Briefings in Bioinformatics, 2025, **26(2)**, bbaf134
https://doi.org/10.1093/bib/bbaf134
Problem Solving Protocol

Data imbalance in drug response prediction: multi-objective optimization approach in deep learning setting

Oleksandr Narykov[1,*], Yitan Zhu[1], Thomas Brettin[1], Yvonne A. Evrard[2], Alexander Partin[1], Fangfang Xia[1], Maulik Shukla[1], Priyanka Vasanthakumari[1], James H. Doroshow[3], Rick L. Stevens[1,4]

[1]Computing, Environment and Life Sciences, Argonne National Laboratory, 9700 S Cass Ave, Lemont, IL 60439, United States
[2]Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD 21702, United States
[3]Developmental Therapeutics Branch, National Cancer Institute, 31 Center Dr, Bethesda, MD 20892, United States
[4]Department of Computer Science, The University of Chicago, 5730 S Ellis Ave, Chicago, IL 60637, United States

[*]Corresponding author. Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL 60439, United States. E-mail: onarykov@anl.gov

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 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 application domains like computer vision or natural language processing domains, limiting current learning capabilities. To combat this issue and improves 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 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 achieve desirable outcomes in the healthcare field.

Keywords: drug response prediction; virtual screening; machine learning; multi-objective optimization; deep learning

Introduction

Cancer is a widely spread genetic disease family with a common characteristic of uncontrolled cell growth and proliferation [1, 2]. This set of complex genetic disorders is highly heterogeneous and notoriously difficult to combat. Artificial intelligence technologies are being incorporated into this field to facilitate our ability to treat patients. For example, machine learning (ML) systems use 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 [3]. 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. 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 [4].

The DRP field is abundant and contains multiple models based on traditional ML approaches – Random Forest [5], AdaBoost [6], XGBoost [7], LightGBM [8], and Support Vector Machine [9]. The quantitative structure–activity relationship methods, such as kernelized Bayesian matrix factorization (KBMF), were used for the drug recommendation system [10]. Conformal prediction methods were used to assign reliability of the model prediction in DRP setting [11]. 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 [12], 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) [13–20].

A pair-input setting introduces known model evaluation pitfalls [21, 22], 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

Received: September 19, 2024. **Revised:** February 7, 2025. **Accepted:** February 18, 2025

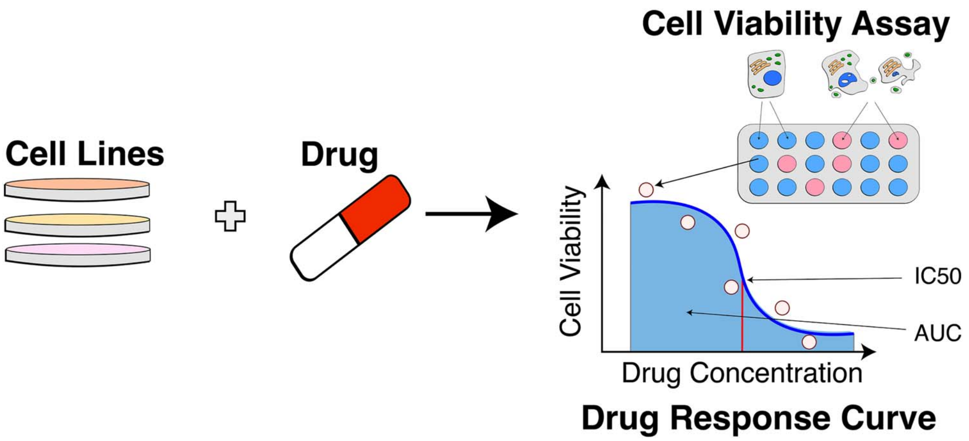

Figure 1. Drug response experiment. Multiple cell viability assays (a combination of membrane-permeable and membrane-impermeable fluorescent proteins that highlight tumor inhibition) are integrated into a single drug response measurement based on the hill-slope model, graphical interpretation of IC50 and AUC drug response measures [23].

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.

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 [4, 24].

Drug discovery [25] setting is one of the most challenging applications for DRP models, as response variability between drugs is much higher than between cell line response variations [26]. 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 the 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 the 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 data structures. While multiple works address the confounding factors for the prediction problem, most of them focus on information from different data modalities, e.g., copy number variations or mutation data [27–29]. However, it is necessary to understand limitation of the proposed evaluation. The drug-blind split approach in our evaluation aims to simulate virtual screening scenarios, but as noted by [30], the resulting performance metrics may not directly reflect real-world effectiveness. This limitation stems from the inherent chemical property variations across training, validation, and test datasets, which persist even when using advanced partitioning methods like scaffold splits or chemistry-based clustering. When test compounds share minimal chemical similarity with the training data, models must make out-of-distribution predictions, which remains challenging for all machine learning approaches. Therefore, for each specific screening application, training and testing splits must be carefully designed to ensure reliable predictions for the target drug classes within the screening library. The model results may vary significantly across different splits and should be compared only to the results obtained on the same split or at least similar one with the same data distribution.

In this work, we are investigating the benefits of explicitly addressing complex substructures arising from the pair-input nature of data with a focus on improving the drug-blind response prediction performance for new drug discovery (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.

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 a response value, such as the area under the dose–response curve (AUC) or the half-maximal inhibitory concentration (IC50) [31], 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 [4].

Methods

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) [32] and Cancer Cell Line Encyclopedia (CCLE) [33]. CCLE dataset contains 8950 experiments based on 474 unique cell lines and 24 drugs. Data comprises

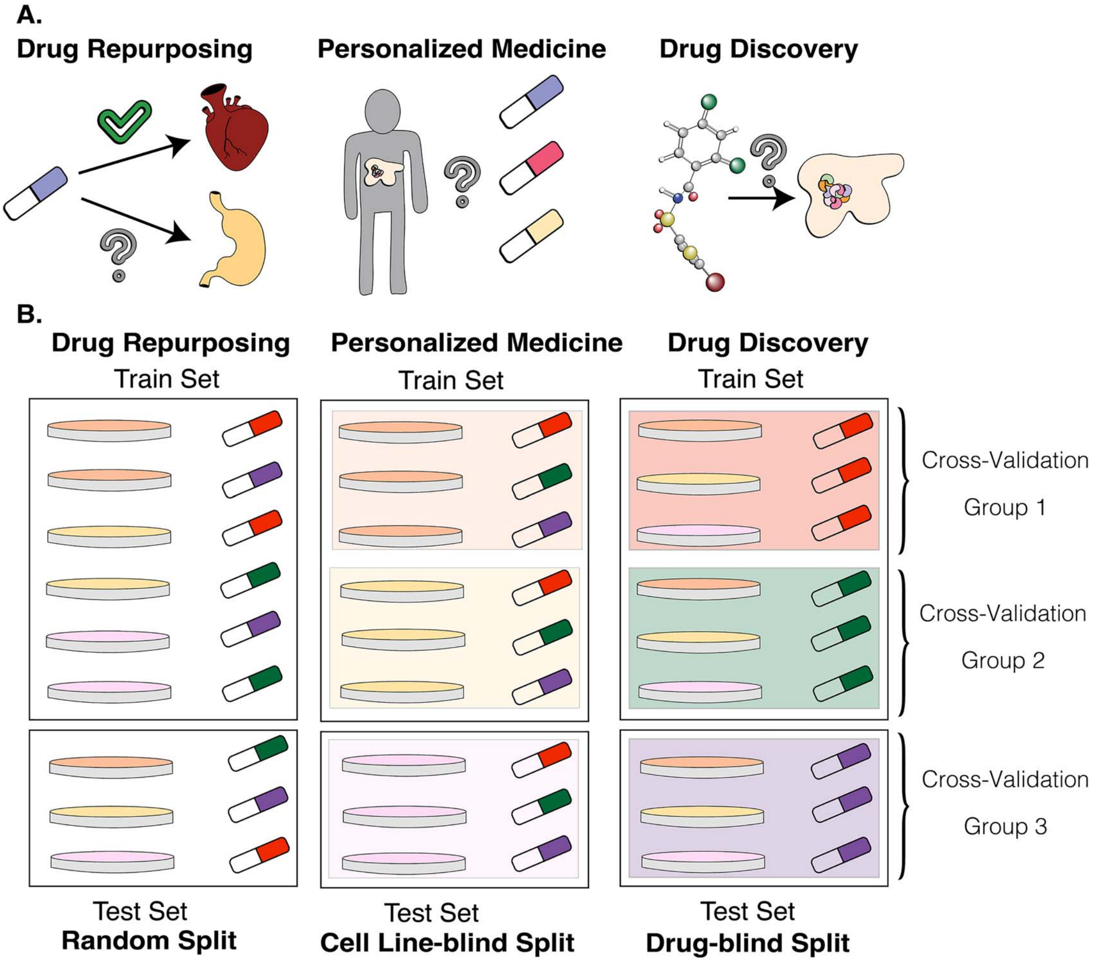

Figure 2. Drug response prediction application areas. A. Real-world tasks that benefit from DRP models. B. Corresponding splits of pair-input entries from drug response datasets. Each petri dish color corresponds to a unique cell line, and each color of the drug corresponds to a unique drug.

RNA-Seq gene expressions, corresponding compounds, and drug response for combining 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. We selected those datasets because they correspond to two different settings – a slight imbalance in the relatively small dataset and a significant imbalance in the large dataset (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.

Drug-level information for both datasets comes in the form of molecular fingerprints computed via Dragon v.7.0 [34] and Simplified Molecular-Input Line-Entry System (SMILES) [35] entries obtained from the PubChem [36] and the web form of Developmental Therapeutic Program (DTP) (https://dtp.cancer.gov/). Data availability: https://zenodo.org/records/13787609 [37].

Learning from imbalanced data

Data imbalance is a common problem in machine learning arising due to the limited amount of available learning data [38, 39]. 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 [40] technique, and its adaptation to the regression problem [41]. 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 less frequent 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 less frequent classes becomes inflated, which may introduce systematic errors to a model [39].

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 [5] and SVM [9], modify the loss function, (particle-swarm optimization network [42], zSVM [43], or refine boosting approaches (AdaC1-AdaC3 [44], RareBoost [45], BABoost [46]. For regression, probability-based methods such as reframing were introduced. This approach focuses on adapting to estimated outputs depending on the context [47].

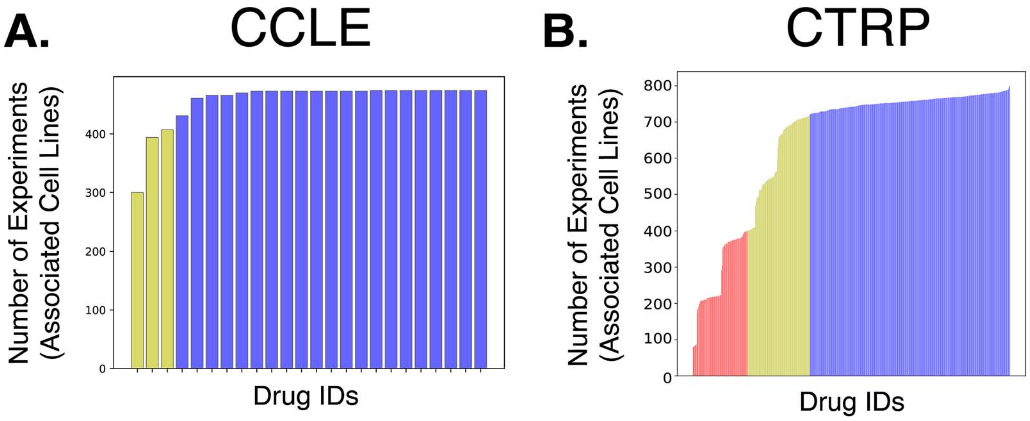

Figure 3. Number of experiments associated with each drug in a dataset. A. CCLE dataset. The highest number of experiments related to a single drug is 474. 12.5% of the drugs (yellow) have a number of experiments related to them that is less than 90% of the highest number. The rest of the drugs are depicted 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 that 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.

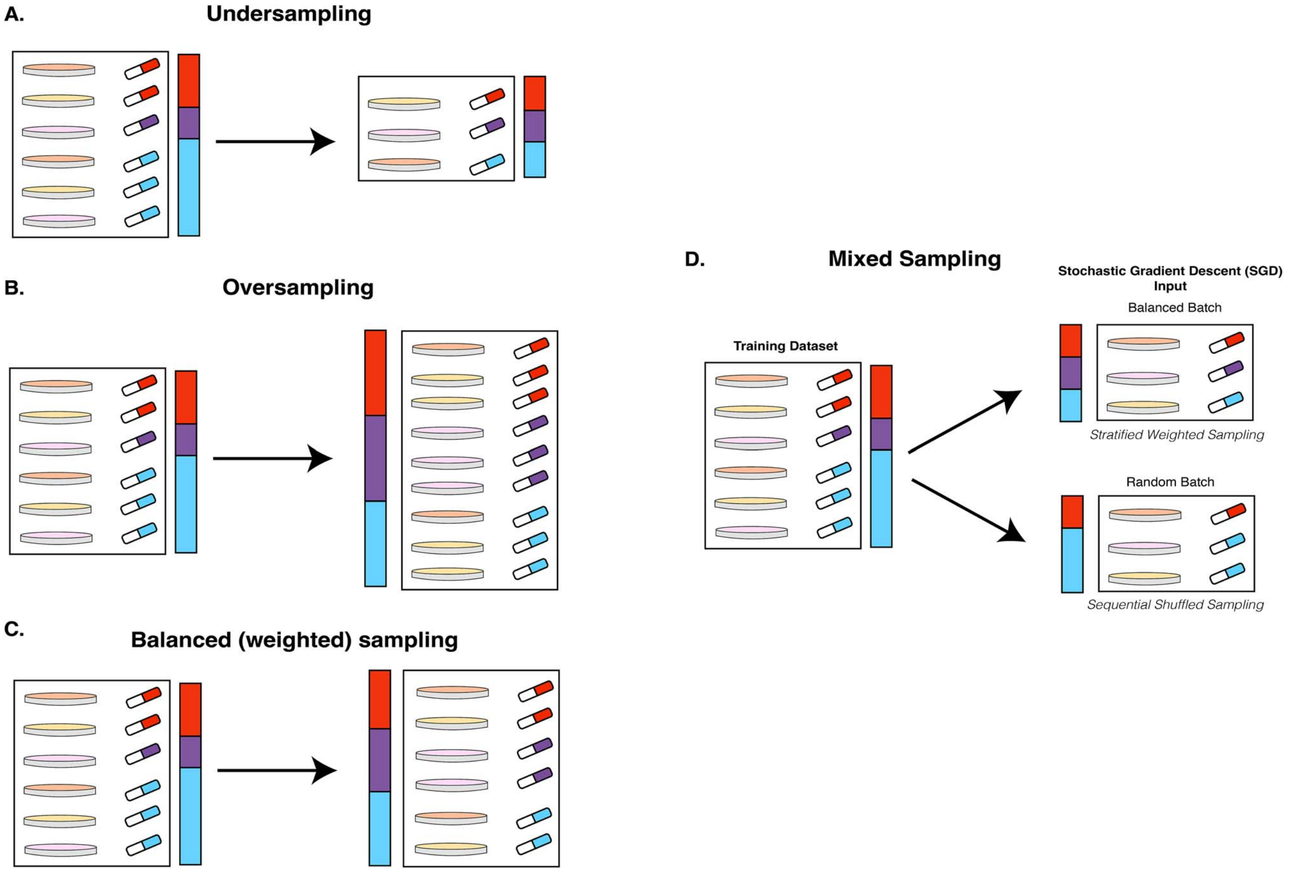
A.
Undersampling
B.
Oversampling
C.
Balanced (weighted) sampling
D.
Mixed Sampling
Training Dataset
Stochastic Gradient Descent (SGD) Input
Balanced Batch
Stratified Weighted Sampling
Random Batch
Sequential Shuffled Sampling

Figure 4. 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 subgroups 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. D. Mixed sampling scheme for stochastic gradient descent-based algorithms, a variation of weighted sampling.

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^2_{avg}. 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

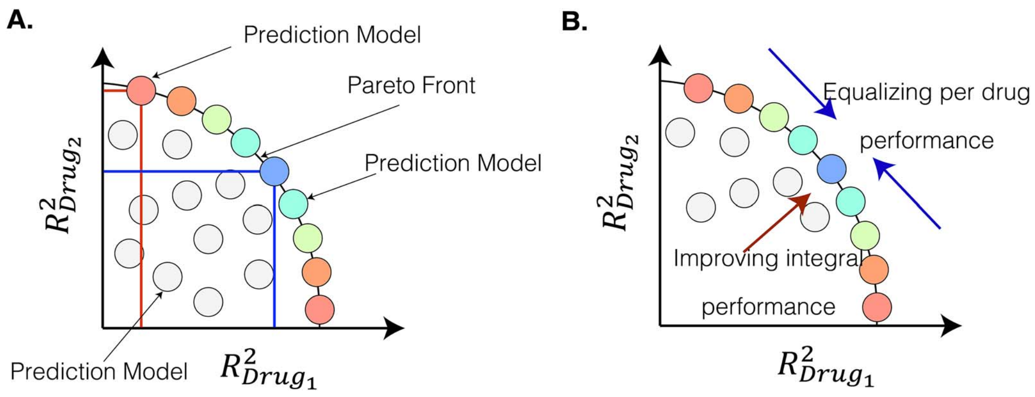

Figure 5. Pareto front of ML models in the space composed of individual drug performance metrics. Colored nodes indicate ML models belonging to the Pareto front, i.e., having integral performance close to the maximum known value. Grey points represent ML models that have worse integral performance.

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 MSE loss function:

$$MSE\left(f(\mathrm{x}), y\right) = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}\left(f\left(\mathrm{x}_{\mathrm{i}}\right) - y_{\mathrm{i}}\right)^2 \tag{2}$$

where $\mathbf{x}$ is the set of features, n – total number of datapoints, f – prediction model, y – ground truth value, and $\hat{y} = f\left(\mathrm{x}_{\mathrm{i}}\right)$.

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 $R^2_{Drug_i}$. 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 $< R^2_{Drug_1}, R^2_{Drug_2} >$ defines the coordinates of the corresponding machine learning model. On top of individual decomposition into performance evaluations $R^2_{Drug_i}$ 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 non-dominated data points across different axes 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 [48]. 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 skewed towards some drugs and thus provide better generalizability between drugs. The reason is that such a model can provide a better association between the unique chemical characteristics of the small compounds and the biological sample's features [37]. Figure 5B provides a graphical visualization of our objective—to maximize the integral performance of the model and to balance individual drug scores.

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

$$MSE_{d_i} = MSE\left(f\left(\mathrm{x}_{d_i}\right), y_{d_i}\right) \tag{3}$$

where $\mathbf{x}$ is the set input of features derived both from cell line and drug information, d_i is the i-th drug, $f(.)$ 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:

$$p_{d_i} = \frac{\exp\left(MSE_{d_i}\right)}{\sum_j \exp\left(MSE_{d_j}\right)} \tag{4}$$

$$P = \left\{p_{d_i} : d_i \in D\right\} \tag{5}$$

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:

$$H(\mathrm{P}) = \sum_{\mathrm{x}_j \in X} -p_j \log\left(p_j\right) \tag{6}$$

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

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

where a is a regularization coefficient, $H(\mathbf{X})$ 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. However, due to the dependency on the model in evaluating $H(P)$ components, the resulting $\boldsymbol{L}(\mathbf{x}, \boldsymbol{D}, \mathbf{y})$ may no longer be convex, even though it consists of a linear combination of convex functions. It is convex only if the model stays unchanged, which is not the case with the continuous weights update or boosting tree addition. It results in a versatile model-agnostic loss function that can be utilized both in classical ML models and deep learning settings, but it can allow more complex deep learning models to undergo more detailed tuning and make the best out of the automatic feature extraction procedure.

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 [49]. 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. See supplementary materials for the detail. We use the mixed-sampling approach as a representative data weighting method to be one of the baselines for the comparison, along with the MSE loss function.

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 (Eq. 7) into the recent state-of-the-art model DeepTTA [20].

DeepTTA consists of three main components. One is an attention-based SMILES encoder subnetwork, and the other is a fully connected neural network (FCNN) [15]. The last part of the network concatenates encodings for drugs and biological samples and performs regression.

Results

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 cannot appear in the training and testing sets simultaneously. It ensures that no information about a 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.

First, we perform hyperparameter tuning in the drug-blind setting of the regularization weight a (Eq. 7) for the MOORLE loss function using validation splits and document its effect on model performance for the test splits (Fig. 6). We consider only drug-blind setting for these experiments due to the high computational cost. Second, 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. We use median value across 10-fold CVs for the visualization purpose due to large variations for R^2 values. The same information is presented via a boxplot that reflects the overall data distribution and supports our conclusions in Supplementary Fig. 1.

To estimate the statistical significance of the effect that proposed strategies have on the drugs we apply two-way repeated measurements ANOVA [50] 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 [51] are reported in the findings.

Insights from hyperparameter tuning for the drug-blind setting

We evaluated the model's performance with the regularization parameter a ranging from 0 to 6 using 10-fold cross-validation on CCLE and CTRP datasets. The data includes averaged performance metrics for bulk data from each cross-validation split (Fig. 6A, D), and averaged model performance for each individual drug (Fig. 6B, E) and drug's mechanism of action (MOA) (Fig. 6C, F) on the test splits. In each case, we estimate the average of the metric and standard deviation, comparing those to the baseline performance of the model with the MSE loss (denoted in red color on Fig. 6). In addition to drug averaging, we analyze MOA grouping to investigate regularization effects on larger meaningful drug subgroups.

We find out that, generally, the effects of the regularization are positive both for CCLE and CTRP datasets; however, due to the subgroups-based nature of the regularization, its weight should be explored and adjusted based on a specific dataset because they may be non-monotonous. Both the CCLE and CTRP weight interval of [3.0,4.0] produce optimal values, improving average performance and reducing standard deviation. Effects on MSE and R^2 mirror each other, which was not an obvious conclusion due to the unbound left side of the R^2 metric. It means that the addition of the MOORLE regularization component can robustly improve its average ability to explain variations for each individual drug. The variability in the results observed for CTRP for the regularization coefficient a of the value of 5.0 highlights the non-linear nature of the proposed loss function and emphasizes the need for fine-tuning it on a validation set. The reduced variability for CCLE is likely due to the smaller dataset size and more balanced drug representation.

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

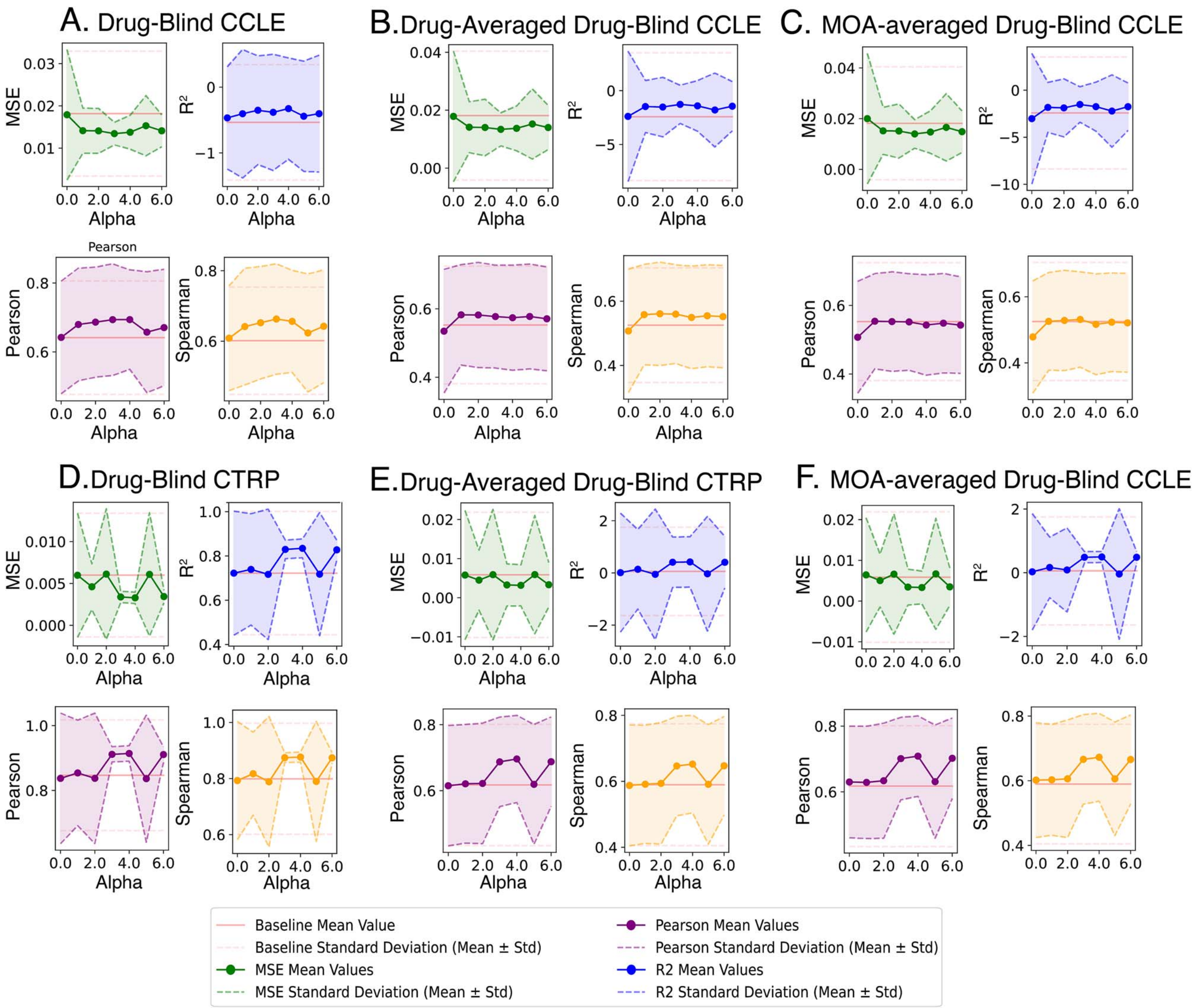

Figure 6. Comprehensive assessment of the regularization weight of MOORLE loss function on model performance. Each assessment includes the coefficient of determination (R^2), mean squared error (MSE), Pearson correlation coefficient (Pearson), and spearman correlation coefficient (spearman) and their standard deviation for the varying value of regularization weight a in MOORLE loss and compared to baseline performance of the model with sequential sampling and MSE loss function. A. Assessment of drug-blind cross-validation splits of the CCLE. B. Assessment of the drug-averaged performance for CCLE. C. Assessment of the MOA-averaged model performance for CCLE. D. Assessment of drug-blind cross-validation splits of the CTRP. B. Assessment of the drug-averaged performance for CTRP. C. Assessment of the MOA-averaged model performance for CTRP.

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 learning the drug repurposing objective. The comprehensive list of scores for Fig. 7 is in Supplementary Table 1.

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 the drug-blind setup is much more challenging for the ML models, 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 = 0.069, while the rest of the values vary from 0.082 to 0.099 for MOORLE with sequential sampling, MSE with mixed sampling, and MSE with sequential sampling. For the CTRP dataset, we see a combination of sequential sampling and MOORLE loss function having a slight edge over the other variations with MAE = 0.039 against 0.052, 0.060, 0.043 for MOORLE with mixed sampling, MSE with mixed sampling, and MSE with sequential sampling for CTRP dataset.

For the CCLE dataset, the two-way repeated measurements ANOVA test highlights no statistical significance of the factors on performance measures for R^2 and MSE. For CTRP, only the sampling strategy was above the significance threshold ($\mathbf{p} - \mathbf{value} = 4.67 \cdot 10^{-3}$).

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

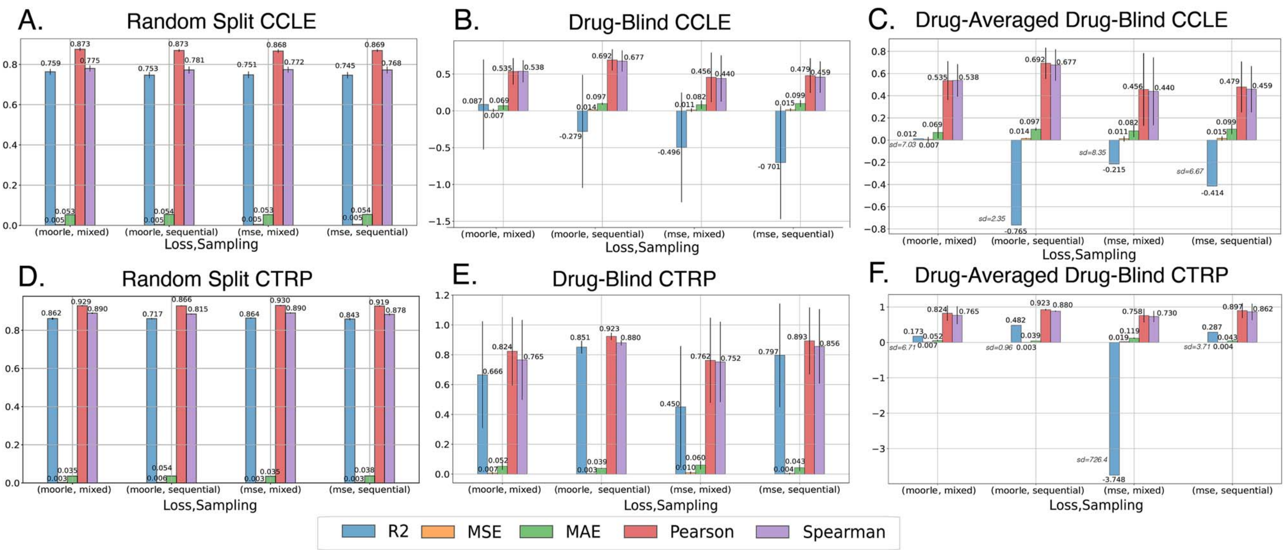

Figure 7. Ablation study on CCLE and CTRP datasets. Performance metrics including coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), Pearson correlation coefficient (Pearson), and spearman correlation coefficient (spearman) are recorded for each combination of factors – Sampling strategy and loss function. Each bar represent median of a 10-fold CV run results. Black bar represents one standard deviation. 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 ('MOOLRE' 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.

the corresponding metric for each drug individually and then average the results. As we can see from Fig. 7E and F, the only measure significantly impacted by this procedure change is R^2. In the case of a drug-wise assessment, this measure reflects a goodness of fit of the trained model for each given drug. However, even when the mean values of the majority of the performance metrics stay the same, taking a look at the problem from a 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 $R^2 = 0.12$, while the rest of the values are -0.765, -0.215, -0.414 for MOORLE with sequential sampling, MSE with mixed sampling, and MSE with sequential sampling for CCLE dataset. For the CTRP dataset, we see a combination of sequential sampling and MOORLE loss function having a significant edge over the other variations with $R^2 = 0.482$ against 0.173, -3.748, 0.287 for MOORLE with mixed sampling, MSE with mixed sampling, and MSE with sequential sampling.

For the drug-averaged evaluation on CTRP dataset (Fig. 7F), two-way repeated measurements ANOVA test corroborated the statistically significant effect of both sampling strategy and loss, adding both a new sampling strategy ($\mathbf{p-value} = 4.41 \cdot 10^{-2}$), and the loss function ($\mathbf{p-value} = 3.39 \cdot 10^{-2}$) on R^2 value, as well as their combination ($\mathbf{p-value} = 2.33 \cdot 10^{-2}$) and the outstanding impact of loss function on MSE score: sampling ($\mathbf{p-value} = 2.88 \cdot 10^{-280}$), loss function ($\mathbf{p-value} = 1.26 \cdot 10^{-123}$), and their combination ($\mathbf{p-value} = 1.27 \cdot 10^{-37}$). For CCLE, no factors were found statistically significant (Fig. 7C). The list of scores for the top-performing drugs are in Supplementary Table 3. The drugs and MOAs with the best scores gains are listed in Fig. 8.

We also conducted experiments on a 5-fold drug-blind cross-validation split using popular DRP methods – LightGBM [8], GraphDRP [16], and fully-connected neural network (FCNN) [12] in addition to DeepTTC with a regularization coefficient $\alpha = 4$ for the DeepTTC, GraphDRP, and FCNN models, and $\alpha = 0.25$ for LightGBM (Table 1). We report the average value from the folds and the standard deviation. We generated a training dataset for each model individually to accommodate for differences in the preprocessing steps, so the performance metrics should not be compared across models but only used to assess the effect of the loss function. The difference in the number of input features in the models contributes to the different outcomes for the CCLE dataset. The results showcase the utility of the MOORLE loss function for DeepTTC and GraphDRP, while its utility for boosting and FCNN remains inconclusive. This behavior may be attributed to the fact that DeepTTC and GraphDRP utilize flexible representation of the small compounds using explainable substructure fingerprints [52] and graphical representation correspondingly, while LightGBM and FCNN rely on Dragon 7 descriptors.

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 data imbalance. 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.

Our contributions include the proposal of the novel model-agnostic multi-objective loss function with entropy-based regularization, and can be utilized both with classical ML algorithms and in Deep Learning. Classes, or domains $\boldsymbol{D}$, in the MOORLE loss function $\boldsymbol{L}(\boldsymbol{x}, \boldsymbol{D}, \boldsymbol{y})$ are interchangeable and can guide model to be aware of other imbalances encoded in data, e.g., sex or age, that regularly appear in biomedical datasets. The proposed approach

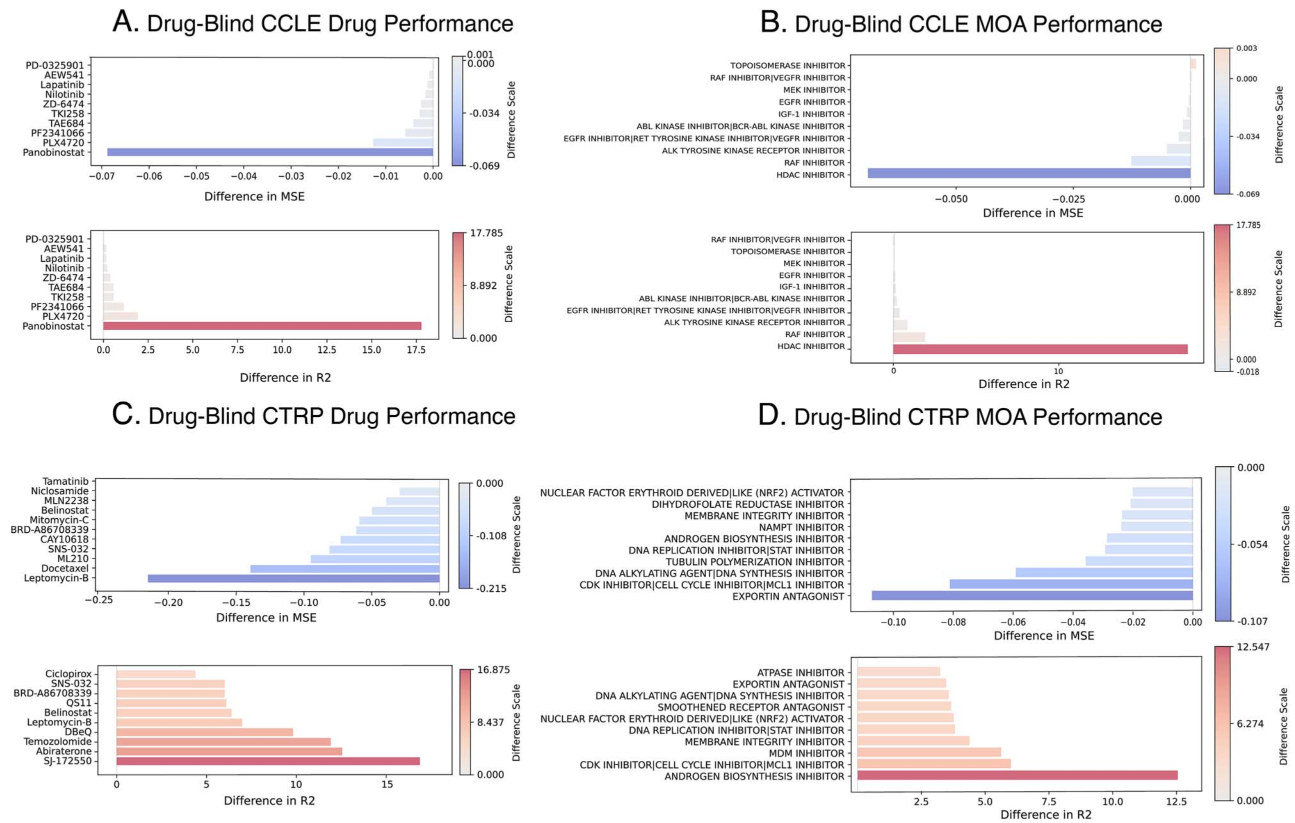

Figure 8. Top-10 drugs' and MOAs' performance improvements. Each panel visualizes the performance improvement for a specific drug or MoA, with blue bars indicating positive improvement and red bars indicating negative change. The horizontal axis represents the magnitude of improvement and the vertical axis corresponds to the individual item, a drug or MOA. We can see that improvement patterns across prediction groupings based on individual drug and molecular MOA are on a comparable scale. A. Top-10 drug performance gains on drug-blind cross-validation splits of the CCLE. B. The drug-averaged performance for CCLE. C. The MOA-averaged model performance for CCLE. D. Drug-blind cross-validation splits performance of the CTRP.

Table 1. Performance of the community models on 5-fold drug-blind cross-validation.

Dataset	CCLE		CTRP	
Performance Metric	**R^2**		**R^2**	
Loss function	**MSE Loss**	**MOORLE**	**MSE Loss**	**MOORLE**
Model				
DeepTTC	−0.48 ± 1.034	0.41 ± 0.170	0.79 ± 0.059	0.84 ± 0.021
LightGBM	−0.92 ± 0.121	−0.39 ± 0.107	0.71 ± 0.073	0.66 ± 0.094
GraphDRP	0.64 ± 0.026	0.66 ± 0.019	0.72 ± 0.011	0.73 ± 0.009
FCNN	0.26 ± 0.078	0.25 ± 0.065	0.83 ± 0.006	0.82 ± 0.004

can be used to promote equitable outcomes in healthcare models. We conduct comprehensive assessments on the effects of regularization weight for drug response prediction problems, compare outcomes to regular MSE loss and weighted sampling, and study a combination of the latter with new loss function via ablation analysis. The MOORLE loss improves performance metrics and reduces standard deviation on all datasets. On top of it, its main feature is the aggregation method for data subgroup losses, which makes it extremely versatile, with the ability to switch out basic MSE loss for any other non-negative loss function. In case when we want our model to prioritize a specific drug group (e.g., ligands for the single protein target) it is possible to adjust individual loss weights to skew performance towards drugs of interest.

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 an order of magnitude. Our future direction is to perform a large-scale analysis of the community models with multiple loss functions and extensive hyperparameter optimization. Though our evaluation using drug-blind splits is intended to approximate a virtual screening problem, the performance evaluation may not be directly translatable to real-world applications, as cautioned by [30]. This happens due to variability in the chemical properties of the compounds that end up in training, validation, and test sets, even in the alternative grouped cross-validation settings like scaffold split and chemistry-based clustering. In case there is no similarity between chemical compounds in training and testing data, the model will be making out-of-distribution predictions, a challenging task for every ML approach. Thus, training and testing splits should be carefully tuned for each real-world application to ensure robust results on the drug class of interest in a screening library. The same considerations apply to the new domain applications, such as drug-target prediction [53] or precision medicine.

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 are needed to control the number of classes sampled in a single batch. At the same time, the multi-objective loss function with entropy regularization was the primary influence for the CTRP dataset.

Key Points

- Data imbalance in dataset subgroups results in systematic errors in machine learning model
- Traditional loss functions like Mean Squared Error do not distinguish between subgroups
- Multi-objective optimization (MOO) gives researchers control over dataset subgroups' treatment
- The proposed model-agnostic loss function (Multi-Objective Optimization Regularized by Loss Entropy) incorporates MOO principles and significantly improves the generalization capability of a drug response prediction artificial intelligence (AI) model in drug discovery setting
- Incorporating MOO principles in AI models can help achieve more equitable outcomes in healthcare

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

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. This research has been funded in whole or in part with federal funding by the NCI-DOE Collaboration established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health, Cancer Moonshot Task Order No. 75N91019F00134 and under Frederick National Laboratory for Cancer Research Contract 75N91019D00024. 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.

Date availability

The full dataset is shared via zenodo: https://zenodo.org/records/13787609. The source code is available on GitHub: https://github.com/AlexandrNP/MOORLE.

References

1. Cronin KA, Lake AJ, Scott S. *et al.* Annual report to the nation on the status of cancer, part I: National cancer statistics. *Cancer* 2018;**124**:2785–800. https://doi.org/10.1002/cncr.31551
2. Bray F, Laversanne M, Weiderpass E. *et al.* The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* 2021;**127**:3029–30. https://doi.org/10.1002/cncr.33587
3. Sharifi-Noghabi H, Zolotareva O, Collins CC. *et al.* MOLI: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 2019;**35**:i501–9. https://doi.org/10.1093/bioinformatics/btz318
4. Partin A, Brettin TS, Zhu Y. *et al.* Deep learning methods for drug response prediction in cancer: Predominant and emerging trends. *Front Med* 2023;**10**:1086097. https://doi.org/10.3389/fmed.2023.1086097
5. Breiman L. Random forests. *Machine learning* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324
6. Hastie T, Rosset S, Zhu J. *et al.* Multi-class adaboost. *Statistics and its Interface* 2009;**2**:349–60. https://doi.org/10.4310/SII.2009.v2.n3.a8
7. Lu J, Chen M, Qin Y. Drug-induced cell viability prediction from LINCS-L1000 through WRFEN-XGBoost algorithm. *BMC bioinformatics* 2021;**22**:1–18. https://doi.org/10.4310/SII.2009.v2.n3.a8
8. Ke G, Meng Q, Finley T. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 2017;**30**:3149–57.
9. Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. *IEEE Trans Neural Netw* 1999;**10**:1055–64. https://doi.org/10.1109/72.788646
10. Ammad-Ud-Din M, Georgii E, Gonen M. *et al.* Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J Chem Inf Model* 2014;**54**:2347–59. https://doi.org/10.1021/ci500152b
11. Hernández-Hernández S, Vishwakarma S, Ballester P. Conformal prediction of small-molecule drug resistance in cancer cell lines. In: *Proceedings of the Conformal and Probabilistic Prediction with Applications*, pp. 92–108, 2022.
12. Menden MP, Iorio F, Garnett M. *et al.* Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PloS One* 2013;**8**:e61318. https://doi.org/10.1371/journal.pone.0061318
13. Joo M, Park A, Kim K. *et al.* A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. *Int J Mol Sci* 2019;**20**:6276. https://doi.org/10.3390/ijms20246276
14. Liu Q, Hu Z, Jiang R. *et al.* DeepCDR: A hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* 2020;**36**:i911–8. https://doi.org/10.1093/bioinformatics/btaa822
15. Zhu Y, Brettin T, Xia F. *et al.* Converting tabular data into images for deep learning with convolutional neural networks. *Sci Rep* 2021;**11**:11325. https://doi.org/10.1038/s41598-021-90923-y
16. Nguyen T, Nguyen GT, Nguyen T. *et al.* Graph convolutional networks for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**19**:146–54.

17. Chu T, Nguyen TT, Hai BD. *et al.* Graph transformer for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**20**:1065–72. https://doi.org/10.1109/TCBB.2022.3206888
18. Oskooei A, Born J, Manica M. *et al.* PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. *arXiv preprint arXiv* 2018;1811.06802.
19. Tao Y, Ren S, Ding MQ. *et al.* Predicting drug sensitivity of cancer cell lines via collaborative filtering with contextual attention. In: *Proceedings of the Machine Learning for Healthcare Conference*, pp. 660–84, 2020.
20. Jiang L, Jiang C, Yu X. *et al.* DeepTTA: A transformer-based model for predicting cancer drug response. *Brief Bioinform* 2022;**23**:bbac100. https://doi.org/10.1093/bib/bbac100
21. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* 2012;**9**:1134–6. https://doi.org/10.1038/nmeth.2259
22. Narykov O, Johnson NT, Korkin D. Predicting protein interaction network perturbation by alternative splicing with semi-supervised learning. *Cell Rep* 2021;**37**:110045. https://doi.org/10.1016/j.celrep.2021.110045
23. Yanagawa E, Nishiyama M, Saeki T. *et al.* Chemosensitivity tests in colorectal cancer patients. *Jpn J Surg* 1989;**19**:432–8. https://doi.org/10.1007/BF02471624
24. Zhu Y, Brettin T, Evrard YA. *et al.* Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci Rep* 2020;**10**:18040. https://doi.org/10.1038/s41598-020-74921-0
25. McGaughey GB, Sheridan RP, Bayly CI. *et al.* Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* 2007;**47**:1504–19. https://doi.org/10.1021/ci700052x
26. Zhu Y, Brettin T, Evrard YA. *et al.* Enhanced co-expression extrapolation (COXEN) gene selection method for building anti-cancer drug response prediction models. *Genes* 2020;**11**:1070. https://doi.org/10.3390/genes11091070
27. Jia P, Hu R, Pei G. *et al.* Deep generative neural network for accurate drug response imputation. *Nat Commun* 2021;**12**:1740. https://doi.org/10.1038/s41467-021-21997-5
28. McNamee R. Regression modelling and other methods to control confounding. *Occup Environ Med* 2005;**62**:500–6. https://doi.org/10.1136/oem.2002.001115
29. He D, Liu Q, Wu Y. *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**:879–92. https://doi.org/10.1038/s42256-022-00541-0
30. Guo Q, Hernandez-Hernandez S, Ballester PJ. Scaffold splits overestimate virtual screening performance. In: *Proceedings of the International Conference on Artificial Neural Networks*, pp. 58–72, 2024.
31. Kurilov R, Haibe-Kains B, Brors B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Sci Rep* 2020;**10**:1–11.
32. Basu A, Bodycombe NE, Cheah JH. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;**154**:1151–61. https://doi.org/10.1016/j.cell.2013.08.003
33. Barretina J, Caponigro G, Stransky N. *et al.* The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7. https://doi.org/10.1038/nature11003
34. Chemoinformatics, K.
35. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6. https://doi.org/10.1021/ci00057a005
36. Kim S, Chen J, Cheng T. *et al.* PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res* 2019;**47**:D1102–9. https://doi.org/10.1093/nar/gky1033
37. Narykov O, Zhu Y, Brettin T. *et al.* Entropy-based regularization on deep learning models for anti-cancer drug response prediction. In: *Proceedings of the Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, pp. 121–2, 2023.
38. Rezvani S, Wang X. A broad review on class imbalance learning techniques. *Appl Soft Comput* 2023;**143**:110415. https://doi.org/10.1016/j.asoc.2023.110415
39. Haixiang G, Yijing L, Shang J. *et al.* Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications* 2017;**73**:220–39. https://doi.org/10.1016/j.eswa.2016.12.035
40. Chawla NV, Bowyer KW, Hall LO. *et al.* SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321–57. https://doi.org/10.1613/jair.953
41. Torgo L, Ribeiro RP, Pfahringer B. *et al.* Smote for regression. In: *Proceedings of the Portuguese conference on artificial intelligence*, pp. 378–89, 2013.
42. Cao P, Zhao D, Zaïane OR. A PSO-based cost-sensitive neural network for imbalanced data classification. In: *Proceedings of the 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*, pp. 452–63, 2013.
43. Abarzadeh M, Kivi HF, Kojabadi HM. A modified SVM switching pattern for Z-source inverter. In: *Proceedings of the 2016 7th Power Electronics and Drive Systems Technologies Conference (PEDSTC)*, pp. 486–91, 2016.
44. Raghuwanshi BS, Shukla S. Class-specific cost-sensitive boosting weighted ELM for class imbalance learning. *Memetic Computing* 2019;**11**:263–83. https://doi.org/10.1007/s12293-018-0267-4
45. Joshi MV, Kumar V, Agarwal RC. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In: *Proceedings of the Proceedings 2001 IEEE international conference on data mining*, pp. 257–64, 2001.
46. Song J, Lu X, Wu X. An improved adaboost algorithm for unbalanced classification data. In: *Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 109–13, 2009.
47. Hern´ Ndez-Orallo, J. Probabilistic reframing for cost-sensitive regression. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2014;**8**:1–55. https://doi.org/10.1145/2641758
48. Teich J. Pareto-front exploration with uncertain objectives. In: *Proceedings of the International Conference on Evolutionary Multi-Criterion Optimization*, pp. 314–28. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.
49. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv* 2014;1412.6980.
50. Potvin PJ, Schutz RW. Statistical power for the two-factor repeated measures ANOVA. *Behav Res Methods Instrum Comput* 2000;**32**:347–56. https://doi.org/10.3758/BF03207805
51. Abdi H. The greenhouse-geisser correction. *Encyclopedia of research design* 2010;**1**:544–8. https://doi.org/10.3758/BF03207805
52. Huang K, Xiao C, Glass L. *et al.* Explainable substructure partition fingerprint for protein, drug, and more. In: *Proceedings of the NeurIPS Learning Meaningful Representation of Life Workshop*. San Diego, CA: Preprint, 2019.
53. Mayr A, Klambauer G, Unterthiner T. *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* 2018;**9**:5441–51. https://doi.org/10.1039/C8SC00148K

Briefings in Bioinformatics, 2025, **26(2)**, bbaf134
https://doi.org/10.1093/bib/bbaf134
Problem Solving Protocol