Machine Learning for Pharmacogenomics and Personalized Medicine: A Ranking Model for Drug Sensitivity Prediction

Shahabeddin Sotudian and Ioannis Ch. Paschalidis, *Fellow, IEEE*

Abstract—It is infeasible to test many different chemotherapy drugs on actual patients in large clinical trials, which motivates computational methods with the ability to learn and exploit associations between drug effectiveness and patient characteristics. This work proposes a machine learning approach to infer robust predictors of drug responses from patient genomic information. Rather than predicting the exact drug response on a given cell line, we introduce an elastic-net regression methodology to compare a drug-cell line pair against an alternative pair. Using predicted pairwise comparisons we rank the effectiveness of different drugs on the same cell line. A total of 173 cell lines and 100 drug responses were used in various settings for training and testing the proposed models. By comparing our approach against twelve baseline methods, we demonstrate that it outperforms the state-of-the-art methods in the literature. In contrast to most other methods, the algorithm is able to maintain its high performance even when we use a large number of drugs and few cell lines.

Index Terms—Drug sensitivity prediction, personalized medicine, elastic net regression, cancer, ranking, score function.

1 Introduction

Performing drug screening and selecting appropriate personalized treatment based on individual genomic or proteomic profiles is one of the paramount goals of precision medicine.

Since the effectiveness of each drug can be varied among patients, finding the right drug for each cancer patient is quite a challenging task [1]. One possible option is to assess the efficacy of drugs using large clinical trials. This path makes it possible to capture most of the pertinent biological features of a patient [2]. However, this approach is impractical for several reasons: (i) large clinical trials are time-consuming and expensive; (ii) a clinical trial typically evaluates one (or very few drugs) at a time; and (iii) even a large clinical trial may not include enough patients to cover all cancer genomic variations.

Cell lines contain a large number of the molecular features of tumors; as a result, they are expected to reflect the properties (mutation status, gene expression, drug sensitivity, and so on) of the original cancer type from which they were cultured [3]. Thus, they provide practical preclinical models in order to analyze strategies for predictive marker development [4]. Consequently, an alternative to clinical trials is drug response prediction using large-scale screenings of cancer cell lines against libraries of pharmacological compounds [5]. This can lead to genomic predictors of drug responses from large panels of cancer cell lines [3], [6].

Hansch et al. [7] were among the first who revealed the existence of mathematical relations between the biological activity of a chemical compound and its physicochemical properties. Pan-cancer repositories (e.g., the molecular analysis for a therapy choice trial at the National Cancer Institute, the Cancer Genome Atlas) provide the foundation for joint analysis of cancer cell lines and their drug responses. Furthermore, over the past few years, two important studies, namely the cancer cell line encyclopedia [3], and the genomics of drug sensitivity projects [8] revealed the applicability of machine learning algorithms in predicting drug response based on panels of cancer cell lines.

Different methodologies have been utilized in the literature for drug sensitivity prediction based on genomic data. Barretina et al. [3] employed elastic net regression [9] to predict drug sensitivity for the *Cancer Cell Line Encyclopedia (CCLE)*. Geeleher et al. [10] used pre-treatment baseline gene expression data to predict the chemotherapeutic response in patients. They applied a ridge regression model to predict drug response for breast cancer cell lines using baseline gene expression data. This method was compared against several methods, such as nearest shrunken centroids, principal component regression, *Least Absolute Shrinkage and Selection Operator (LASSO)* regression [11], elastic net regression, and random forests [12]. They observed that ridge regression yielded the best performance. However, other studies comparing multiple algorithms on a drug sensitivity database have observed that random forests perform better

- *S. Sotudian is with the Department of Electrical and Computer Engineering, Division of Systems Engineering, Boston University, Boston, MA, 02215, USA.*
  E-mail: sotudian@bu.edu
- *I. C. Paschalidis is with the Department of Electrical and Computer Engineering, Division of Systems Engineering, Department of Biomedical Engineering, and Faculty of Computing & Data Sciences, Boston University, Boston, MA, 02215, USA.*
  E-mail: yannisp@bu.edu, sites.bu.edu/paschalidis.

The research was partially supported by the NSF under grants DMS-1664644, CNS-1645681, and IIS-1914792, by the ONR under grant N00014-19-1-2571, by the NIH under grants R01 GM135930 and UL54 TR004130, by the DOE under grant DE-AR-0001282, and by the Boston University Institute for Health System Innovation & Policy.

than ridge regression or other regularized linear regression approaches [13]. Undoubtedly, the random forest algorithm has been one of the top-performing algorithms in drug sensitivity prediction, and this has been proven in multiple other drug sensitivity studies [14], [15], [16], [17]. Nevertheless, in all of these studies, the authors did not consider the effect of molecular feature data, the type of drug used, the sensitivity of the response (discrete or continuous), and methods for summarizing compound sensitivity values.

In a comprehensive comparative study in drug sensitivity prediction, Jang et al. [6] considered these modeling factors to compare various machine learning algorithms, including principal component regression, partial least square regression, least squares support vector machine regression with linear kernels, random forests, LASSO, ridge regression, and elastic net regression. They considered more than 110,000 various models based on a multifactor experimental design. Their analysis suggested that elastic net or ridge regression will most likely yield the most accurate predictors. However, some studies argue that the existing regression-based methods for drug selection may sacrifice performance on very few but sensitive drugs to achieve better performance on the majority of insensitive drugs [18]. Therefore, when using the regression-based models to predict how a cell line responds to sensitive drugs, the prediction could lead to incorrect drug selection or prioritization [18].

Motivated by the high performance of regression-based methods in drug sensitivity prediction, we propose a novel elastic-net-based regression that utilizes gene expression features and drug sensitivity data to build a predictor. Most of the conventional regression methods learn a vector of coefficients for each cell line or drug and use these vectors to predict the sensitivity values. We believe this is the main reason that may lead to sacrificing performance on very few but sensitive drugs to achieve better performance on the majority of insensitive drugs. Moreover, because the number of cell lines is significantly less than the number of features, these approaches usually lead to "overfitting." In addition, they fail to capture the drug-cell line relation across various cell lines. As a result, they are unable to effectively prioritize sensitive drugs over insensitive drugs.

Our method uses regularized regression to model the difference in sensitivity of a given drug acting on different cell lines or different drugs acting on the same cell line. Specifically, we do not predict the exact drug response for each drug, but rather develop a model that estimates the difference in sensitivity obtained by comparing two different drug-cell line pairs. This score function can compare drug sensitivities for a given cell line and then use these pairwise comparisons to rank the drugs. Using this model, we can rank the sensitivity of different drugs acting on a novel cell line. The use of regularization as part of the regression is motivated by the earlier work we reviewed and recent results on the connection between regularization and robustness to the potential presence of outliers [19], [20].

The remainder of this paper is organized as follows. Section 2 outlines our method, presents the data we used, and discusses model training and performance evaluation. Section 3 presents our main results and how our method compares to alternatives. Section 4 discusses the results and draws some conclusions.

Notational conventions: We use boldfaced lowercase letters to denote vectors, ordinary lowercase letters to denote scalars, boldfaced uppercase letters to denote matrices, and calligraphic capital letters to denote sets. All vectors are column vectors. For space saving reasons, we write $\mathbf{x} = (x_1, \ldots, x_n)$ to denote the column vector $\mathbf{x} \in \mathbb{R}^n$. For any matrix $\mathbf{A}$, we let a_{ij} denote its (i, j) element, $\mathbf{a}_i$ the ith row and, with some abuse of our conventions, $\mathbf{A}_j$ the jth column. $\mathbf{I}$ denotes the identity matrix. We use prime to denote the transpose of a vector, and $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for the ℓ_p norm, where $p \geq 1$. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\|\mathbf{A}\|_p$ will denote the ℓ_p norm of a vectorized form of the matrix, i.e., $\|\mathbf{A}\|_p = (\sum_{i=1}^n \sum_{j=1}^m |a_{ij}|^p)^{1/p}$. We also use $\mathbf{e}_i$ to denote the ith unit vector, a vector of all zeros except the ith element which is set to 1.

2 Materials and Methods

2.1 Model description

We are given a matrix $\mathbf{S} \in \mathbb{R}^{N_C \times N_G}$, where each row corresponds to one of N_C cell lines (patients). Each cell line is characterized by a gene expression vector containing N_G features. We are also given a matrix $\mathbf{R} \in [0, 1]^{N_C \times N_D}$ of drug responses to N_D different drugs; each row corresponds to a cell line and each column lists the response of a specific drug to all cell lines. Drug responses are in $[0, 1]$ and a higher response implies higher sensitivity of the cell line to the drug (i.e., a more effective drug). Our ultimate goal is to rank the N_D drugs based on their response, so that the most effective ones are ranked on top. Importantly, we care about the relative ordering of the most effective drugs; predicting the exact value of the drug response is of no consequence.

In our specific application, we do not have any information about the drugs (e.g., their mechanism of action, physical and chemical properties, metabolism, therapeutics, and toxicity). We just know the name of each drug, which makes the prediction even more difficult. Therefore, we use so-called "one-hot" encoding to represent drugs; i.e., for each of the N_D drugs we create an indicator variable taking values in $\{0, 1\}$. Application of the ith drug will simply be represented by the ith unit vector. If we had access to the drug-specific features, we can easily substitute the one-hot encoding vector with such features.

Define now the $(N_D + N_G)$-dimensional vector $\mathbf{p}_{ij} = (\mathbf{e}_i, \mathbf{s}_j)$, $i = 1, \ldots, N_D$, $j = 1, \ldots, N_C$, formed as the concatenation of the one-hot vector representing drug i and the j row vector of $\mathbf{S}$ containing the gene expression for cell line j. Per our notational conventions, r_{ij} denotes the (i, j) element of the response matrix $\mathbf{R}$, that is, the response of drug j in cell line i. Denote by $N = N_C N_D$ the number of all possible combinations of a drug with a cell line. Define the matrix $\mathbf{X} \in \mathbb{R}^{N\times(N_D+N_G)}$ with rows $\mathbf{p}_{11}, \mathbf{p}_{21}, \ldots, \mathbf{p}_N$ corresponding to drug-cell line pairs and let $\mathbf{y} = (r_{11}, r_{21}, \ldots, r_N)$ be the vector of corresponding responses. For any $i = 1, \ldots, N$, $(\mathbf{x}_i, y_i)$ will consist of the ith row of $\mathbf{X}$ and the ith element of $\mathbf{y}$.

We will view $\{(\mathbf{x}_i, y_i);\ i = 1, \ldots, N\}$ as a training set and we are interested in training a model that predicts the difference in sensitivity between two different drug-cell line

combinations $\mathbf{x}_i, \mathbf{x}_j$. We start by defining appropriate labels. Define:

$$a_{ij} = \xi_{ij}(y_i - y_j) \times 100, \qquad i, j = 1, \ldots, N,$$

where

$$\xi_{ij} = \begin{cases} \phi_1, & \text{when } i \text{ and } j \text{ refer to same drug in different cell lines,} \\ \phi_2, & \text{when } i \text{ an } j \text{ refer to different drugs in the same cell line,} \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

and ϕ_1, ϕ_2 are positive constants. Let also

$$\psi_{ij} = \begin{cases} 1, & \text{when } i \text{ and } j \text{ refer to the same drug in different cell lines or different drugs in the same cell line,} \\ 0, & \text{otherwise.} \end{cases}$$

The training problem amounts to finding a matrix $\mathbf{W}$ that solves the following optimization problem

$$\min_{\mathbf{W}} \frac{1}{Z} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} [\psi_{ij}(\mathbf{x}_i' \mathbf{W} \mathbf{x}_j - a_{ij})]^2 + \lambda_1 \|\mathbf{W}\|_2^2 + \lambda_2 \|\mathbf{W}\|_1, \quad (2)$$

where $Z = N(N_D + N_C - 2)$ and $\lambda_1, \lambda_2 \geq 0$ are regularization parameters.

Having solved the regression problem in (2), we can rank the drugs on how they affect a new cell line using the following procedure. Let $\mathbf{s}$ be the gene expression vector of the new cell line. For any $i = 1, \ldots, N_D$, we compare the drug-cell line pairs $\mathbf{g}_i = (\mathbf{e}_i, \mathbf{s})$ against themselves and other pairs $\mathbf{x}_n$, $n = 1, \ldots, N$, already seen in the training set. Specifically, the score T_i of drug $i = 1, \ldots, N_D$, on the cell line $\mathbf{s}$ is obtained by:

$$T_i = \frac{1}{(N_D + N_C - 1)} \left(\sum_{\substack{j=1 \\ j \neq i}}^{N_D} \mathbf{g}_i' \mathbf{W} \mathbf{g}_j + \sum_{n=1}^{N} \psi_{in} \mathbf{g}_i' \mathbf{W} \mathbf{x}_n \right). \quad (3)$$

Ranking these scores T_i provides a ranking of the effectiveness of the drugs on the new cell line $\mathbf{s}$.

We present an example to illustrate the scoring process. Suppose that we have 3 drugs D_1, D_2, and D_3 and 4 cell lines C_1, C_2, C_3, and C_4. The structure of the training matrix $\mathbf{X}$ is shown in Figure 1. The arrows in this figure depict the comparisons our model performs during training for D_1 in C_1. To rank the drugs in a new cell line C_{Test}, use (3) to compute T_i, $i = 1, 2, 3$. Figure 2 demonstrates the comparisons (3) performs to evaluate the score for D_1.

2.2 Data sets

We designed, evaluated, and trained our method using the cell line data and drug sensitivity data from the *Cancer Cell Line Encyclopedia (CCLE)* [21] and the *Cancer Therapeutics Response Portal (CTRP v2)* [22]. We used the "Act Area" (the area above the fitted dose-response curve) to quantify drug sensitivity. With this metric, lower response indicates higher drug sensitivity. It worth mentioning that approximately 20% of the drug responses are missing; this portion of the data were excluded from our analysis.

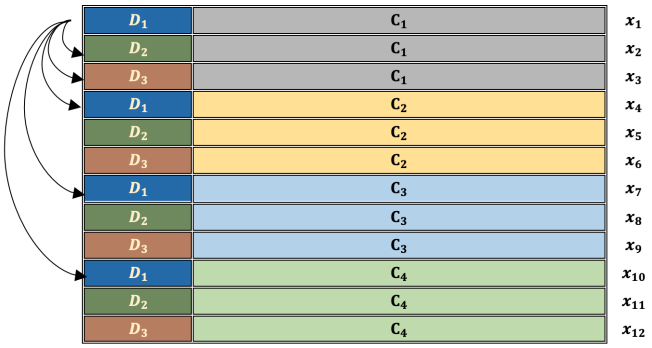

Fig. 1: The structure of $\mathbf{X}$ and the comparisons our model will carry out for D_1 in C_1.

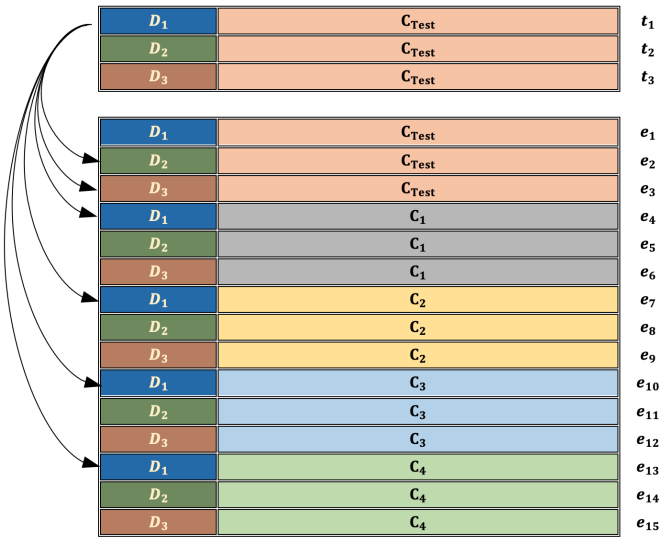

Fig. 2: The process of calculating the score for D_1 in a new cell line.

2.3 Gene selection scheme

Each gene in our data set has approximately 20,000 features. A snapshot of all the transcriptional activity in a cell line can be obtained by gene expression microarrays. Please refer to Appendix A and [23] for more information. Since the number of features is significantly higher than the number of training samples, using all these features may result in overfitting. Instead, we selected an appropriate subset of features to use for all our experiments. To that end, and consistent with the discussion in Section 1, we used LASSO regression to select a subset of features.

Specifically, consider the matrix $\mathbf{X}$ and the response vector $\mathbf{y}$ defined in Section 2.1. For the rows $(\mathbf{e}_i, \mathbf{s}_j)$, $i = 1, \ldots, N_D$, $j = 1, \ldots, N_C$, of $\mathbf{X}$ define a common coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $\boldsymbol{\beta}_1 \in \mathbb{R}^{N_D}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{N_C}$. To develop a regression-based model for predicting $\mathbf{y}$, we solve the following loss minimization problem:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}_2\|_1, \quad (4)$$

where $\lambda > 0$ is a scalar that modulates the strength of the regularizer. Gene features were eliminated using a recursive feature elimination procedure. In particular, we solve (4) us-

ing cross-validation to select λ. We eliminate 5% of the gene features whose corresponding coefficient in $\boldsymbol{\beta}_2$ is among the 5% smaller absolute values. We reformulate (4) using the remaining gene features and repeat this process while the validation loss keeps decreasing. Further details of the gene selection scheme and the list of the selected genes can be found in Appendix A.

2.4 Converting the model to elastic net regression

The following Lemma is almost immediate; hence, we skip the proof. It establishes that Problem (2) can be reformulated as a standard elastic net regression.

Lemma 2.1 *Formulation (2) can be converted to the following elastic net regression problem:*

$$\min_{\mathbf{w}} \frac{1}{Z} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} [\psi_{ij}(\mathbf{v}_{ij}^{'}\mathbf{w} - a_{ij})]^2 + \lambda_1 \|\mathbf{w}\|_2^2 + \lambda_2 \|\mathbf{w}\|_1, \quad (5)$$

where $\mathbf{v}_{ij}$ *is the vectorization of* $\mathbf{x}_i\mathbf{x}_j'$*, and* $\mathbf{w}$ *is the vectorization of* $\mathbf{W}$*.*

2.5 Iterative thresholding for complexity reduction

Regularizing with the ℓ_1-norm is known to induce sparsity in the coefficient vector $\mathbf{w}$. To reduce the complexity of our model, we can set to zero some elements of $\mathbf{w}$ that appear less important in making a prediction; this has also the effect of reducing running time and providing a higher level of interpretability [24] to the model. To that end, we used an iterative procedure similar to the one mentioned in Section 2.3.

More specifically, in each iteration, we use cross-validation to select λ_1, λ_2 in (5) that minimize the loss on the validation set. Then, we select 10% of the elements of $\mathbf{w}$ with the smallest absolute value and set them to zero. We continue iterating in this fashion until we do not see any considerable decrease in the (validation) performance of our model. The end result will be a parameter vector $\mathbf{w}$ with much less non-zero elements than the total number of $(N_D+N_C)^2$ elements. In this way, we not only prevent over fitting, but also decrease the complexity of the model.

2.6 Closed form solution

After fixing a number of elements of $\mathbf{w}$ to zero using the approach described in Section 2.5, we remove the ℓ_1-norm penalty from the problem formulation (5). The resulting problem is equivalent to ridge regression. Let $\mathbf{V} \in \mathbb{R}^{N(N-1)\times(N_D+N_G)^2}$ be the matrix with rows $\{\psi_{ij}\mathbf{v}_{ij};\ i,j = 1,\ldots,N,\ i \neq j\}$. Define also the vector $\boldsymbol{\alpha} = (\psi_{ij}a_{ij};\ i,j = 1,\ldots,N,\ i \neq j) \in \mathbb{R}^{(N_D+N_G)^2}$. Then, Problem (5), with $\lambda_2 = 0$, is written as:

$$\min_{\mathbf{w}} \frac{1}{Z}\|\mathbf{V}\mathbf{w} - \boldsymbol{\alpha}\|_2^2 + \lambda_1\|\mathbf{w}\|_2^2. \quad (6)$$

The following lemma provides a closed-form solution to Problem (6). A proof is provided in Appendix B.

Lemma 2.2 *The closed form solution for Problem (6) can be written as follows:*

$$\mathbf{w} = (\mathbf{V}'\mathbf{V} + \lambda_1 Z\mathbf{I})^{-1}\mathbf{V}'\boldsymbol{\alpha}. \quad (7)$$

Notice that the matrix $\mathbf{V}'\mathbf{V}$ is positive semi-definite and, for large enough λ_1, $\mathbf{V}'\mathbf{V} + \lambda_1 Z\mathbf{I}$ is positive definite and thus invertible. When $\mathbf{w}$ is dense, the asymptotic time complexity of the closed form solution would be $O((N_D + N_G)^6 + N(N-1)(N_D + N_G)^4)$. However, as we discussed in Section 2.5, we can greatly decrease the number of non-zero elements of $\mathbf{w}$ and instead solve for the remaining non-zero elements of $\mathbf{w}$, where the matrix to be inverted has dimension $n \times n$ with n being the number of non-zero elements of $\mathbf{w}$. A schematic of the proposed model is summarized in Figure 3.

2.7 Experimental Setting

The data were standardized so that variables lie between zero and one. The dataset was further split into a training and a test set. Details of the pre-processing steps can be found in Appendix C.

We applied iterative thresholding on the training set and found the non-zero elements of $\mathbf{w}$. Performance was evaluated on the test set. To calculate some of the metrics we classified each drug into two classes – sensitive/insensitive – for each cell line. Specifically, a fixed percentile of all drug response values in the training set for each cell line was used as a threshold to determine drug sensitivity in that cell line.

2.8 Performance metrics

To evaluate the accuracy of our approach, we used two ranking metrics, namely $AH@k$, and CI^s. The first metric is the average-hit at k, which is the average number of sensitive drugs that are ranked among the top k of a ranking list [25], [26]. This metric is important because in this specific application we care more about sensitive drugs. It evaluates the ability of a model to place sensitive drugs on top of insensitive drugs. However, a high value for $AH@k$ does not necessarily guarantee that the ordering among the top-ranked drugs is correct. For that purpose, we consider the concordance index CI. The concordance index measures the ratio of correctly ordered drug pairs among all possible pairs [25]. In other words, it measures whether the ordering structure of a ranking list is close to its ground truth or not. Since we care more about the sensitive drugs, we use CI to evaluate the ranking structures among only sensitive drugs and denote it by CI^s. If we use $d_m \succ d_n$ to represent that drug d_m is more sensitive, and thus ranked higher than drug d_n, CI^s can be defined as follows:

$$CI^s = \frac{1}{|d_m \overset{\mathsf{T}}{\succ} d_n|} \sum_{d_m \succ d_n} \mathbb{I}(d_m \overset{\mathsf{P}}{\succ} d_n),$$

where $d_m \overset{\mathsf{T}}{\succ} d_n$ represents that drug d_m is ranked higher than drug d_n based on the ground truth ranking, $d_m \overset{\mathsf{P}}{\succ} d_n$ represents that drug d_m is ranked higher than drug d_n based on the predicted ranking list of sensitive grants, $\mathbb{I}$ is the indicator function, and $|\cdot|$ denotes the cardinality of a set [18].

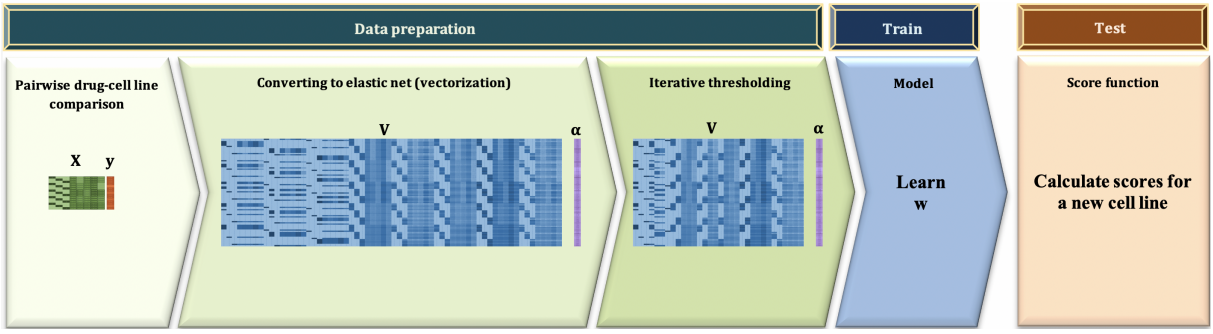

Fig. 3: Schematic overview of the proposed method. In the first step (Data preparation), the model creates the pairwise drug-cell line comparison matrix using one-hot vectors of drugs and the gene expressions for training cell lines. Then, the proposed model converts the problem to an elastic net regression by changing and expanding the pairwise drug-cell line comparison matrix. Finally, iterative thresholding reduces the complexity and running time of the model. In the training step, $\mathbf{w}$ will be learned using Equation (7). Finally, in the last step, the scores of drugs for a new cell line are calculated using Equation (3). Eventually, the ranking of the drugs in that cell line is obtained as a permutation π which sorts the drugs in the decreasing order of their score.

2.9 Baseline methods

We compared the proposed method against twelve state-of-the-art approaches for drug sensitivity prediction including *Elastic Net (EN)* [9], *Kernel Ridge Regression (KRR)* [27], *Bayesian Multitask Multiple Kernel Learning (BMTMKL)* [28], *Partial Least Square* regression (PLS) [29], *Principal Component Regression (PCR)* [24], *Sparse Principal Component Regression (SPCR)* [30], *Gaussian Process Regression (GPR)* [31], LASSO [11], *Robust Support Vector Regression (RSVM)* [32], *LambdaMART* [33], *Neural Network Regression (NNR)* [34] and *Deep neural network (DNN)* [35]. For more information about these algorithms, please refer to the above-mentioned references. Furthermore, the details of LambdaMART and NNR can be found in the supplementary materials. In general, the typical practice for regression models is to design supervised predictive models for each drug based on the gene expression profiles of cell lines. Then, the ranking of the drugs in a specific cell line is obtained as a permutation, sorting the drugs in the decreasing order of predicted drug response values.

2.10 Hyper-parameter optimization

There are three hyper-parameters in our model, namely ϕ_1, ϕ_2 (cf. (1)) and λ_1 (cf. (2)). We tuned these hyper-parameters with a grid search on the training set, using cross-validation. To that end, we tried different values to find the best values of these hyper-parameters. Similarly, we conducted a grid search for each of the parameters of the baseline methods. The details of the parameter-tuning procedure can be found in Appendix D.

3 Results

We consider three case studies for ranking cancer drugs acting on cancer cell lines: (i) lung cancer cell lines, (ii) blood cancer cell lines, and (iii) myeloma and lymphoma cell lines. In each case, we compare the performance of the proposed method on ranking drugs with baseline methods introduced in Sec. 2.9. The source code is provided on GitHub. [1]

In most of the prior work in the literature, all the cell lines and drugs in CCLE and CTRP v2 (see Sec. 2.2) were used for ranking. In this work, we elected to develop separate drug lists for each of the cancer cases studies outlined above. One reason is that cancer care remains specialized to the type of tissue where the primary cancer developed. The drawback is that the ranking problem is more challenging, since, in each case, we work with fewer cell lines. As a result, part of the goal is to assess whether the proposed model is able to overcome this challenge. From a histological angle perspective, there are six broad categories of cancers based on tissue type: carcinoma, sarcoma, myeloma, leukemia, lymphoma, and mixed types [36]. Moreover, depending on the primary site of origin, cancers may be of specific types such as lung cancer, liver cancer, breast cancer, etc.

In Case Study 1, we used the cell lines which were derived from the lung cancer tumors (carcinoma group). After removing missing data, we have 51 cell lines and 100 drugs. According to our gene selection results, 88 genes are selected as a minimal subset of all genes. The details of the gene selection scheme and the list of the selected genes can be found in Appendix A. We present two different experiments, one involving the ranking of 50 drugs, and the other involving 100 drugs. In all experiments, we set the sensitivity threshold so that the number of sensitive drugs in that experiment is exactly 10 drugs. Therefore, even when we have 100 drugs, we would have 10 sensitive drugs and we will measure the ability of the models to rank these sensitive drugs at the very top of the ranking list. It worth mentioning that in the second experiment, we added 50 new drugs to the existing drugs in the first experiment. However, the sensitive drugs in the first experiment are not exactly the same as the sensitive drugs in the second experiment because in each experiment, the threshold is defined based on a different set of drugs.

In Table 1, we compare the proposed method with the baseline methods for 50 and 100 drugs. In this and the tables

1. https://github.com/noc-lab/Drug-Response-Prediction.

TABLE 1: Performance Comparison of Ranking Methods for 50 and 100 Drugs (Lung Cancer Cell Lines).

	Methods	Parameters				AH@5	AH@10
50 Drugs	**RSVM**	$\Omega = 0.1$	$\gamma = 100$	Kernel= polynomial		**3.7**	**6.4**
	LASSO	$\lambda = 10$				3.6	6
	KRR	$\lambda = 10^5$	$\sigma^2 = 10$	Kernel= RBF		3.6	6
	EN	$\alpha = 0.1$		$\lambda = 100$		3.6	6
	BMTMKL	$\alpha_b = 10^{-10}$		$\beta_b = 10^{10}$		3.3	6.1
	PLS	PLScomp = 28				3.4	5.2
	PCR	Pcomp = 32				3.4	5.4
	SPCR	Pcomp = 29		Card = 25		3.4	**6.4**
	GPR	Kernel= Matern 5/2				**3.7**	6.1
	NNR	Epochs= 50		Batch size= 5		3.6	**6.4**
	LambdaMART	$\tau = 0.01$	$n_T = 100$	$D_{max} = 5$	$Sub = 0.7$	**3.8**	**6.4**
	DNN	Epochs= 100	Batch size= 5	$Drop = 0.3$	$\ell_2 = 0.001$	3.6	**6.4**
	Proposed Model	$\phi_1 = 2.5$	$\phi_2 = 1$	NonZ= 5010	$\lambda_1 = 1$	**3.8**	**6.6**
100 Drugs	**RSVM**	$\Omega = 0.1$	$\gamma = 1$	Kernel= polynomial		3.4	**5.4**
	LASSO	$\lambda = 100$				3.4	5
	KRR	$\lambda = 10$	$\sigma^2 = 10$	Kernel= RBF		3.4	5.2
	EN	$\alpha = 0.1$		$\lambda = 10$		3.4	**5.4**
	BMTMKL	$\alpha_b = 10^{-10}$		$\beta_b = 10^{10}$		3.2	5.2
	PLS	PLScomp = 26				3	4.5
	PCR	Pcomp = 22				3.2	4.9
	SPCR	Pcomp = 28		Card = 75		**3.5**	4.9
	GPR	Kernel= exponential				**3.5**	5
	NNR	Epochs= 100		Batch size= 5		3.4	5.1
	LambdaMART	$\tau = 0.01$	$n_T = 1000$	$D_{max} = 5$	$Sub = 1$	3.2	5.3
	DNN	Epochs= 100	Batch size= 5	$Drop = 0.1$	$\ell_2 = 0.01$	3.2	4.9
	Proposed Model	$\phi_1 = 4$	$\phi_2 = 1$	NonZ= 2694	$\lambda_1 = 1$	**3.8**	**5.5**

corresponding to the other two case studies, the last two columns report the average AH@5 and AH@10 for testing data that correspond to the best parameter combinations. Bold and underlined numbers indicate the best performance among all methods for each metric. Bold numbers demonstrate the second-best performance among all methods for each metric. In all tables, Ω denotes the kernel parameter for RSVM, NPComp is the number of principle components, PLScomp is the number of PLS components, Pcomp is the number of principle components, Card is the desired number of non-sparse components (cardinality) of output for each principal component, τ is the learning rate, D_{max} is the maximum depth of a tree, Sub refers to the subsample ratio of the training instances, and n_T is the number of trees in the model. Additionally, $Drop$ refers to hidden dropout ratio, and ℓ_2 is the parameter of the ℓ_2-norm penalty for the DNN method.

In general, the predictive performance obtained by our model for both scenarios is significantly higher than the baseline methods. As mentioned earlier, only the top few most sensitive drugs will be of great interest in practice. Clearly, the proposed method performs considerably well in this aspect. Evidently, the difference between the best and the second-best AH@5 for 100 drugs, is greater than what we find for 50 drugs. In our experiment on lung cancer cell lines, RSVM shows a reasonably good overall performance and it is the second-best methods. Moreover, LambdaMART, DNN, and NNR which perform well in terms of AH@5 and AH@10 for 50 drugs, performed poorly for 100 drugs. As can be seen, due to the limited number of samples available for training and testing, the performance of the baseline methods decreases significantly when we increase the number of drugs. Nevertheless, the proposed method is able to maintain its high performance. The data augmentation step (i.e., comparing each drug with other drugs in the same cell line and comparing that drug with itself in other cell lines) and predicting scores instead of the exact drug responses are two main reasons behind this superior performance.

In Case Study 2, we used the cell lines which were derived from the blood cancer tumors (leukemia group). After removing missing data, we retain 63 cell lines and 100 drugs. Based on our gene selection results, 134 genes are selected as a minimal subset of all genes. The details of the gene selection scheme and the list of the selected genes can be found in Appendix A. The performance comparison of ranking methods for both 50 and 100 drugs is presented in Table 2. The experimental setting for these experiments is exactly like Case Study 1.

For 50 drugs, the proposed method consistently outperforms all baseline methods across all performance metrics except AH@5 where it is tied with several methods (RSVM, LASSO, KRR, EN, SPCR, GPR, and NNR). However, these methods do not have a comparable performance in other metrics. As we can see, AH@10 is the highest for the proposed method and the difference between the best and the second-best AH@10 is considerably high. For 100 drugs, the performance of the proposed method is even better. Most of baseline methods demonstrate equal performance, to which our model compares favorably. In these experiments,

TABLE 2: Performance Comparison of Ranking Methods for 50 and 100 Drugs (Blood Cancer Cell Lines).

	Methods	Parameters				AH@5	AH@10
50 Drugs	**RSVM**	$\Omega = 0.1$	$\gamma = 0.1$	Kernel= polynomial		**3.42**	5.50
	LASSO	$\lambda = 1000$				**3.42**	5.50
	KRR	$\lambda = 1000$	$\sigma^2 = 10$	Kernel= RBF		**3.42**	5.50
	EN	$\alpha = 0.3$		$\lambda = 10$		**3.42**	5.50
	BMTMKL	$\alpha_b = 10^{-10}$		$\beta_b = 10^{10}$		3.17	5.42
	PLS	PLScomp = 30				3.00	5.00
	PCR	Pcomp = 26				3.25	5.33
	SPCR	Pcomp = 25		Card = 50		**3.42**	5.50
	GPR	Kernel= Matern 3/2				**3.42**	5.50
	NNR	Epochs= 100		Batch size= 5		**3.42**	5.50
	LambdaMART	$\tau = 0.1$	$n_T = 1000$	$D_{max} = 5$	$Sub = 0.7$	**3.33**	**5.58**
	DNN	Epochs= 100	Batch size= 5	$Drop = 0.3$	$\ell_2 = 0.001$	**3.33**	5.42
	Proposed Model	$\phi_1 = 5$	$\phi_2 = 1$	NonZ= 2886	$\lambda_1 = 10^5$	**3.42**	**5.92**
100 Drugs	**RSVM**	$\Omega = 2$	$\gamma = 0.1$	Kernel= polynomial		**3.42**	**5.08**
	LASSO	$\lambda = 100$				**3.42**	**5.08**
	KRR	$\lambda = 100$	$\sigma^2 = 100$	Kernel= RBF		**3.42**	**5.08**
	EN	$\alpha = 0.8$		$\lambda = 10$		**3.42**	**5.08**
	BMTMKL	$\alpha_b = 10^{-10}$		$\beta_b = 10^{10}$		2.83	**5.08**
	PLS	PLScomp = 29				2.92	4.83
	PCR	Pcomp = 27				3.17	**5.08**
	SPCR	Pcomp = 25		Card = 25		**3.42**	**5.08**
	GPR	Kernel= squared exponential				**3.42**	**5.08**
	NNR	Epochs= 50		Batch size= 5		**3.42**	**5.08**
	LambdaMART	$\tau = 0.01$	$n_T = 100$	$D_{max} = 5$	$Sub = 0.7$	**3.42**	**5.17**
	DNN	Epochs= 50	Batch size= 5	$Drop = 0.3$	$\ell_2 = 0.01$	**3.42**	**5.08**
	Proposed Model	$\phi_1 = 2.5$	$\phi_2 = 1$	NonZ= 2664	$\lambda_1 = 10$	**3.58**	**5.17**

LambdaMART, NNR, RSVM, and EN show a reasonably good overall performance and are the second-best methods.

In Case Study 3, we used the cell lines which were derived from myeloma and lymphoma groups. Lymphoma and myeloma are cancers of the immune system. After removing missing data, we have 59 cell lines and 100 drugs. According to our gene selection results, 115 genes are selected as a minimal subset of all genes. The details of gene selection scheme and the list of the selected genes can be found in Appendix A. Table 3 shows the performance comparison for 50, and 100 drugs.

For 50 drugs, the proposed method outperforms all baseline methods across all performance metrics except AH@5 where it is tied with two methods (PLS and PCR). However, these methods do not have a comparable AH@10. When we increase the number of drugs to 100, the predictive performance obtained by our method is found to be significantly higher than the baseline methods. In our experiment for myeloma and lymphoma cell lines, EN shows a reasonably good overall performance and it is the second-best method. Furthermore, LambdaMART and PLS that perform really well in terms of AH@10 for 50 drugs, demonstrated a poor performance for 100 drugs.

Even though in personalized medicine and drug selection, we just care about the top few most sensitive drugs, knowing the right order of those sensitive drugs will be of great interest. In this regard, CI^s measures whether the ordering structure of a ranking list for sensitive drugs is close to its ground truth or not. As demonstrated in Figure 4, the proposed method has a higher ability to put the sensitive drugs in the right order compared to the baseline methods. LambdaMART is the only method that has a comparable or better CI^s. However, when we consider the main evaluation metrics (AH@5 and AH@10), it does not have a comparable performance.

In a nutshell, these results suggest that predicting drug sensitivities with the proposed method leads to superior predictive performance than the baseline methods across all performance metrics. Moreover, the proposed method is not only able to push the most sensitive drugs to the top of the ranking list, but it can put them in the right order.

4 Discussion and Conclusion

Since the effectiveness of medicines and therapies varies among patients, a conventional clinical practice is to treat cancer patients with a variety of therapeutic options. However, using molecular data, we can use the biological differences among patients' cancers to choose precise and individualized therapeutic options. To that end, we proposed a novel elastic-net-based regression that utilizes gene expression features and drug sensitivity data to build a predictor. In general, this model provides the following considerable advantages:

- Instead of predicting the exact drug response for each drug, we calculate a score for each drug and then use these scores to rank the drugs for a specific cancer cell line. This approach is able to solve the problem of fitting on the insensitive drugs that is usually seen in regression-based models.

TABLE 3: Performance Comparison of Ranking Methods for 50 and 100 Drugs (Myeloma and Lymphoma Cell Lines).

	Methods	Parameters				AH@5	AH@10
50 Drugs	**RSVM**	$\Omega = 1$	$\gamma = 1$	Kernel= polynomial		**4.00**	6.82
	LASSO	$\lambda = 10$				**4.00**	**6.91**
	KRR	$\lambda = 1$	$\sigma^2 = 1$	Kernel= RBF		**4.00**	**6.91**
	EN	$\alpha = 0.9$		$\lambda = 10$		**4.00**	**6.91**
	BMTMKL	$\alpha_b = 10^{-10}$		$\beta_b = 10^{10}$		**4.00**	6.54
	PLS	PLScomp = 24				**4.09**	**6.91**
	PCR	Pcomp = 26				**4.09**	6.82
	SPCR	Pcomp = 24		Card = 75		**4.00**	6.82
	GPR	Kernel= Matern 3/2				**4.00**	**6.91**
	NNR	Epochs= 100		Batch size= 25		**4.00**	6.82
	LambdaMART	$\tau = 0.01$	$n_T = 100$	$D_{max} = 5$	$Sub = 1$	**4.00**	**7.00**
	DNN	Epochs= 50	Batch size= 5	$Drop = 0.3$	$\ell_2 = 0.01$	**4.00**	6.54
	Proposed Model	$\phi_1 = 0.1$	$\phi_2 = 1$	NonZ= 3980	$\lambda_1 = 0.1$	**4.09**	**7.00**
100 Drugs	**RSVM**	$\Omega = 1$	$\gamma = 0.1$	Kernel= polynomial		3.82	6.64
	LASSO	$\lambda = 10$				4.00	6.64
	KRR	$\lambda = 10$	$\sigma^2 = 10$	Kernel= RBF		**4.09**	6.64
	EN	$\alpha = 0.1$		$\lambda = 10$		**4.09**	**6.73**
	BMTMKL	$\alpha_b = 10^{-10}$		$\beta_b = 10^{10}$		3.82	6.64
	PLS	PLScomp = 28				4.00	6.45
	PCR	Pcomp = 32				3.82	6.45
	SPCR	Pcomp = 28		Card = 100		3.82	6.64
	GPR	Kernel=exponential				4.00	6.64
	NNR	Epochs= 100		Batch size= 5		3.82	6.64
	LambdaMART	$\tau = 0.1$	$n_T = 10$	$D_{max} = 10$	$Sub = 0.7$	3.64	6.36
	DNN	Epochs= 100	Batch size= 5	$Drop = 0.3$	$\ell_2 = 0.01$	3.82	6.45
	Proposed Model	$\phi_1 = 0.3$	$\phi_2 = 1$	NonZ= 4180	$\lambda_1 = 50$	**4.18**	**6.82**

- We learn a coefficient matrix for all the cell lines and drugs, and then define a score function using that matrix to prioritize drugs within each cell line. In this way, we are able to capture drug-cell line relations between various cell lines.
- Since we usually do not have access to enough data, we need to achieve high performance with a limited amount of data. The specific structure of this model (comparing each drug with other drugs in the same cell line and comparing that drug with itself in other cell lines) substantially improves the accuracy of the drug prediction model under limited data.
- The proposed model is able to maintain its high performance even when we use a large number of drugs and a few cell lines. As we saw in Section 3, when we increase the number of drugs, we increase the complexity of the problem and predicting the true ranking becomes more difficult. We observed that when increasing the number of drugs, the ability of other methods to predict the right ranking decreases substantially. However, the proposed method maintains its high performance even when we run it for 100 drugs.
- A great number of methods in the literature try to learn either a vector or a matrix for each drug or cell line. Such methods need to keep a huge amount of information to find the ranking for new samples. In contrast, the proposed method maintains one sparse matrix.
- Learning one sparse matrix for all cell lines and drugs reduces the possibility of over fitting in the proposed model.

Although our model is quite promising, it also suffers from some limitations that can be addressed in future work. First, in the proposed model, we only used gene expression data. However, one way to increase the prediction accuracy of this model is to incorporate various genomic information, such as epigenomic characterizations and protein level information. Furthermore, an interesting future direction is to incorporate the toxicity of drugs in our predictions. Due to the biological differences among patients, the side effects of medicines and therapies may vary. However, when we predict the most effective drugs, we do not consider the side effects and toxicity of each drug. Therefore, by considering this factor, we can further optimize our predictions. Finally, graph neural networks have been shown to yield state-of-the-art results in some applications compared to other deep learning-based approaches [37]. It has been demonstrated that representing compound structures as molecular graphs can improve the performance of drug response prediction [37]. Thus, in future work, and assuming we have drug composition information, we can represent drug molecules as graphs instead of using one-hot encoding vectors.

REFERENCES

[1] I. Bayer, P. Groth, and S. Schneckener, "Prediction Errors in Learning Drug Response from Gene Expression Data – Influence of Labeling, Sample Size, and Machine Learning Algorithm," *PLoS ONE*, vol. 8, no. 7, Jul. 2013.

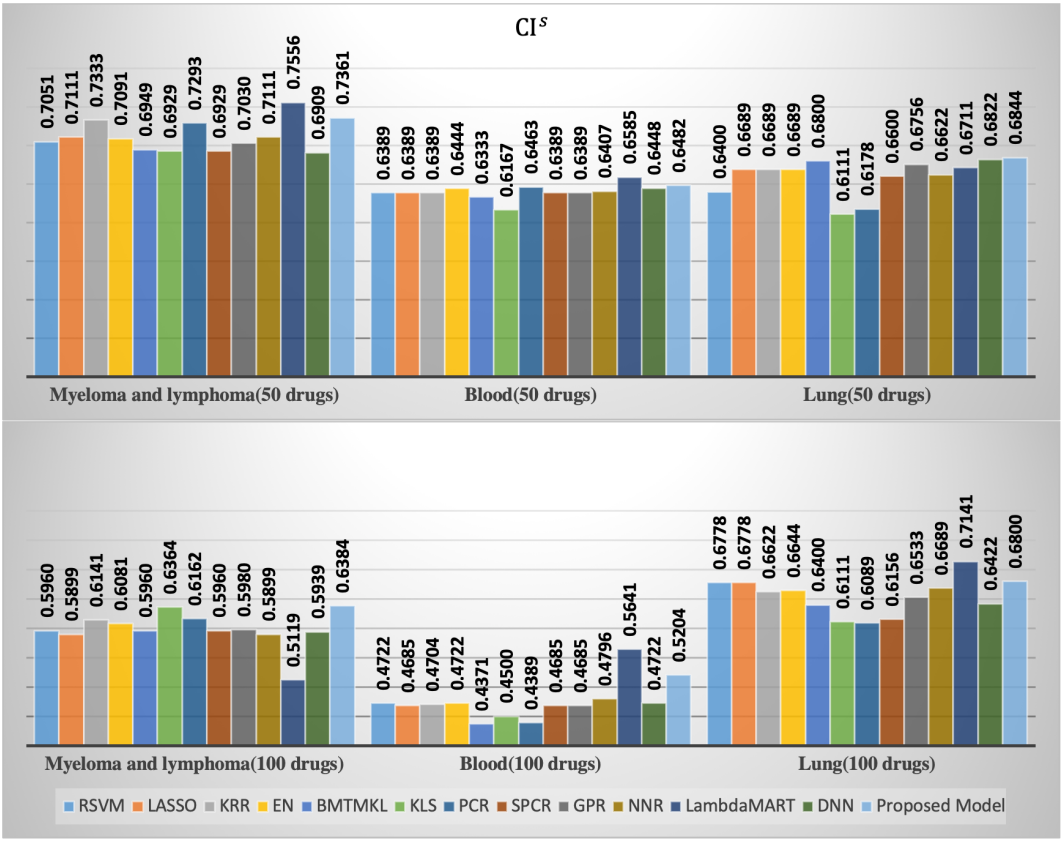

Fig. 4: CI^s comparison of ranking methods for 50 and 100 drugs.

[2] Z. Dong, N. Zhang, C. Li, H. Wang, Y. Fang, J. Wang, and X. Zheng, "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection," *BMC Cancer*, vol. 15, no. 1, p. 489, Jun. 2015.

[3] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012, number: 7391 Publisher: Nature Publishing Group.

[4] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, Z. Sanborn, F. Pepin, S. Durinck, J. E. Korkola, M. Griffith, J. S. Hur, N. Huh, J. Chung, L. Cope, M. J. Fackler, C. Umbricht, S. Sukumar, P. Seth, V. P. Sukhatme, L. R. Jakkula, Y. Lu, G. B. Mills, R. J. Cho, E. A. Collisson, L. J. van't Veer, P. T. Spellman, and J. W. Gray, "Modeling precision treatment of breast cancer," *Genome Biology*, vol. 14, no. 10, p. R110, Dec. 2013.

[5] L. Pusztai, K. Anderson, and K. R. Hess, "Pharmacogenomic Predictor Discovery in Phase II Clinical Trials for Breast Cancer," *Clinical Cancer Research*, vol. 13, no. 20, pp. 6080–6086, Oct. 2007, publisher: American Association for Cancer Research Section: Imaging, Diagnosis, Prognosis.

[6] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," in *Biocomputing 2014*. WORLD SCIENTIFIC, Nov. 2013, pp. 63–74.

[7] C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir, "Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients," *Nature*, vol. 194, no. 4824, pp. 178–180, Apr. 1962.

[8] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O'Brien, J. L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott, and C. H. Benes, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, pp. 570–575, Mar. 2012, number: 7391 Publisher: Nature Publishing Group.

[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[10] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines," *Genome Biology*, vol. 15, no. 3, p. R47, Mar. 2014.

[11] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[12] L. Breiman, "Random forests machine learning, vol. 45," 2001.

[13] C. De Niz, R. Rahman, X. Zhao, and R. Pal, "Algorithms for Drug Sensitivity Prediction," *Algorithms*, vol. 9, no. 4, p. 77, Dec. 2016, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[14] S. Haider, R. Rahman, S. Ghosh, and R. Pal, "A Copula Based Approach for Design of Multivariate Random Forests for Drug Sensitivity Prediction," *PLoS ONE*, vol. 10, no. 12, Dec. 2015.

[15] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine, "Predicting in vitro drug sensitivity using Random Forests," *Bioinformatics*, vol. 27, no. 2, pp. 220–224, Jan. 2011, publisher: Oxford Academic.

[16] J. D. Ospina, J. Zhu, C. Chira, A. Bossi, J. B. Delobel, V. Beckendorf,

B. Dubray, J.-L. Lagrange, J. C. Correa, A. Simon, O. Acosta, and R. de Crevoisier, "Random Forests to Predict Rectal Toxicity Following Prostate Cancer Radiation Therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 89, no. 5, pp. 1024–1031, Aug. 2014.

[17] Y. Ma, Z. Ding, Y. Qian, X. Shi, V. Castranova, E. J. Harner, and L. Guo, "Predicting Cancer Drug Response by Proteomic Profiling," *Clinical Cancer Research*, vol. 12, no. 15, pp. 4583–4589, Aug. 2006, publisher: American Association for Cancer Research Section: Imaging, Diagnosis, Prognosis.

[18] Y. He, J. Liu, and X. Ning, "Drug Selection via Joint Push and Learning to Rank," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 1, pp. 110–123, Jan. 2020.

[19] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 517–564, 2018.

[20] ——, "Distributionally robust learning," *Foundations and Trends® in Optimization*, vol. 4, no. 1-2, pp. 1–243, 2020.

[21] "Broad Institute Cancer Cell Line Encyclopedia (CCLE)." [Online]. Available: https://portals.broadinstitute.org/ccle

[22] "Cancer Therapeutics Response Portal." [Online]. Available: https://portals.broadinstitute.org/ctrp.v2.1/

[23] D. K. Slonim and I. Yanai, "Getting started in gene expression microarray analysis," *PLoS Comput Biol*, vol. 5, no. 10, p. e1000543, 2009.

[24] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd ed., ser. Springer Series in Statistics. New York: Springer-Verlag, 2009.

[25] T.-Y. Liu, "Learning to Rank for Information Retrieval," Mar. 2009.

[26] W. A. Abbasi, A. Asif, S. Andleeb, and F. U. A. A. Minhas, "Camels: In silico prediction of calmodulin binding proteins and their binding sites," *Proteins: Structure, Function, and Bioinformatics*, vol. 85, no. 9, pp. 1724–1740, 2017.

[27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, Sep. 2012, google-Books-ID: RC43AgAAQBAJ.

[28] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, and G. Stolovitzky, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014, number: 12 Publisher: Nature Publishing Group.

[29] R. Rosipal and L. J. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space," *Journal of Machine Learning Research*, vol. 2, no. Dec, pp. 97–123, 2001.

[30] M. Hein and T. Bühler, "An Inverse Power Method for Nonlinear Eigenproblems with Applications in 1-Spectral Clustering and Sparse PCA," *arXiv:1012.0774 [cs, math, stat]*, Dec. 2010, arXiv: 1012.0774.

[31] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Mass: The MIT Press, Nov. 2005.

[32] K. De Brabanter, K. Pelckmans, J. De Brabanter, M. Debruyne, J. A. K. Suykens, M. Hubert, and B. De Moor, "Robustness of Kernel Based Regression: A Comparison of Iterative Weighting Schemes," in *Artificial Neural Networks – ICANN 2009*, ser. Lecture Notes in Computer Science, C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, Eds. Berlin, Heidelberg: Springer, 2009, pp. 100–110.

[33] C. J. Burges, "From ranknet to LambdaRank to LambdaMART: An overview," *Learning*, vol. 11, no. 23-581, p. 81, 2010.

[34] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv:1508.05326 [cs]*, Aug. 2015, arXiv: 1508.05326.

[35] T. Sakellaropoulos, K. Vougas, S. Narang, F. Koinis, A. Kotsinas, A. Polyzos, T. J. Moss, S. Piha-Paul, H. Zhou, E. Kardala *et al.*, "A deep learning framework for predicting response to therapy in cancer," *Cell reports*, vol. 29, no. 11, pp. 3367–3373, 2019.

[36] P. A. Trott, "International Classification of Diseases for Oncology," *Journal of Clinical Pathology*, vol. 30, no. 8, p. 782, Aug. 1977.

[37] T.-T. Nguyen, G. T. T. Nguyen, T. Nguyen, and D.-H. Le, "Graph convolutional networks for drug response prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

Shahabeddin Sotudian received the B.S. and M.S. degrees both in industrial and systems engineering from Tehran Polytechnic, in 2015 and 2017, respectively. He is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering, Division of Systems Engineering, Boston University, Boston, MA, USA. His research interests lie in optimization, applied probability and statistics, and machine learning, with applications in healthcare and biology.

Ioannis Ch. Paschalidis received the M.S. and Ph.D. degrees both in electrical engineering and computer science from the Massachusetts Institute of Technology(MIT), Cambridge, MA, USA, in 1993 and 1996, respectively. In September 1996 he joined Boston University where he has been ever since. He is a Professor at Boston University with appointments in the Department of Electrical and Computer Engineering, the Division of Systems Engineering, and the Department of Biomedical Engineering. He is the Director of the Center for Information and Systems Engineering (CISE). He has held visiting appointments with MIT and Columbia University, New York, NY, USA. His current research interests lie in the fields of systems and control, networking, applied probability,optimization, operations research, computational biology, and medical informatics. Dr. Paschalidis is a recipient of the NSF CAREER award (2000), several best paper and best algorithmic performance awards, and a 2014 IBM/IEEE Smarter Planet Challenge Award. He was an invited participant at the 2002 Frontiers of Engineering Symposium, organized by the U.S. National Academy of Engineering and the 2014 U.S. National Academies Keck Futures Initiative (NAFKI) Conference. He is the inaugural Editor-in-Chief of the IEEE Transactions on Control of Network Systems.