AAC values between cell lines in common between GDSCv1/CTRPv2 and gCSI -- CTRPv2 baseline in Blue and GDSCv1 baseline in Red).

Multivariable analysis reveals that the performance of models decreases when moving from within-domain to cross-domain analysis: To study the generalization of drug sensitivity predictors, we trained different models on CTRPv2 dataset and tested them of GDSCv2 and gCSI (Figure 5A). The cross-domain performance for the majority of the studied drugs is decreased significantly compared to the within-domain performance (Figure 5B-D). Elastic Net performance decreased from 0.4±0.17 (within-domain AAC) to 0.34±021 in GDSCv2 and 0.34±0.21 in gCSI (both in AAC- Figure 5C). Similarly, Random Forest decreased from 0.41±0.1 (within-domain AAC) to 0.33±0.2 in GDSCv2 and 0.35±0.21 in gCSI (both in AAC- Figure 5D). Ridge Regression demonstrated a different trend and increased from 0.24±0.17 (within-domain AAC) to 0.33±0.17 in GDSCv2 and 0.4±0.21 in gCSI (both in AAC- Figure 5A). This was due to some outlier predictions; Pearson correlation is sensitive to outliers and when we looked at the Spearman correlation results, the performance of Ridge Regression also decreased in cross-domain analysis (Figure S1).

These results suggest that even when the train and test data of a model utilized the same drug screening assay and were preprocessed similarly, it does not necessarily guarantee generalization. Moreover, the within-domain and cross-domain analysis together suggest that making models more complex improves the performance compared to the univariable analysis.

For example, ERRB2 had an estimated association of 0.41 in GDSCv2 and 0.44 in gCSI, however, Elastic Net achieved 0.61 in GDSCv2 and Ridge Regression achieved 0.6 in gCSI which demonstrates the power of multivariate analysis. Finally, the cross-domain analysis using multivariable methods also reconfirmed that AAC is a better metric compared to IC50 (Table S7) because datasets are more consistent on AAC (baseline correlations of 0.45±0.18 in GDSCv2 and 0.43±0.21 in gCSI for AAC in contrast to 0.38±0.27 and 0.42±0.13 in IC50, respectively) and methods are more accurate (for example, in gCSI, Ridge Regression achieved 0.4±0.21 in AAC vs. 0.36±0.18 in IC50; Elastic Net achieved 0.34±0.21 in AAC vs. 0.33±0.18 in IC50; DNN achieved 0.35±0.2 in AAC vs. 0.32±0.21 in IC50; Random Forest achieved 0.35±0.2 in AAC vs. 0.37±0.16 in IC50). We observed a similar pattern in the Spearman correlation (Figure S3A-D). We also observed similar patterns in both Pearson and Spearman results when comparing the best performing within-domain model to the best performing cross-domain model for each drug (Figure S3E-F).

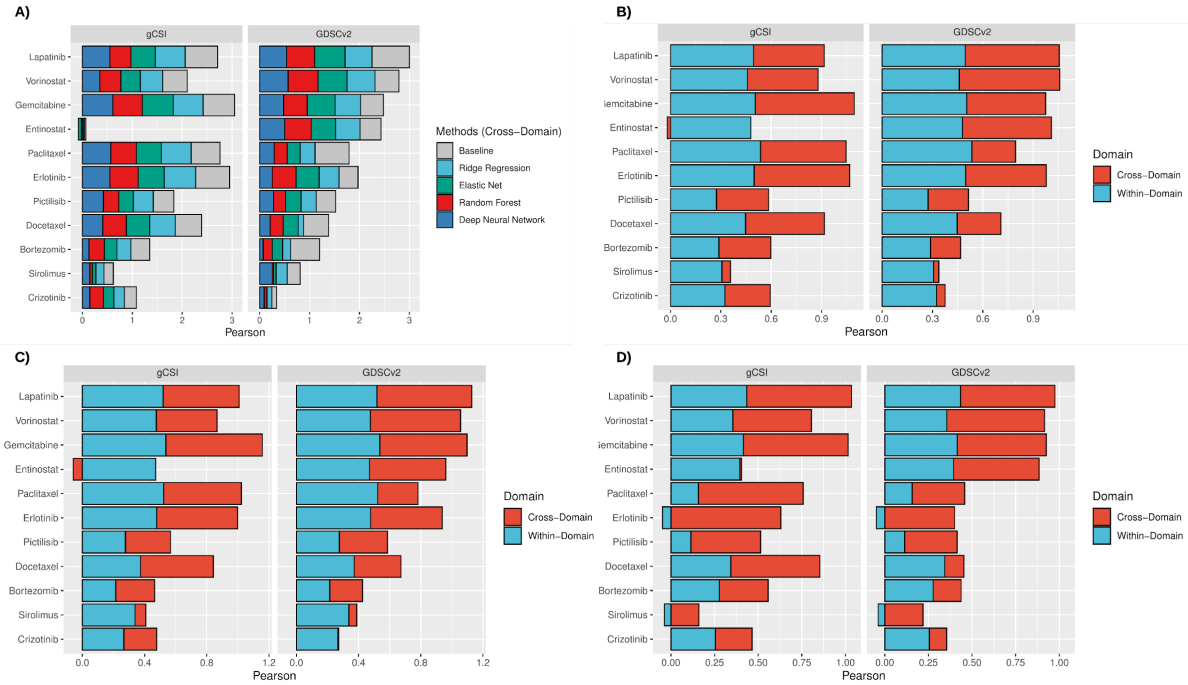

Figure 5. Comparison of multivariable methods in terms of Pearson correlation in cross-domain analysis trained on CTRPv2 to predict AAC, tested on GDSCv2 and gCSI (A). Comparison of the within-domain performance of Ridge Regression (B), Elastic Net (C), and Random Forest (D) to their cross-domain performance in terms of Pearson correlation.

Non-solid tissue types influence the performance of models: The majority of pharmacogenomics datasets are pan-cancer with solid and non-solid tissue types. We studied the molecular and pharmacological profiles of non-solid tissues (haematopoietic and lymphoid tissue types) in CTRPv2, GDSCv2, gCSI . We observed that the sensitivity outcome (AAC) in non-solid samples is significantly different compared to solid samples and they tend to be more sensitive than solid samples (Figure S4A-C) which aligns with the previous studies [1,52,53].

Similarly, the non-solid samples also clustered differently compared to solid samples (Figure S4D-F). These results raise the question of whether including both liquid and solid lines in the training set is beneficial for learning models to predict drug sensitivity.

To answer this, we trained two Ridge Regression models to predict AAC using protein coding genes as follows: one trained on all samples (solid and non-solid together) in CTRPv2, and the other one trained only on solid samples (non-solid samples were removed) in CTRPv2. We measured the associations between the predictions and the binary status of tissue type (solid vs. non-solid) in GDSCv2 and gCSI using the area under precision-recall curve (AUPR). The predictions of the model which was trained on all samples demonstrated a very high AUPR compared to the model that was only trained on solid samples. This suggests that by including non-solid samples, models predict the tissue type rather than the drug sensitivity itself (Figure 6). To confirm this, we trained another Ridge Regression model after removing a random subset of solid samples with the same size as the non-solid samples to make sure that the observed result was not because of sample size (we repeated the random selection 10 times and reported the average value). We also reported the Baseline AUPR which indicates the ground truth association between actual AAC and the binary status of tissue type. We observed that models that were trained on all samples (including solid and non-solid tissues) and the one with a random subset removed, had the highest AUPR in both GDSCv2 and gCSI for the majority of the drugs compared to the model that was trained only on solid samples (Figure 6). This confirms previous results that the difference in molecular profiles and drug sensitivity of non-solid samples have a negative impact on the drug sensitivity prediction task and it is crucial to remove all non-solid tissue types before any machine learning modeling.

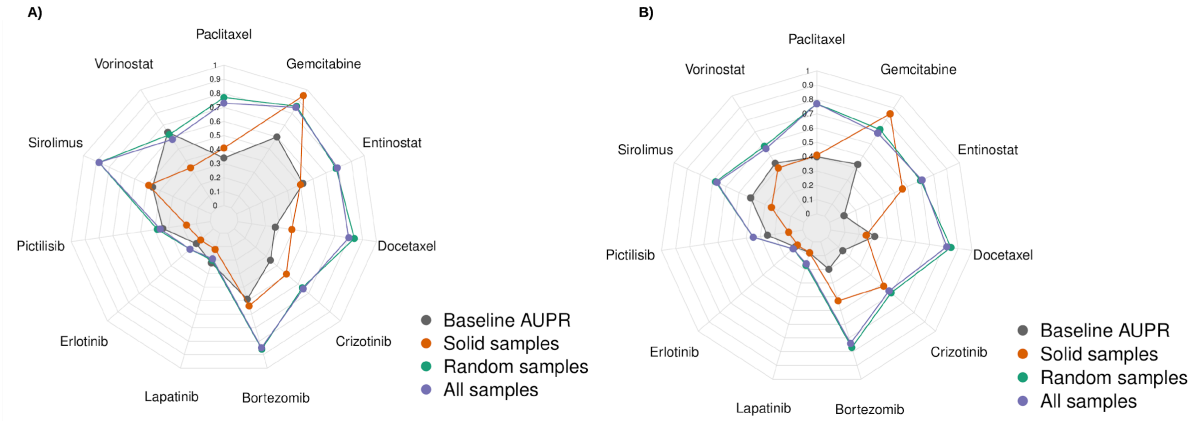

Figure 6. The association between predictions of a Ridge Regression model trained on CTRPv2 to predict AAC and tissue binary type status of GDSCv2 (A) and gCSI (B) in terms of AUPR. Training on cell lines with solid (Red) tissue types had lowest AUPR compared to training on all samples (solid and non-solid--Purple) and training on a randomly selected subset with the same size as solid samples (average AUPR over 10 different subsets--Green). The results confirmed that removing non-solid samples decreases the association of predictions with tissue type.

DISCUSSION

In this study, we investigated the fundamental challenges of developing machine learning models to predict drug sensitivity from cell line pharmacogenomics data. We named our study PGx Guidelines because we believe answers to these questions provide actionable guidelines for developing predictors of drug sensitivity.

The guidelines show that the performance of machine learning models decreases when moving from within-domain multivariable modeling to cross-domain multivariable modeling. This is particularly important because it shows that models face generalization difficulties when trained and tested on cell line datasets with comparable molecular and pharmacological profiles. Consequently, such models are highly unlikely to generalize to clinical samples when they cannot even generalize to (more similar) preclinical samples.

The PGx guidelines also demonstrate that the AAC is a more suitable measure of drug sensitivity as opposed to the IC50, and CTRPv2 is a more suitable training dataset as opposed to GDSCv1 due to the larger sample size available with RNA sequencing and the difference in the drug screening assay. This is particularly important because employing IC50 and GDSCv1 is the common trend in machine learning for drug sensitivity prediction.

The PGx guidelines also demonstrate the necessity of removing non-solid tissue types from the datasets before any modeling which is often not considered in training models. This is especially true for post-hoc interpretation of feature importance. Our results suggest that models trained on a mixture of these two tissue types actually learn to predict non-solid tumour status, bringing into question whether importance scores will be relevant to the task of drug sensitivity prediction.

The PGx Guidelines demonstrate the impact of feature selection on the training process in both within-domain and cross-domain scenarios and show the utility of the mRMR method. Although we demonstrated that AAC is a more suitable metric compared to IC50, we provided a comprehensive univariable analysis on how to preprocess IC50 and found that log truncating archives the most accurate and consistent results, overall.

Our goal was not to provide a comparative study -- train models to achieve the highest possible performance -- in terms of method, we selected some of the most basic available methods to focus on the importance of data for drug sensitivity prediction. It is likely to improve the reported results by investing more time on hyper-parameter tuning or adopting more complex training schemes or objective functions (particularly for Deep Neural Networks) the results improve slightly. However, our experiments shed light on some of the current issues with machine learning for drug sensitivity prediction.

Although Ridge Regression and Random Forest showed slightly better performance compared to Elastic Net, overall, these methods showed a competitive performance. We focused on Ridge Regression for the majority of the analyses of PGx Guidelines because it is less sensitive to the setting of hyper-parameters than other methods, in particular Deep Neural Networks, which have the highest number of hyper-parameters and are more sensitive to the choice of their values. Therefore, we included Deep Neural Networks for cross-domain analysis but did not consider them for feature selection experiments. We also did not perform within-domain analysis for Deep Neural Networks because of the limited sample size of the nested cross validation for

the high number of parameters of each network and the early stopping regularization which makes the comparison difficult for within-domain analysis. We also utilized feature selection to reduce the input dimensionality (number of genes) and tried focusing only on the L1000 landmark genes [54], or focusing on top genes selected by the Minimum Redundancy--Maximum Relevance (mRMR) method [55]. However, we did not observe any significant difference (Figure S5).

Some of the major limitations of this study are as follows: we assumed a similar concentration range across the studied datasets, this can be an important factor in generalization but the problem is that by focusing on the samples with the same range, we will not have enough samples to train models especially given the high dimensionality of the data. We believe this is a very important factor that should be considered when more samples are available. Similarly, we did not consider multi-omics data because comparable omics data types (mutation, proteomics, copy number aberration, etc.) are not available in the studied pharmacogenomics datasets to investigate the impact of multi-omics data on cross-domain generalization. For PGx Guidelines, we only focused on monotherapy models and did not investigate multi-output or multi-task learning, or drug combination which can be promising future direction, similar to incorporating the chemical structure representation of the drugs as input. Also, we did not consider the pathway transformation of the genes as input features and employed genes themselves. Such transformations can improve the prediction performance [33].

Funding

This work was supported by Natural Sciences and Engineering Research Council (NSERC) and Canadian Institutes of Health Research (CIHR).

ACKNOWLEDGEMENTS

We would like to thank Hossein Asghari and Colin C. Collins (Vancouver Prostate Centre) for their kind support.

REFERENCES

1. Basu A, Bodycombe NE, Cheah JH, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 2013; 154:1151–1161
2. Seashore-Ludlow B, Rees MG, Cheah JH, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015; 5:1210–1223
3. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012; 483:570–575
4. Iorio F, Knijnenburg TA, Vis DJ, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016; 166:740–754
5. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. J. Clin. Oncol. 2013; 31:1803–1805
6. Twomey JD, Brahme NN, Zhang B. Drug-biomarker co-development in oncology - 20 years and counting. Drug Resist. Updat. 2017; 30:48–62
7. Pal SK, Miller MJ, Agarwal N, et al. Clinical Cancer Advances 2019: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology. J. Clin. Oncol. 2019; 37:834–849
8. Marquart J, Chen EY, Prasad V. Estimation of the Percentage of US Patients With Cancer Who Benefit

From Genome-Driven Oncology. JAMA Oncol 2018; 4:1093–1098
9. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012; 483:603–607
10. Klijn C, Durinck S, Stawiski EW, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat. Biotechnol. 2015; 33:306–312
11. Haverty PM, Lin E, Tan J, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. Nature 2016; 533:333–337
12. Gao H, Korn JM, Ferretti S, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat. Med. 2015; 21:1318–1325
13. Mer AS, Ba-Alawi W, Smirnov P, et al. Integrative Pharmacogenomics Analysis of Patient-Derived Xenografts. Cancer Res. 2019; 79:4539–4550
14. Shi R, Radulovich N, Ng C, et al. Organoid Cultures as Preclinical Models of Non-Small Cell Lung Cancer. Clin. Cancer Res. 2020; 26:1162–1174
15. Smirnov P, Kofia V, Maru A, et al. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. Nucleic Acids Res. 2018; 46:D994–D1002
16. Haibe-Kains B, El-Hachem N, Birkbak NJ, et al. Inconsistency in large pharmacogenomic studies. Nature 2013; 504:389–393
17. Safikhani Z, Smirnov P, Freeman M, et al. Revisiting inconsistency in large pharmacogenomic studies. F1000Res. 2016; 5:2333
18. Geeleher P, Gamazon ER, Seoighe C, et al. Consistency in large pharmacogenomic studies. Nature 2016; 540:E1–E2
19. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature 2015; 528:84–87
20. Mpindi JP, Yadav B, Östling P, et al. Consistency in drug response profiling. Nature 2016; 540:E5–E6
21. Niepel M, Hafner M, Mills CE, et al. A Multi-center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. Cell Syst 2019; 9:35–48.e5
22. Bouhaddou M, DiStefano MS, Riesel EA, et al. Drug response consistency in CCLE and CGP. Nature 2016; 540:E9–E10
23. Mammoliti A, Smirnov P, Nakano M, et al. ORCESTRA: a platform for orchestrating and sharing high-throughput pharmacogenomic analyses. Cold Spring Harbor Laboratory 2020; 2020.09.18.303842
24. Luna A, Elloumi F, Varma S, et al. CellMiner Cross-Database (CellMinerCDB) version 1.2: Exploration of patient-derived cancer cell line pharmacogenomics. Nucleic Acids Res. 2021; 49:D1083–D1093
25. Adam G, Rampášek L, Safikhani Z, et al. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 2020; 4:19
26. Chen J, Zhang L. A survey and systematic assessment of computational methods for drug response prediction. Brief. Bioinform. 2020;
27. Güvenç Paltun B, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. Brief. Bioinform. 2019;
28. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol. 2014; 15:R47
29. Geeleher P, Zhang Z, Wang F, et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. Genome Res. 2017; 27:1743–1751
30. Huang EW, Bhope A, Lim J, et al. Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. PLoS Comput. Biol. 2020; 16:e1007607
31. Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics 2016; 32:2891–2895
32. Riddick G, Song H, Ahn S, et al. Predicting in vitro drug sensitivity using Random Forests. Bioinformatics 2011; 27:220–224
33. Costello JC, NCI DREAM Community, Heiser LM, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology 2014; 32:1202–1212
34. He X, Folkman L, Borgwardt K. Kernelized rank learning for personalized drug recommendation. Bioinformatics 2018; 34:2808–2816
35. Mourragui S, Loog M, van de Wiel MA, et al. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. Bioinformatics 2019; 35:i510–i519

36. Mourragui S, Loog M, Vis DJ, et al. PRECISE+ predicts drug response in patients by non-linear subspace-based transfer from cell lines and PDX models. bioRxiv 2020;
37. 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; 16:269–278
38. Rampášek L, Hidru D, Smirnov P, et al. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. Bioinformatics 2019; 35:3743–3751
39. Sakellaropoulos T, Vougas K, Narang S, et al. A Deep Learning Framework for Predicting Response to Therapy in Cancer. Cell Rep. 2019; 29:3367–3373.e4
40. Sharifi-Noghabi H, Zolotareva O, Collins CC, et al. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 2019; 35:i501–i509
41. Sharifi-Noghabi H, Peng S, Zolotareva O, et al. AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics. Bioinformatics 2020; 36:i380–i388
42. Snow O, Sharifi-Noghabi H, Lu J, et al. BDKANN - Biological Domain Knowledge-based Artificial Neural Network for drug response prediction. Cold Spring Harbor Laboratory 2020; 840553
43. Kuenzi BM, Park J, Fong SH, et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell 2020; 38:672–684.e6
44. Zhu Y, Brettin T, Evrard YA, et al. Ensemble transfer learning for the prediction of anti-cancer drug response. Sci. Rep. 2020; 10:18040
45. Born J, Manica M, Oskooei A, et al. Paccmann rl: Designing anticancer drugs from transcriptomic data via reinforcement learning. International Conference on Research in Computational Molecular Biology 2020; 231–233
46. Ma J, Fong SH, Luo Y, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. Nature Cancer 2021;
47. Jiang Y, Rensi S, Wang S, et al. DrugOrchestra: Jointly predicting drug response, targets, and side effects via deep multi-task learning. Cold Spring Harbor Laboratory 2020; 2020.11.17.385757
48. Jang INS, Neto EC, Guinney J, et al. SYSTEMATIC ASSESSMENT OF ANALYTICAL METHODS FOR DRUG SENSITIVITY PREDICTION FROM CANCER CELL LINE DATA. Biocomputing 2014 2013; 63–74
49. Bray NL, Pimentel H, Melsted P, et al. Erratum: Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 2016; 34:888
50. Smirnov P, Safikhani Z, El-Hachem N, et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. Bioinformatics 2016; 32:1244–1246
51. Sebaugh JL. Guidelines for accurate EC50/IC50 estimation. Pharm. Stat. 2011; 10:128–134
52. Yao F, Tonekaboni SAM, Safikhani Z, et al. Tissue specificity of in vitro drug sensitivity. J. Am. Med. Inform. Assoc. 2017; accepted:
53. Gerdes H, Casado P, Dokal A, et al. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. Nat. Commun. 2021; 12:1850
54. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 2017; 171:1437–1452.e17
55. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 2005; 3:185–205

Supplementary Material

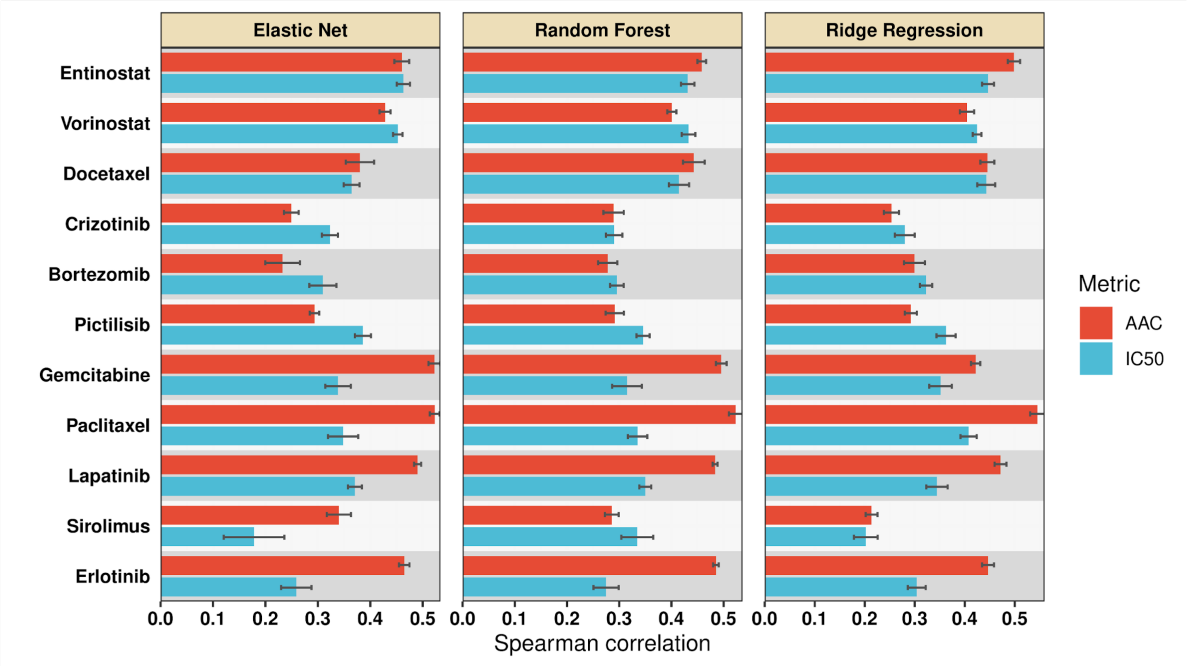

Figure S1. **Mean Spearman correlation (over 10 runs) for different multivariable methods in within-domain analysis trained to predict AAC or IC50. The results indicated that AAC outperforms IC50.**

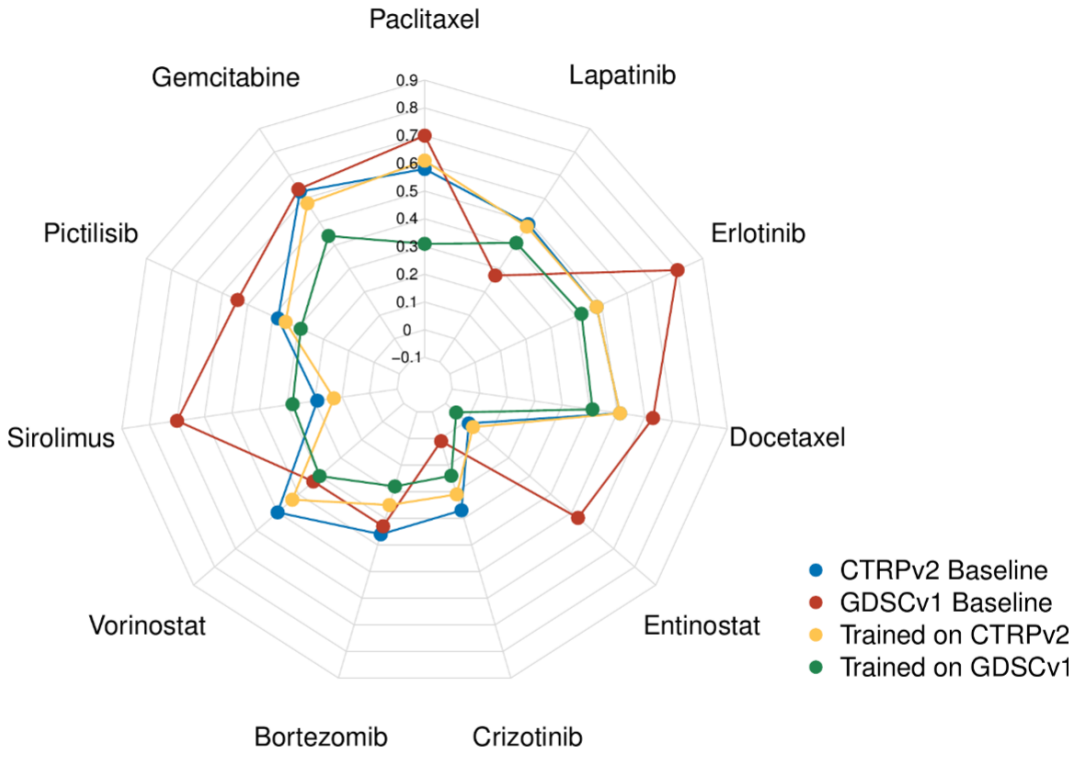

Figure S2. **Comparison of CTRPv2 and GDSCv1 as training datasets for Ridge Regression models in terms of Spearman correlation in cross-domain analysis to predict AAC on gCSI. For the majority of the drugs, models trained on CTRPv2 (Yellow) achieved better performance compared to those trained on GDSCv1 (Green) even though GDSCv1 has higher baseline correlation (correlation of AAC values between cell lines in common between GDSCv1/CTRPv2 and gCSI--CTRPv2 baseline in Blue and GDSCv1 baseline in Red).**

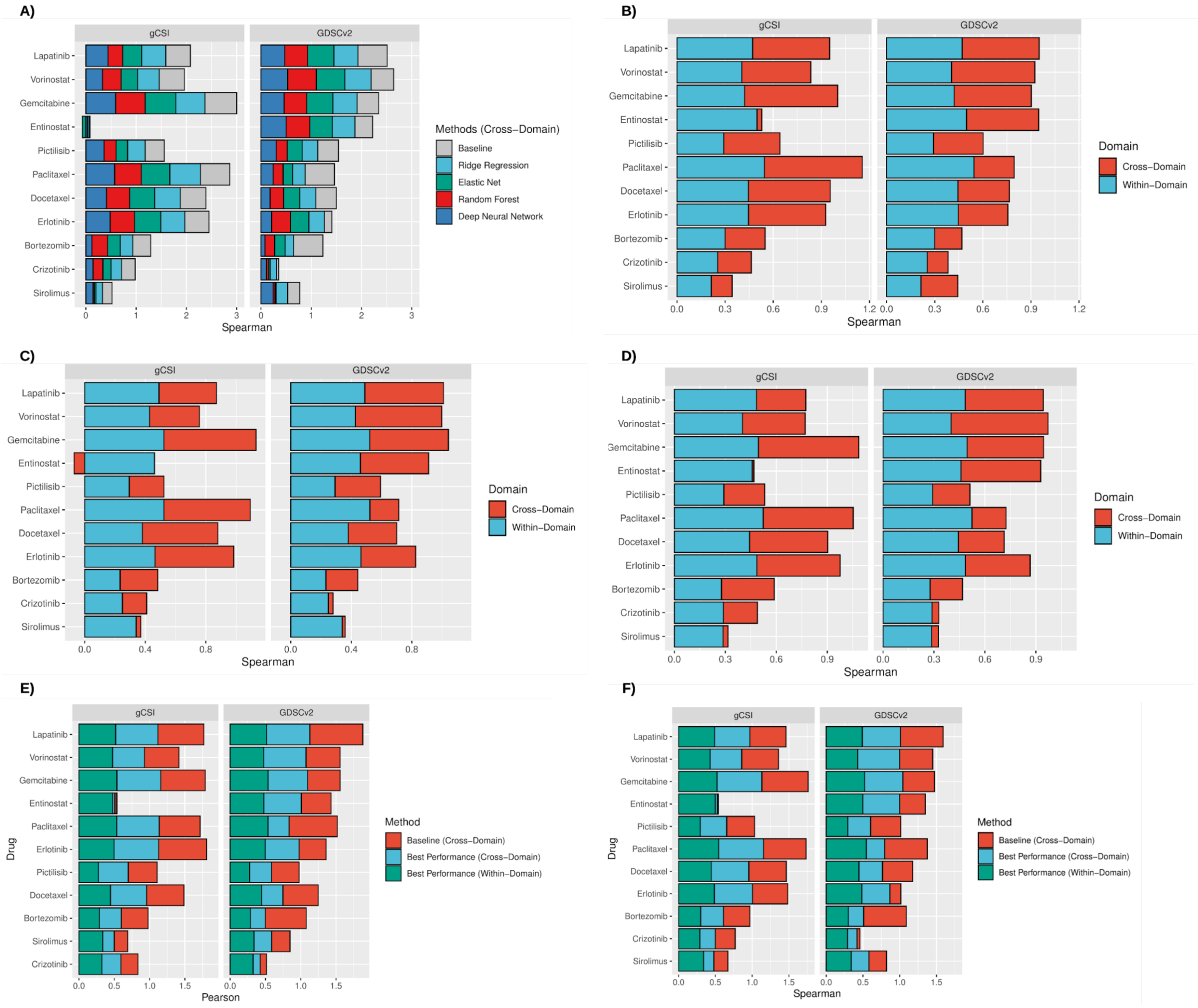

Figure S3. **Comparison of multivariable methods in terms of Spearman correlation in cross-domain analysis trained on CTRPv2 to predict AAC, tested on GDSCv2 and gCSI (A). Comparison of the within-domain performance of Ridge Regression (B), Elastic Net (C), and Random Forest (D) to their cross-domain performance in terms of Spearman correlation. Comparison of the best performing multivariable method for each drug in within-domain and cross-domain in terms of Pearson (E) and Spearman correlation (F).**

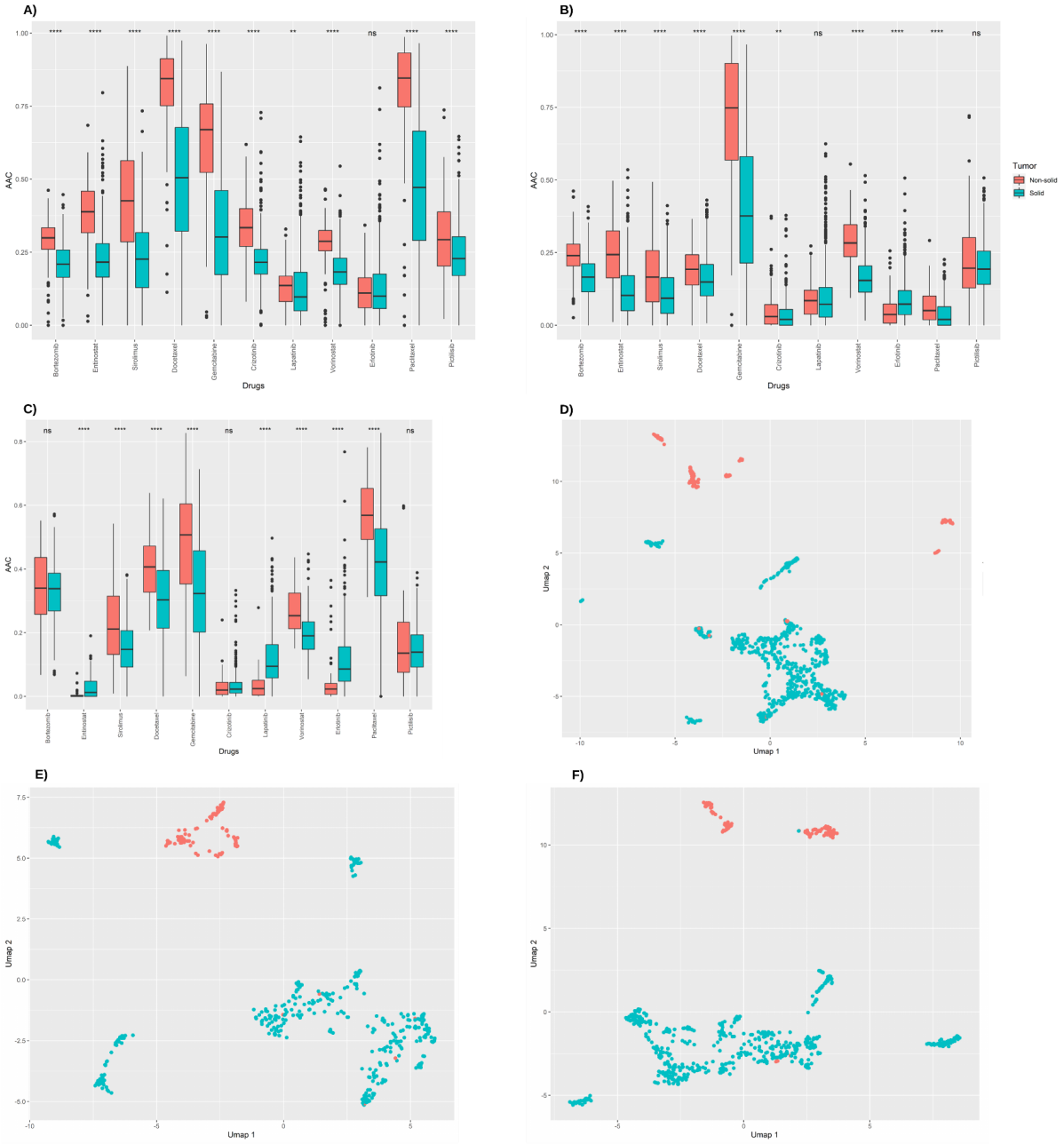

Figure S4. **Comparison of the AAC values of solid (Blue) and non-solid (Red) samples across CTRPv2 (A), GDSCv2 (B), and gCSI (C) for 11 drugs in common with them. Comparison of the molecular profiles of solid (Blue) and non-solid (Red) samples via UMAP plot across CTRPv2 (D), GDSCv2 (E), and gCSI (F).**

Feature Selection

To see if reducing the number of input genes improves the prediction performance, we investigated two feature selection approaches: II) focusing on L1000 landmark genes [54], and focusing on top 200 genes selected by the Minimum Redundancy--Maximum Relevance (mRMR) method [55]. The idea of mRMR method is to select a subset of features (genes) that have high correlations with the output (drug sensitivity) and minimum correlations among themselves. We trained Ridge Regression models on CTRPv2 using the reduced features and tested them on GDSCv2 and gCSI in a cross-domain fashion. The experiments of L1000 landmark genes (961 genes were in common across applied datasets) and mRMR top 200 genes were implemented in R via the mRMRe (version 2.1.0) and Caret packages (version 6.0.86). On average, we did not observe a significant difference between feature selection cross-domain results and the cross-domain results obtained from protein coding genes for AAC (Figure S5).

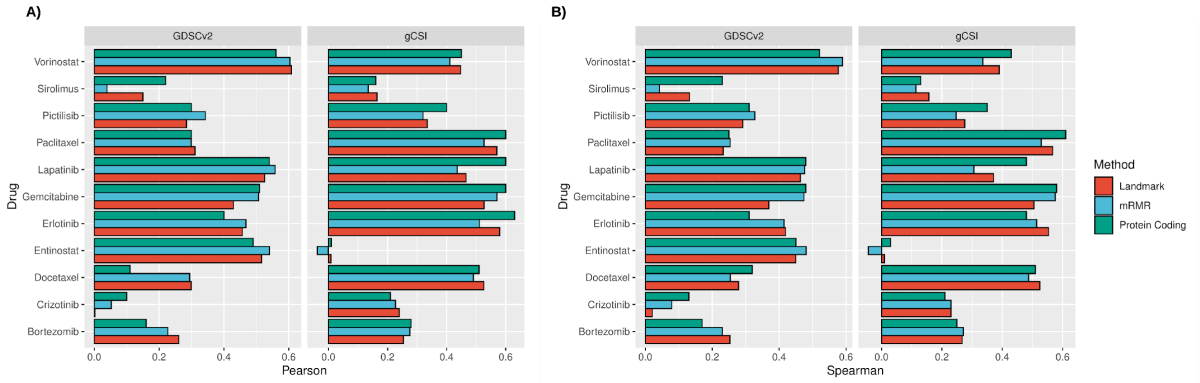

Figure S5. Comparison of top 200 genes selected by mRMR (Blue) and L1000 landmark genes (Red) with protein coding genes (Green) on GDSCv2 and gCSI for AAC in terms of Pearson correlation (A) and Spearman correlation (B).