

Data Imbalance in Drug Response Prediction – Multi-Objective Optimization Approach in Deep Learning Setting

Oleksandr Narykov^{1,*}, Yitan Zhu¹, Thomas Brettin¹, Yvonne A. Evrard², Alexander Partin¹, Fangfang Xia¹, Maulik Shukla¹, Priyanka Vasanthakumari¹, James H. Doroshow³ and Rick L. Stevens^{1,4}

¹Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL 60439, USA, ²Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA, ³Developmental Therapeutics Branch, National Cancer Institute, Bethesda, MD 20892, USA, ⁴Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA

*To whom correspondence should be addressed.

Abstract

Drug response prediction (DRP) methods tackle the complex task of associating the effectiveness of small molecules with the specific genetic makeup of the patient. Anti-cancer DRP is a particularly challenging task requiring costly experiments as underlying pathogenic mechanisms are broad and associated with multiple genomic pathways. The scientific community has exerted significant efforts to generate public drug screening datasets, giving a path to various machine learning (ML) models that attempt to reason over complex data space of small compounds and biological characteristics of tumors. However, the data depth is still lacking compared to computer vision or natural language processing domains, limiting current learning capabilities. To combat this issue and increase the generalizability of the DRP models, we are exploring strategies that explicitly address the imbalance in the DRP datasets. We reframe the problem as a multi-objective optimization across multiple drugs to maximize deep learning model performance. We implement this approach by constructing Multi-Objective Optimization Regularized by Loss Entropy (MOORLE) loss function and plugging it into a Deep Learning model. We demonstrate the utility of proposed drug discovery methods and make suggestions for further potential application of the work to promote equitable outcomes in the healthcare field.

Availability: <https://github.com/AlexandrNP/MOORLE>

Contact: onarykov@anl.gov

1 Introduction

Cancer is a widely spread genetic disease family with a common characteristic of uncontrolled cell growth and proliferation (Bray, et al., 2021; Cronin, et al., 2018). This set of complex genetic disorders is highly heterogeneous and notoriously difficult to combat. Artificial intelligence (AI) technologies are being incorporated into this field to facilitate our ability to treat patients. E.g., machine learning (ML) systems are used doctors in processing radiological images and histopathological information.

Drug response prediction (DRP) is an important application of ML as it projects our estimates for the small ligand efficacy in treating cancer (Fig.1). Designing efficient DRP models can help with real-world problems of drug repurposing, personalized medicine, and virtual drug screening by reducing the number of costly wet lab experiments required to devise novel treatment protocols or develop new drugs.

However, all those settings require different approaches for assessment. Some models approach the problem in drug-specific scenarios, corresponding to personalized medicine settings, e.g., MOLI (Sharifi-Noghabi, et al., 2019). They predict drug response for a particular small molecule-based exclusively on biological information. These models cannot make inferences based on previously unseen chemical compounds. However, this setting severely limits the amount of information available for training and the strength of the model. To alleviate this issue and extend the application to different scenarios, most works approach DRP as a pair-input problem. This setting is called pan-drug DRP (Partin, et al., 2023).

The DRP field is abundant and contains multiple models based on traditional ML approaches – Random Forest (Breiman, 2001), AdaBoost (Hastie, et al., 2009), XGBoost (Lu, et al., 2021), LightGBM (Ke, et al., 2017), and Support Vector Machine (Chapelle, et al., 1999). The recent trend is the extensive usage of Deep Learning (DL) models that utilize automatic feature extraction associated with multi-layer Neural Networks (NN). One of the first approaches in this direction was described by (Menden, et al., 2013), who proposed a single-layer NN for predicting IC50. In recent years, the DRP field has had multiple models based on various architectures – convolutional neural networks (DeepIC50, DeepCDR, IGTD), graph neural networks (GraphDRP, GraTransDRP), attention-based models (PaccMann, CADRE, DeepTTA) (Chu, et al., 2022; Jiang, et al., 2022; Joo, et al., 2019; Liu, et al., 2020; Nguyen, et al., 2021; Oskooei, et al., 2018; Tao, et al., 2020; Zhu, et al., 2021).

A pair-input setting introduces known model evaluation pitfalls (Park and Marcotte, 2012), and it is important to make appropriate train/test splits to get generalizable performance estimates. So, for a drug repurposing scenario, it is natural for the model to have prior information on both biological samples and ligands. It means that the training set may include the response of the drug in question on another cell line and the response of some other drugs on a given cell line. As long as the combination of biological data and ligands is unique, including it in the test set is appropriate. However, for virtual drug screening, ensuring that the model has no prior information on the small molecule is essential. This means that if a drug appears in a test set entry, no pairs should be involved in the training set. Otherwise, we would observe information leakage and have over-optimistic results.

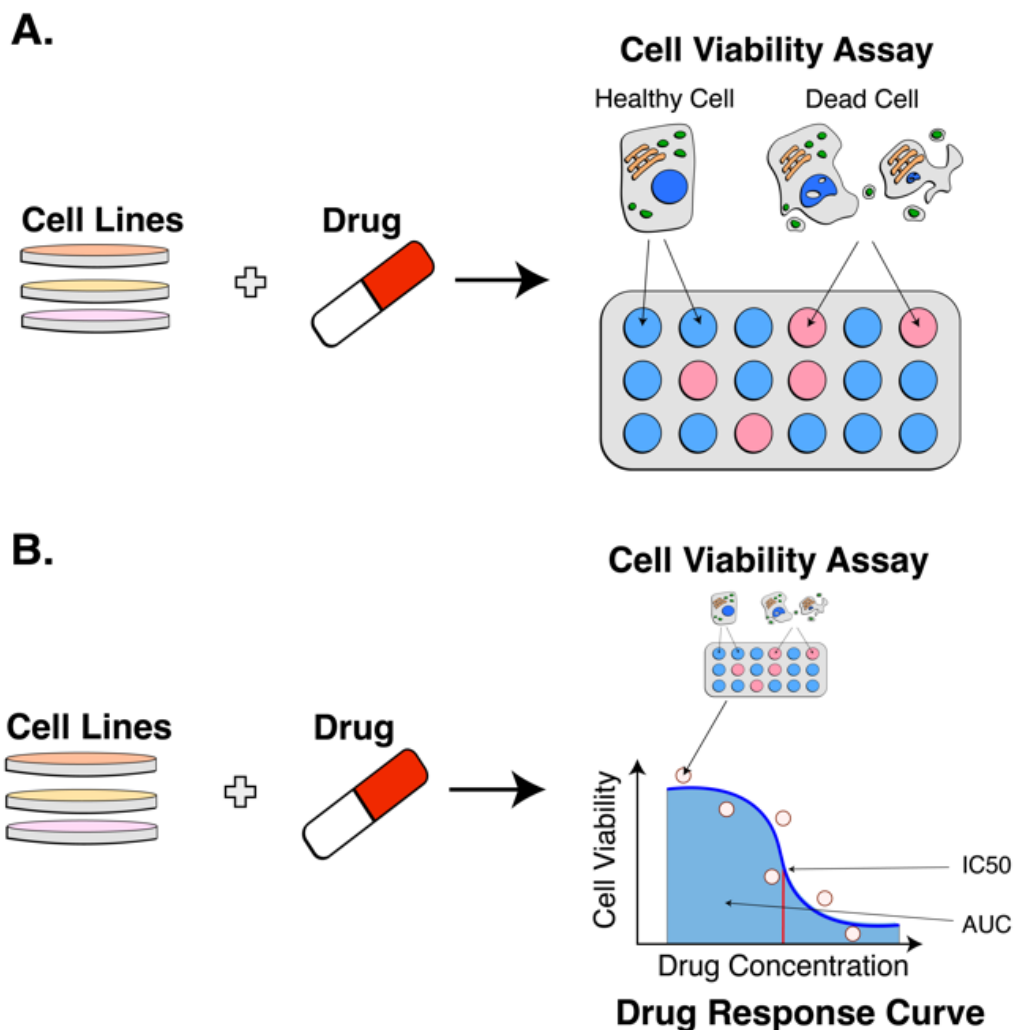


Fig. 1. Drug response experiment. A. A single cell viability assay. A combination of membrane-permeable and membrane-impermeable fluorescent proteins can estimate the degree of tumor inhibition. B. Multiple cell viability assays are integrated into a single drug response measurement based on the hill-slope model, graphical interpretation of IC₅₀ and AUC drug response measures (Yanagawa, et al., 1989).

Personalized medicine aims to find a treatment plan best suited for a specific patient based on their biological characteristics – genetic makeup, disease history, and style of living. DRP applications in this area aim to detect drug resistivity and, for cancer, find the most efficient drug to combat tumors specific to a given patient (Partin, et al., 2023; Zhu, et al., 2020).

Virtual screening (McGaughey, et al., 2007) setting is one of the most challenging applications for DRP models, as drug response variability between drugs is much higher than between cell line response variations (Zhu, et al., 2020). It is vital for advancing drug discovery capabilities. A significant number of works in the field focus on a one-size-fits-all optimization approach when training models. In most cases, the target is to minimize Mean Squared Error (MSE) over all pair-inputs. The performance of pan-cancer pan-drug models is commonly evaluated on a cross-validation (CV) holdout test set using performance metrics like Pearson Correlation Coefficient (PCC) and the coefficient of determination (R²). This approach assumes the ability of ML algorithms to uncover relationships between variables automatically and hides the complexity of underlying dataset structures. While multiple works address the confounding factors for the regression, most of them focus on information from different modalities, e.g., copy number variations (CNV) or mutation data (He, et al., 2022; Jia, et al., 2021; McNamee, 2005).

MOORLE

In this work, we are investigating the benefits of explicitly addressing complex substructures arising from p from the dataset construction with a focus on improving the virtual screening problem (Fig.2). We discuss existing approaches for learning from imbalanced data and propose an outlook on drug response prediction as a multi-objective optimization (MOO) task, attempting to maximize the prediction performance over different drugs and cancers. MOO approaches usually address problems that have multiple criteria for their evaluation. Recent work proposed its usage in contrastive learning (Moukafih, et al., 2023).

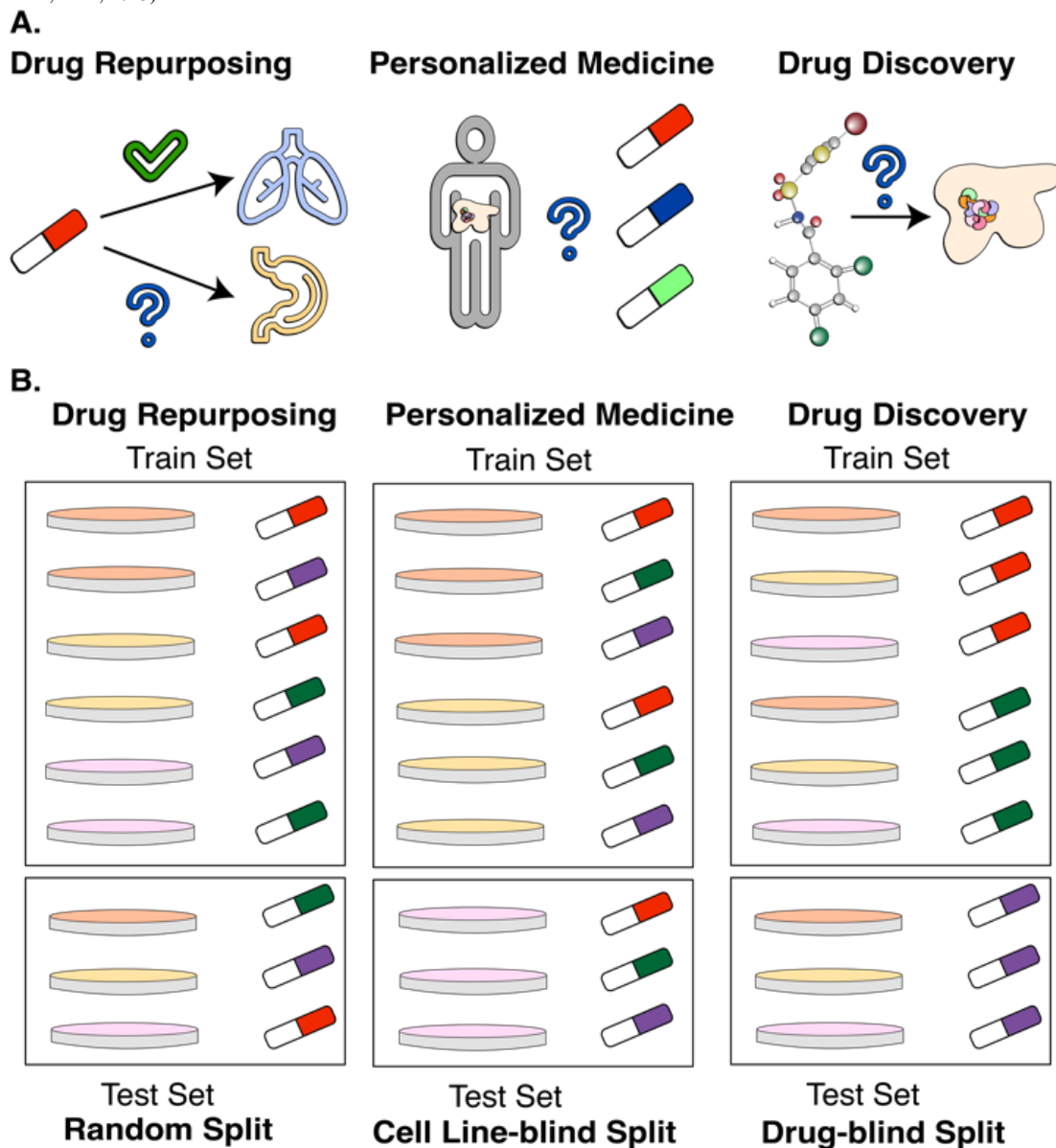


Fig. 2. Drug response prediction application areas. A. Real-world tasks that benefit from DRP models. B. Corresponding splits of pair-input entries from drug response datasets. Each Petri dish color corresponds to a unique cell line, and each color of the drug corresponds to a unique drug.

Due to the pair-input way of constructing datasets for DRP, we can approach this problem as a hybrid formulation between classification and regression tasks. While the final goal of DRP models is to predict the continuous value of area under the drug response curve (AUC) or 50% inhibition (IC50) (Kurilov, et al., 2020), the dataset imbalance follows conducted experiments across discrete cell lines and ligand names, which can be understood as classes. Our current work focuses on drug discovery applications, corresponding to the drug-blind split of the datasets. In addition, we are also providing an assessment for the drug repurposing task to validate proposed methods, as it is one of the standard formulations of the DRP problem (Partin, et al., 2023).

2 Methods

2.1 Data

The primary data source in the DRP field is cell line experiments that measure tumor inhibition via cell viability assays. Multiple metrics characterize experimental tumor inhibition results, the most widespread being the cutoff for drug concentration that provides IC50 and AUC (Fig.1). Those are continuous metrics, so regression models are traditionally used to estimate them.

In this study, we use standard DRP cell line datasets – Cancer Therapeutics Response Portal (CTRP) (Basu, et al., 2013) and Cancer Cell Line Encyclopedia (CCLE) (Barretina, et al., 2012). CCLE dataset contains 8,950 experiments based on 474 unique cell lines and 24 drugs. Data comprises RNA-Seq gene expressions, corresponding compounds, and drug response for the combination of those two entries. The CTRP dataset does not contain gene expression data but utilizes standard commercially available cell lines and contains a much larger number of drug response experiments - 254,566. It is based on 812 cell lines and 495 ligands. Gene expression data for biological samples came from different sources, including CCLE. While dataset CCLE is mainly balanced, the variability of the experiment number is much higher for CTRP (Fig.3). Having CCLE as one of the test sets also helps us assess whether datasets with a high number of classes benefit from the proposed methodologies, regardless of the imbalance presence.

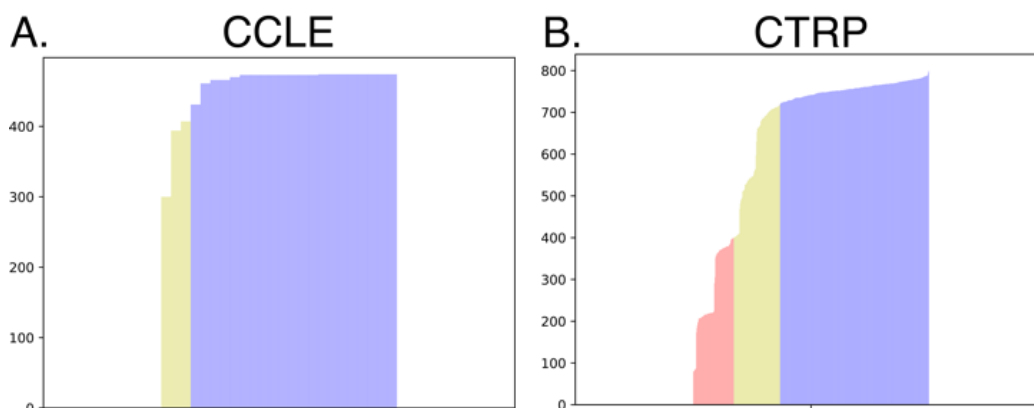


Fig 3. Number of experiments associated with each drug in a dataset. A. CCLE dataset. The highest number of experiments related to a single drug is 474. 12.5% of the drugs (yellow) have a number of experiments related to them, which is less than 90% of the highest number. The rest of the drugs are depicted in blue. B. CTRP dataset. The highest number of experiments associated with a single drug is 799. 17.2% of the drugs (red) have a number of experiments related to them, which is less than 50% of this highest number. 19.6% of the drugs (yellow) have a number of experiments between 50% and 90% of the highest number of experiments. The rest of the drugs are depicted in blue.

Drug-level information for both datasets comes in the form of molecular fingerprints computed via Dragon v.7.0 (Cheminformatics) and Simplified Molecular-Input Line-Entry System (SMILES) (Weininger, 1988) entries obtained from the PubChem (Kim, et al., 2019) and the web form of Developmental Therapeutic Program (DTP) (<https://dtp.cancer.gov/>).

2.2 Learning from imbalanced data

Data imbalance is a common problem in machine learning arising due to the limited amount of available learning data (Haixiang, et al., 2017; Rezvani and Wang, 2023). This issue garnered more attention in the context of classification. Traditionally, two major approaches have been developed to handle data imbalance in the datasets: sampling (Fig.4) and cost adaptation.

The first is focused on data preprocessing and includes various sampling techniques and synthetic data generation. It includes undersampling, oversampling, a combination of these approaches (e.g., weighted sampling) (Fig.4), the SMOTE (Chawla, et al., 2002) technique, and its adaptation to the regression problem (Torgo, et al., 2013). The undersampling strategy balances data by discarding excessive data in overrepresented classes. It works well when samples of the same class are similar and additional data points from that class are not crucial for making precise predictions. It is unsuitable for the DRP problem because it leads to severe data loss. Oversampling randomly draws instances from underrepresented classes with replacement until the number of examples from each class is balanced. In this case, we have no data loss; however, the importance of data points from underrepresented classes becomes inflated, which may bias the model (Haixiang, et al., 2017). Depending on the number of classes and present imbalance, oversampling may significantly inflate the size of the dataset. Balanced (weighted) sampling can be seen as a combination of the previous two approaches. Each data point is assigned a weight inversely proportional to the number of corresponding class instances in the dataset.

The SMOTE technique is based on k nearest neighbors and generates synthetic examples as a weighted average between a selected point and each of its neighbors (closest data points) from the same class (Chawla, et al., 2002). The algorithm was adopted for regression; however, it heavily relies on the assumption of linearity between features and response value, as well as the convexity of the clusters formed by different classes, which are not observed in DRP problem data. It is also known that in high-dimensional space, SMOTE tends to be severely biased towards underrepresented classes (Rezvani and Wang, 2023). Because of these points, we do not further consider SMOTE in our work.

The second major approach focuses on the learning algorithm modifications. There is a large body of works for classification problems that attempts to introduce class weights (weighted variations of random forest (Breiman, 2001) and SVM (Chapelle, et al., 1999), modify the loss function, (particle-swarm optimization network (Cao, et al., 2013), zSVM (Abarzadeh, et al., 2016), or refine boosting approaches (AdaC1-AdaC3 (Raghuwanshi and

MOORLE

Shukla, 2019), RareBoost (Joshi, et al., 2001), BABoost (Song, et al., 2009). For regression, probability-based methods such as reframing were introduced. This approach focuses on adapting to estimated outputs depending on the context (Hernández-Orallo, 2014).

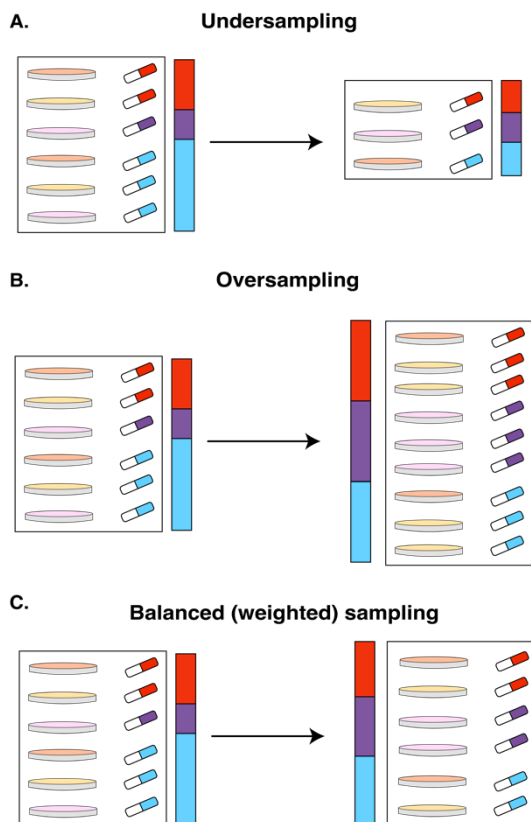


Fig. 4. Common sampling strategies for an imbalanced dataset in the context of the DRP problem. Each rectangle represents a dataset. A Petri dish of a distinct color corresponds to the unique cell line. Drugs of different colors represent unique small molecules and compose distinct groups in data. The proportion of unique drugs is also displayed in a color bar near each dataset. In our study, we treat each drug as a class. A. Undersampling. B. Oversampling. C. Balanced, or weighted, sampling.

2.3 Drug Response Prediction as Multiple Objective Optimization

As we discussed earlier, the most common model evaluation is based on integral performance, e.g., R^2 , PCC, concordance index, etc. In this paragraph, we are using R^2 as an example of performance measurement, and we refer to the standard evaluation of the entire hold-out portion of the dataset as R_{avg}^2 . This measure compares estimates of the residual sum of squares produced by the model

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \quad (1)$$

where y is the ground truth value, \hat{y} is the model prediction, and \bar{y} is the expected value of the response variable in the test dataset. Maximizing R^2 is a common target for DRP models, and can be considered a single-objective optimization problem. However, directly using R^2 for training the ML algorithms is not a common approach, as the coefficient of determination is not a convex function. It is more feasible to disregard the total sum of squares and solve the optimization problem directly for the residual sum of squares. It results in a common mean squared error (MSE) loss function:

$$MSE(f(x), y) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 \quad (2)$$

where x is the set of features, n – total number of datapoints, f – prediction model, y – ground truth value.

This formulation is suitable for drug repurposing task, as it considers each unique combination of biological sample and ligand a unique standalone data sample. However, when we discuss the virtual drug screening application, we are interested in the model's ability to reason over individual small molecules. It means that the performance of each drug in the dataset can be considered a standalone optimization problem. Let's consider that for each drug we are attempting to maximize $R_{Drug_i}^2$. We can construct a data space where each coefficient of determination corresponding to i -th drug forms an orthonormal basis. Then each individual machine learning model can be uniquely described based on the performance it achieves for the corresponding small molecule (Fig.5A). E.g., in two dimensional case with two drugs $Drug_1$ and $Drug_2$ vector $\langle R_{Drug_1}^2, R_{Drug_2}^2 \rangle$ defines the coordinates of the corresponding machine learning model. On top of individual decomposition into performance evaluations $R_{Drug_i}^2$ we can also associate an integral

performance metric with each machine learning model. It can be either R^2 over all datapoint or an average of individual performances $R^2_{Drug_i}$. In this work, we choose the latter because this measure is less sensitive to the number of experiments associated with each separate drug.

We can see (Fig.5A) that under these assumptions, multiple data points may correspond to the machine learning models with the same integral performance but different tradeoffs between individual drugs. This is a subset of the class-composed Pareto front that is defined as all possible models with extreme performance (Teich, 2001). Now, an important question is which ML model from the Pareto front is preferable. We hypothesize that selecting models closer to the center of this set (cyan and green points from Fig.5) will result in better generalization - a capability of predicting responses to new drugs not included in the training set - as the trained models are not biased towards some drugs and thus provide a better generalizability between drugs. The reason is that a better association behind unique chemical characteristics of the small compounds and biological samples features. Preliminary findings were discussed in (Narykov, et al., 2023). Fig.5B provides a graphical visualization of our objective – to maximize the integral performance of the model and to balance individual drug scores.

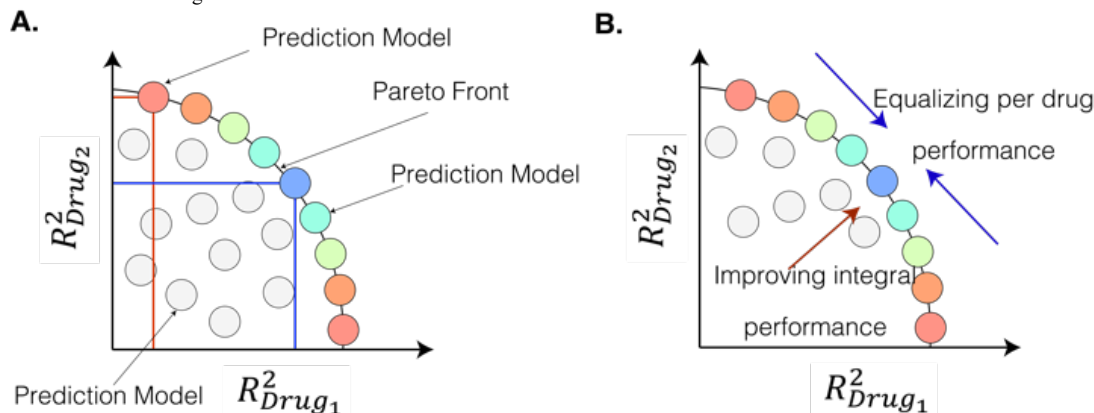


Fig. 5. Pareto front of ML models in the space composed of individual drug performance metrics. Colored nodes indicate ML models belonging to the Pareto front, i.e., having integral performance close to the maximum known value. Grey points represent ML models that have worse integral performance.

As in the case with the regular R^2 , the current formulation does not fit to be directly used for ML algorithms training. To realize this strategy, we can define a loss function based on the MSE of individual drugs and an entropy-like regularization component. We will define loss for individual drugs as

$$MSE_{d_i} = MSE(f(x_{d_i}), y_{d_i}) \quad (3)$$

where x is the set of features, d_i is the i -th drug, $f(\cdot)$ is the prediction model that produces AUC, y is the ground truth values. Then, we will calculate normalized losses by applying the softmax function to individual scores and put them in the set P :

$$p_{d_i} = \frac{\exp(MSE_{d_i})}{\sum_j \exp(MSE_{d_j})}, \quad (4)$$

$$P = \{p_{d_i}; d_i \in D\} \quad (5)$$

where D is the set of all drugs and p_{d_i} is normalized loss for drug d_i . As we can see, set P can be treated as a probability distribution. As we want to incentivize equal loss for individual drugs, we can use a regularization based on the entropy function:

$$H(P) = \sum_{x_j \in X} -p_j \log(p_j), \quad (6)$$

However, the function $H(P)$ is maximized when our desired property is achieved, and is concave. So, in order to adopt it for a loss function we will use it in form $\ln|D| - H(P)$, where $\ln|D|$ is the maximum value that discrete entropy can take for the distribution with $|D|$ entries. This transformation minimizes loss when drug-specific losses are equal and results in a convex function.

We will define the loss function over the set of features x , response values y , and classes D as

$$L(x, D, y) = \frac{1}{|D|} \sum_{i=1}^{|D|} MSE_{d_i} + a(\ln|D| - H(P)), \quad (7)$$

where a is a regularization coefficient, $H(X)$ is the entropy of a distribution. The averaged sum of drug-specific losses maximizes integral performance, while the entropy-based regularization component promotes evening-out loss across drugs. We call this construction Multi-Objective Optimization Regularized by Loss Entropy (MOORLE) loss function.

Each individual part of the equation (7) is convex. It allows us to state that $L(x, D, y)$ function is also convex, as it consists of a linear combination of convex functions. That is not a required but desired property for loss function. It results in a versatile model-agnostic loss function that can be utilized both in classical ML models and deep learning settings.

2.4 Mixed Sampling Approach

The practical consideration regarding loss function modification proposed in 2.3 is the modern approach for NN training. Instead of updating gradients for the entire dataset in the gradient descent (GD) algorithm, continuous updates of model weights are made based on mini-batches. This approach allows to train models significantly faster and is a main feature of stochastic gradient descent (SGD) or its improvements, e.g., ADAM (Kingma and Ba, 2014).

MOORLE

When we use sequential shuffled sampling that draws each element from the dataset once in random order, and is de facto a standard for deep learning frameworks like PyTorch and Keras, there is a high possibility that underrepresented drugs would have only a small influence on the objective function. Because of these considerations, we explored different sampling strategies described in 2.2. Undersampling, as expected, led to unsatisfactory performance for both CCLE and CTRP datasets, as most of the data was discarded. Oversampling was suitable for smaller-sized CCLE, but in the case of CTRP, this approach inflated the training dataset 20 times its original size and was not computationally feasible. Weighted sampling allowed us to control the size of the training dataset manually. However, maintaining the dataset size two to four times the original training set resulted in data loss and performance deterioration.

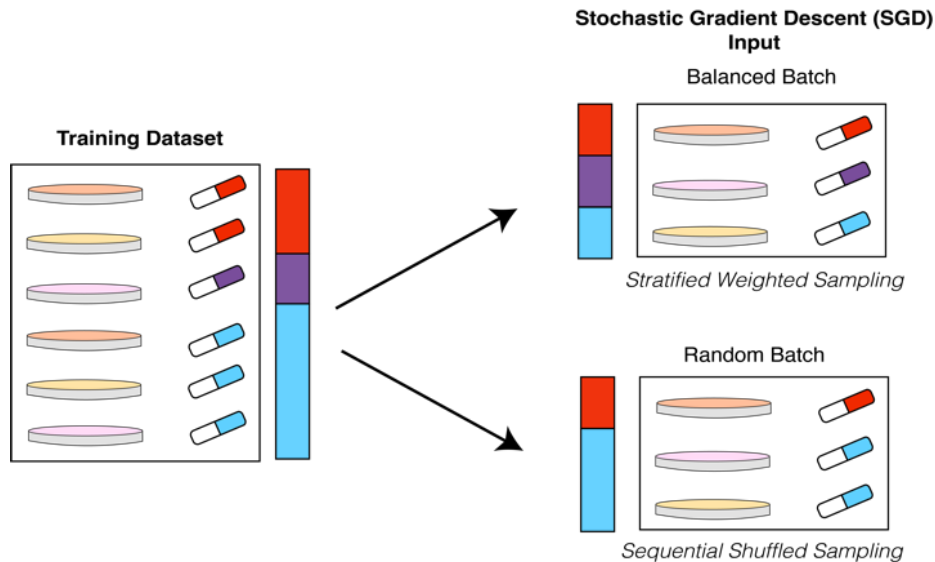


Fig. 6. Mixed sampling scheme for stochastic gradient descent-based algorithms.

This led us to develop a hybrid sampling strategy. Based on the common training dataset, we are deriving two sets of batches – one is the balanced batch based on weighted sampling, and another is a batch composed via sequential shuffle split (Fig.6). This mixed strategy inflates the size of the original training set twice and allows us to present model all samples from training dataset while ensuring regular update of cost function based on all present groups (classes) in the dataset. In practice, performing weighted sampling for each epoch significantly increases training time. To offset the cost of performing it, we cache obtained batches and store them for 10 epochs (including current), shuffling batch order before each new epoch. Using this heuristic, we reduce the number of times we have to perform weighted sampling.

2.5 Machine Learning Algorithms

The approach described in section 2.3 is model-agnostic and can be potentially adopted for any ML algorithm; due to the latest trends in the field, we are focusing on its adaptation for Deep Learning settings. We incorporate loss function (7) into the recent state-of-the-art model DeepTTA (Jiang, et al., 2022).

DeepTTA consists of three main components. One is an attention-based SMILES encoder subnetwork, and the other is a fully connected neural network (FCNN) (Zhu, et al., 2021). The last part of the network concatenates encodings for drugs and biological samples and performs regression.

3 Results

3.1 Experimental setup

We analyze the effect of adopting a mixed sampling approach and entropy-regularized loss function in DeepTTA models under random split and drug-blind split CV model evaluation strategies (Fig.2B). In a drug-blind setting, a ligand can not appear in the training and testing sets simultaneously. It ensures that no information about particular ligand is present in the test set. Each model run is evaluated by 10-fold cross-validation with fixed split that is shared among the studies for the same dataset and the same split type.

We perform an ablation study to investigate the influence of sampling strategy and loss function on model performance under various conditions. We consider two sampling strategies – standard sequential and mixed sampling- proposed in Section 2.4. We also consider two loss functions – widely used MSE and multi-objective loss function with entropy regularization proposed in Section 2.3. It results in four possible combinations of the influencing factors for each model run.

To estimate the statistical significance of the effect that proposed strategies have on the drugs we apply two-way repeated measurements ANOVA (Potvin and Schutz, 2000) algorithm from *pengouin* Python package. For random split and drug-blind settings, results for each CV iteration are considered

repeated measurement; for drug-averaged drug-blind settings, each individual drug plays this role. Greenhouse-Geisser corrected p-values (Abdi, 2010) are reported in the findings. Data from Fig. 7 is available in Supplementary Table 1.

3.2 Random split evaluation (drug repurposing)

The random split was mainly introduced as a baseline setting to observe model behavior under the standard for the field experiment setup. We expected that the proposed change in the objective would not significantly influence this scenario, as the model has abundant information on the related ligands, unless we encounter a high imbalance of drug representations in batches. Indeed, as we can see, for CCLE (Fig.7A), with R^2 varying from 0.745 to 0.759 and MSE staying around 0.005, while two-way repeated measurements ANOVA did not detect statistically significant effects of loss function or sampling strategy on the performance. The situation with CTRP (Fig.7B) is very similar, except for the combination of multi-objective function and sequential sampling strategy. We hypothesize that due to the completely random proportion of classes (drugs) in training batches, the model has difficulties to learn the current objective (drug repurposing)

3.3 Drug-blind evaluation (virtual screening, bulk assessment)

As described in Fig.2, drug-blind evaluation corresponds to the virtual screening problem, where we assess the ligand's performance previously not seen by the model. In this scenario, performance metrics are calculated for the entire hold-out portion of the cross-validation set.

As drug-blind setup is much more challenging for ML model, we see a sharp drop in the performance. For the CCLE dataset, the best-performing combination is MOORLE loss function with mixed sampling with MAE = 0.09, while the rest of the values are 0.09, 0.11, 0.11 for MOORLE with sequential sampling, MSE with mixed sampling, and MSE with sequential sampling for CTRP dataset. We see a combination of sequential sampling and MOORLE loss function having a slight edge over the other variations with MAE = 0.056 against 0.064, 0.112, 0.067 for MOORLE with mixed sampling, MSE with mixed sampling, and MSE with sequential sampling for CTRP dataset.

For CCLE dataset, two-way repeated measurements ANOVA test highlight sampling strategy as a statistically significant performance factor (bul-letsp – value = $2.88 \cdot 10^{-2}$) for R^2 , as well as interaction of sample strategy and loss function (p – value = $6.93 \cdot 10^{-3}$). For MSE, the same combination of two factors mentioned above is statistically significant (p – value = $3.18 \cdot 10^{-2}$). For CTRP, both sampling and loss function were not a statistically significant factors for all performance metrics.

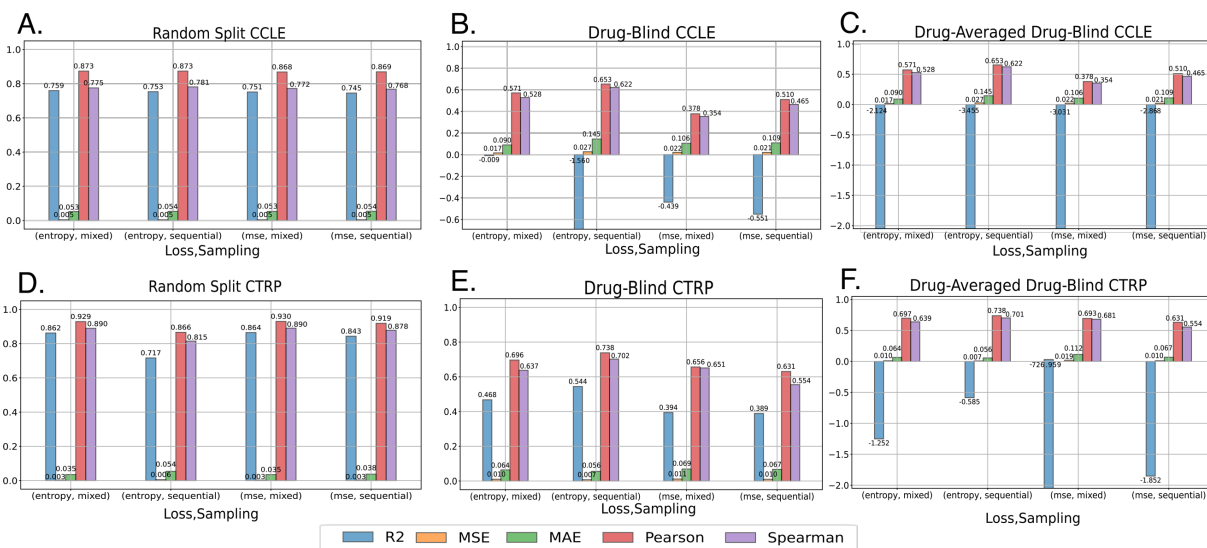


Fig. 7. Ablation study on CCLE and CTRP datasets. Performance metrics coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), Pearson correlation coefficient (Pearson), and Spearman correlation coefficient (Spearman) are recorded for each combination of factors – sampling strategy and loss function. Sampling strategies consist of sequential random sampling (denoted ‘sequential’ in the figure) and hybrid strategy introduced in 2.4 (‘mixed’ in the figure). Loss functions are represented by MSE (‘mse’ in the figure) and MOORLE - multi-objective loss function regularized by loss entropy (‘entropy’ in the figure). A. Random split evaluation strategy, CCLE dataset. B. Drug-blind split evaluation strategy, CCLE dataset. C. Drugwise evaluation under drug-blind split, CCLE dataset. D. Random split evaluation strategy, CTRP dataset. E. Drug-blind split setting, CTRP dataset. F. Drugwise evaluation under drug-blind split, CTRP dataset.

MOORLE

3.4 Drugwise scoring under drug-blind split (virtual screening, drug-specific assessment)

Most works that discuss drug-blind evaluation perform bulk assessment, as described in 3.3. However, to be completely thorough with our assessment, we first attempt to calculate the corresponding metric for each drug individually and then average the results. As we can see from Fig.7F and Fig.7E, the only measure significantly impacted by this procedure change is R^2 . However, even when the mean values of the majority of the performance metrics stay the same, taking a look at the problem from drug-by-drug perspective allows us to better reason over the influence that changes in ML model have on the performance. As MAE score remains very close to the previous scenario, we are making comparisons based on R^2 in this section.

For the CCLE dataset, the best-performing combination is MOORLE loss function with mixed sampling with $R^2 = -2.12$, while the rest of the values are -3.46, -3.03, -2.87 for MOORLE with sequential sampling, MSE with mixed sampling, and MSE with sequential sampling for CTRP dataset. We see a combination of sequential sampling and MOORLE loss function having a slight edge over the other variations with $R^2 = -0.59$ against -1.25, -726.96, -1.85 for MOORLE with mixed sampling, MSE with mixed sampling, and MSE with sequential sampling for CTRP dataset.

For the drug-averaged evaluation, two-way repeated measurements ANOVA test corroborated the statistically significant effect of adding both a new sampling strategy and the loss function on MSE value (p - value = $1.56 \cdot 10^{-3}$), and the outstanding impact of loss function on Pearson correlation coefficient (p - value = $9.16 \cdot 10^{-4}$) in CCLE dataset case (Fig.7C). For CTRP, the most influential factor on MSE value was sampling (p - value = $3.73 \cdot 10^{-5}$), with loss function also playing a significant role (p - value = $4.1 \cdot 10^{-4}$) (Fig.7F). As R^2 metric is unbound on the left, it exhibits great variability across individual drugs, which we can see in the examples of top-performing and worst-performing drugs from the Table 1.

Table 1. Ten drugs with the best and ten with the worst R^2 score from the CTRP dataset.

Drug Name	R^2	MSE	Pearson	Spearman
YK-4-279	0.839	0.006	0.903	0.871
FQI-1	0.831	0.003	0.929	0.865
Triazolothiadiazine	0.822	0.005	0.855	0.798
Parbendazole	0.818	0.005	0.860	0.856
KX2-391	0.809	0.008	0.798	0.794
BRD-K70511574	0.780	0.008	0.798	0.794
Tivantinib	0.775	0.003	0.929	0.865
Rigosertib	0.768	0.005	0.860	0.856
FQI-2	0.744	0.005	0.837	0.795
Vorinostat	0.743	0.006	0.903	0.871
TG-100-115	-10.028	0.014	0.367	0.229
BMS-270394	-10.154	0.014	0.367	0.229
Procarbazine	-10.204	0.014	0.367	0.229
MI-1	-10.378	0.014	0.367	0.229
Azacitidine	-10.408	0.014	0.367	0.229
CAY10576	-10.873	0.016	0.068	0.067
SZ4TA2	-16.939	0.014	0.367	0.229
BRD-K96431673	-21.211	0.014	0.367	0.229
ABT-888	-39.453	0.016	0.068	0.067
SJ-172550	-48.139	0.016	0.068	0.067

The examples of the best- and worst-performing drugs are from the drug-blind assessment of the model with sequential sampling and multi-objective loss function with entropy-based regularization.

4 Discussion

Current drug response prediction approaches implicitly rely on the ability of deep learning algorithms to find a true relationship between features and response values while simultaneously correcting for biases present in the dataset. However, as the amount of available data for this biomedical problem is limited, we investigated potential ways to improve predictive value by explicitly addressing the data imbalance problem.

O.Narykov et al.

The proposed multi-objective loss function with entropy regularization is model-agnostic and can be utilized both with classical ML algorithms and in Deep Learning. Classes, or domains \mathbf{D} , in the MOORLE loss function $\mathbf{L}(\mathbf{x}, \mathbf{D}, \mathbf{y})$ are interchangeable and can guide model to be aware of other biases encoded in data, e.g., gender, race, or age that regularly appear in biomedical datasets due to present sources of bias in our society. The proposed approach can be used to promote equitable outcomes in healthcare models.

One of the drawbacks of the proposed methodology is the dependency on the regularization coefficient. It should be derived in the inner cross-validation loop to achieve the best possible performance. However, nested cross-validation for the proposed datasets is exceptionally computationally demanding, increasing the runtime by the order of magnitude. Another limitation of this work is performing an ablation study on a single model. Our future direction is to perform a large-scale analysis of the community models with multiple loss functions and hyperparameter optimization.

Surprisingly, the mixed sampling strategy positively impacted the small, better-balanced CCLE dataset and was mostly detrimental to CTRP. It is possible that because of the large number of drugs in the latter, each balanced batch did not contain enough representative samples for the corresponding drug (class). Further adjustments in controlling the number of classes sampled in a single batch are needed. At the same time, the multi-objective loss function with entropy regularization was the primary influence for the CTRP dataset.

Funding

Argonne National Laboratory's work was supported by Leidos Biomedical Research, Inc. under Acknowledgement of Agreement No. A21154, through U.S. Department of Energy contract DE-AC02-06CH11357. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Conflict of Interest: none declared.

References

- Abarzadeh, M., Kivi, H.F. and Kojabadi, H.M. A modified SVM switching pattern for Z-source inverter. In, *2016 7th Power Electronics and Drive Systems Technologies Conference (PEDSTC)*. IEEE; 2016. p. 486-491.
- Abdi, H. The greenhouse-geisser correction. *Encyclopedia of research design* 2010;1(1):544-548.
- Barretina, J., et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;483(7391):603-607.
- Basu, A., et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;154(5):1151-1161.
- Bray, F., et al. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* 2021;127(16):3029-3030.
- Breiman, L. Random forests. *Machine learning* 2001;45:5-32.
- Cao, P., Zhao, D. and Zaiiane, O.R. A PSO-based cost-sensitive neural network for imbalanced data classification. In, *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTh, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers 17*. Springer; 2013. p. 452-463.
- Chapelle, O., Haffner, P. and Vapnik, V.N. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks* 1999;10(5):1055-1064.
- Chawla, N.V., et al. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 2002;16:321-357.
- Cheminformatics, K. https://chm.kode-solutions.net/products_dragon.php
- Chu, T., et al. Graph Transformer for drug response prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2022;20(2):1065-1072.
- Cronin, K.A., et al. Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. *Cancer* 2018;124(13):2785-2800.
- Haixiang, G., et al. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications* 2017;73:220-239.
- Hastie, T., et al. Multi-class adaboost. *Statistics and its Interface* 2009;2(3):349-360.
- He, D., et al. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence* 2022;4(10):879-892.
- Hernandez-Orallo, J. Probabilistic reframing for cost-sensitive regression. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2014;8(4):1-55.
- Jia, P., et al. Deep generative neural network for accurate drug response imputation. *Nature communications* 2021;12(1):1740.
- Jiang, L., et al. DeepTTA: a transformer-based model for predicting cancer drug response. *Briefings in Bioinformatics* 2022;23(3):bbac100.
- Joo, M., et al. A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. *International journal of molecular sciences* 2019;20(24):6276.
- Joshi, M.V., Kumar, V. and Agarwal, R.C. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In, *Proceedings 2001 IEEE international conference on data mining*. IEEE; 2001. p. 257-264.
- Ke, G., et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 2017;30.
- Kim, S., et al. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* 2019;47(D1):D1102-D1109.
- Kingma, D.P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* 2014.
- Kurilov, R., Haibe-Kains, B. and Brors, B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Scientific reports* 2020;10(1):1-11.
- Liu, Q., et al. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* 2020;36(Supplement_2):i911-i918.

MOORLE

- Lu, J., Chen, M. and Qin, Y. Drug-induced cell viability prediction from LINCS-L1000 through WRFEN-XGBoost algorithm. *BMC bioinformatics* 2021;22:1-18.
- McGaughey, G.B., et al. Comparison of topological, shape, and docking methods in virtual screening. *Journal of chemical information and modeling* 2007;47(4):1504-1519.
- McNamee, R. Regression modelling and other methods to control confounding. *Occupational and environmental medicine* 2005;62(7):500-506.
- Menden, M.P., et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one* 2013;8(4):e61318.
- Moukafih, Y., Ghogho, M. and Smali, K. Supervised contrastive learning as multi-objective optimization for fine-tuning large pre-trained language models. In, *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2023. p. 1-5.
- Narykov, O., et al. Entropy-Based Regularization on Deep Learning Models for Anti-Cancer Drug Response Prediction. In, *Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*. 2023. p. 121-122.
- Nguyen, T., et al. Graph convolutional networks for drug response prediction. *IEEE/ACM transactions on computational biology and bioinformatics* 2021;19(1):146-154.
- Oskooei, A., et al. PaccMann: prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. *arXiv preprint arXiv:1811.06802* 2018.
- Park, Y. and Marcotte, E.M. Flaws in evaluation schemes for pair-input computational predictions. *Nature methods* 2012;9(12):1134-1136.
- Partin, A., et al. Deep learning methods for drug response prediction in cancer: predominant and emerging trends. *Frontiers in Medicine* 2023;10:1086097.
- Potvin, P.J. and Schutz, R.W. Statistical power for the two-factor repeated measures ANOVA. *Behavior Research Methods, Instruments, & Computers* 2000;32(2):347-356.
- Raghuwanshi, B.S. and Shukla, S. Class-specific cost-sensitive boosting weighted ELM for class imbalance learning. *Memetic Computing* 2019;11:263-283.
- Rezvani, S. and Wang, X. A broad review on class imbalance learning techniques. *Applied Soft Computing* 2023:110415.
- Sharifi-Noghabi, H., et al. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 2019;35(14):i501-i509.
- Song, J., Lu, X. and Wu, X. An improved adaboost algorithm for unbalanced classification data. In, *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE; 2009. p. 109-113.
- Tao, Y., et al. Predicting drug sensitivity of cancer cell lines via collaborative filtering with contextual attention. In, *Machine Learning for Healthcare Conference*. PMLR; 2020. p. 660-684.
- Teich, J. Pareto-front exploration with uncertain objectives. In, *International Conference on Evolutionary Multi-Criterion Optimization*. Springer; 2001. p. 314-328.
- Torgo, L., et al. Smote for regression. In, *Portuguese conference on artificial intelligence*. Springer; 2013. p. 378-389.
- Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 1988;28(1):31-36.
- Yanagawa, E., et al. Chemosensitivity tests in colorectal cancer patients. *The Japanese journal of surgery* 1989;19:432-438.
- Zhu, Y., et al. Ensemble transfer learning for the prediction of anti-cancer drug response. *Scientific reports* 2020;10(1):18040.
- Zhu, Y., et al. Enhanced co-expression extrapolation (COXEN) gene selection method for building anti-cancer drug response prediction models. *Genes* 2020;11(9):1070.
- Zhu, Y., et al. Converting tabular data into images for deep learning with convolutional neural networks. *Scientific reports* 2021;11(1):11325.