Deep Learning for Healthcare Final Report (Spring 2022):

Reproduction Study for RefDNN Cancer Drug Resistance Prediction Model

Delong Meng
delongmeng@hotmail.com

Presentation link: https://youtu.be/jDHLPnswYU8
Code link: https://github.com/delongmeng/DL4H-RefDNN

1 Introduction

The goal of precision medicine is to develop personalized treatment plans for patients according to their specific characteristics including their genetic background, making it critical to precisely predict the response of disease-related cells to certain drugs. Recent years, machine learning and deep learning strategies have been widely applied to the drug response problem in various human diseases including cancer (Baptista et al., 2021). For instance, the RefDNN project (Choi et al., 2020) aims to predict cancer drug resistance and proposed a "reference drug" based neural network architecture to achieve this goal. Specifically, the task is to build a binary classifier to predict whether a cancer cell line is sensitive or resistant to a certain drug, based on both the gene expression profile of cell lines and molecular fingerprints of drugs. The key idea of this work is to use a set of drugs (so-called "reference drugs" by the authors) to learn the representations of cell lines and drugs. Here I attempt to reproduce the key components of this work to elucidate how deep learning models can be utilized to help make such predictions.

2 Scope of reproducibility

This paper proposed a novel architecture based on a collection of reference drugs and learn representation of cell lines and drugs based on the reference drugs, and it can overcome the cold-start problem and help identify drug-resistance related gene markers.

2.1 Addressed claims from the original paper

- Claim 1: The RefRNN model has better prediction performance including accuracy, AUCROC, precision, recall, f1 score and AUCPR, as compared to baseline models such as Random Forest and Elastic Net.
- Claim 2: The RefRNN model has better performance (such as AUCROC and AUCPR) making prediction on unseen cancer types or drugs, compared with baseline models such as Random Forest and Elastic Net.
- Claim 3: Weights of the cell line representation step of the RefDNN model can be used to identify biomarkers to certain drugs.

3 Methodology

3.1 Model description

The model takes cell lines and drugs as inputs. Each cell line is originally represented by the expression levels of a range of genes, and then an ElasticNet layer is used to get the representation of probability of resistance to each drug of the reference drugs. Note that the weights of this step are later used to identify biomarkers, and partial loss is also computed here by comparing with the true drug resistance response. On the other hand, drugs are originally represented by their fingerprints, and then converted to their structural similarity compared with the reference drugs. Next, the cell line representation and drug representation will be multiplied element-wisely and serve as the input to a deep neural network (DNN), with two fully connected hidden layers with equal number of units, before making a final prediction of the resistance. Partial loss of the DNN part will also be computed. The weights of the ElasticNet layer and the DNN classifiers are updated by the gradient of the total loss consisting of both the ElasticNet and the DNN losses. For more details of the model architecture and learning objective described above, please refer to the original paper (Choi et al., 2020).

The total number of parameters in the model depends on the exact hyperparameters of the model,

Table 1: Overview of the datasets used in this study

Data Type	GDSC dataset	CCLE dataset
Gene expression of cell lines (predictor)	983 cell lines x 17780 genes	491 cell lines x 18926 genes
Fingerprint of drugs (predictor)	222 drugs x 3072 dimensions	12 drugs x 3072 dimensions
Drug response (response)	190036 pairs (120606 resistance; 69430 sensitive)	5724 pairs (3402 resistance; 2322 sensitive)

Table 2: Computational resource usage and running time for this study

Section	Model	Dataset	Device (memory)	Running time
Initial Bayesian optimization and performance evaluation	Elastic Net	CCLE	CPU (96 GB)	25 min
	Elastic Net	GDSC	CPU (96 GB)	9 h
	Random Forest	CCLE	CPU (96 GB)	25 min
	Random Forest	GDSC	CPU (96 GB)	10 h
	RefDNN	CCLE	GPU (16 GB)	3h
	RefDNN	GDSC	GPU (16 GB)	21 h
Leave-One-Cancer-type-Out Cross Validation (LOCOCV)	Elastic Net	CCLE	CPU (96 GB)	2 min
	Elastic Net	GDSC	CPU (96 GB)	1 h
	Random Forest	CCLE	CPU (96 GB)	1 min
	Random Forest	GDSC	CPU (96 GB)	1.5 h
	RefDNN	CCLE	GPU (16 GB)	20 min
	RefDNN	GDSC	GPU (16 GB)	2 h
Leave-One-Drug-Out Cross Validation (LODOCV)	Elastic Net	CCLE	CPU (96 GB)	1 min
	Elastic Net	GDSC	CPU (96 GB)	8 h
	Random Forest	CCLE	CPU (96 GB)	1 min
	Random Forest	GDSC	CPU (96 GB)	10 h
	RefDNN	CCLE	GPU (16 GB)	10 min
	RefDNN	GDSC	GPU (16 GB)	14 h
Biomarker identification	RefDNN	CCLE	CPU (96 GB)	<1min
	RefDNN	GDSC	CPU (96 GB)	<1min

for example the number of hidden units in the neural network. When this hyperparameter is set to be 128, the total number of parameters is 737,313.

3.2 Data description

There are two pharmacogenomics datasets used in this study: the Cancer Cell Line Encyclopaedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) datasets. The data is provided in this paper's Github repo (`https://github.com/mathcom/RefDNN`). Each dataset has three components as shown in Table 1.

3.3 Hyperparameters

During the initial performance evaluation stage, this paper uses the Bayesian optimization and nested cross-validation approach to search for optimal values for hyperparameters. For the RefDNN model, the hyperparameters include 1) the number of units in hidden layers, 2) learning rate for the FTRL optimizer, 3) learning rate for the Adam optimizer, 4) L1 and 5) L2 regularization strength for the FTRL optimizer. The hyperparameters of the baseline models are the number of estimators and max depth for the Random Forest model, and regularization strength, l1 ratio, and learning rate for the Elastic Net model. The best set of parameters of each model for each dataset will be used for the leave-one-group-out cross validation stage. The detailed searching space, hyperparameters used in the initial performance evaluation and leave-one-group-out cross validation stage can be found in my Github repo.

For other hyperparameters, default values from the original author's repo were used: outer fold = 5, inner fold = 3, number of Bayesian search = 20, number of training steps = 5000, batch size = 64.

3.4 Implementation

The existing code for this paper (`https://github.com/mathcom/RefDNN`) serves as a reference and template. I needed to re-organize the existing scripts and write up more scripts to perform the model training and evaluation tasks. In addition, I wrote up all of the analysis code for output data wrangling, analysis and visualization. All of these scripts can be found in my Github repo (`https://github.com/delongmeng/DL4H-RefDNN`).

3.5 Computational requirements

My previous expectation for computational requirements was: "The sample size seems to be not too large. The architecture is also not too complicated. I don't foresee computational challenges here and think it should be feasible." The original authors' repo suggested that "RefDNN requires system memory larger than 24GB. If you want to use tensorflow-gpu, GPU memory of more than 4GB is required."

For this reproduction study, a GPU machine equipped with a NVIDIA T4 GPU (16 GB memory) and a CPU machine with 96 GB memory were used to train the RefDNN model and baseline models (Elastic Net and Random Forest), respectively. It takes around 36 seconds or 4 minutes to train one round (including 5000 training steps) of the RefDNN model for CCLE dataset and GDSC dataset, respectively. Thus, it takes over 20 hours to perform the initial Bayesian optimization and performance evaluation for the RefDNN model on the GDSC dataset (4 minutes x 3 inner fold x 20 hyperparameter combinations x 5 outer fold). Detailed running time for all of the steps can be found in Table 2. The RefDNN model was run using Tensorflow 1.12.

4 Results

The prediction performance of the RefDNN model was first evaluated and compared with baseline models in both GDSC and CCLE datasets. It was further evaluated for untrained cancer types or drugs. Later, the weights of different genes for certain drugs were extracted to investigate their potential usage as biomarkers for drug resistance.

4.1 Prediction performance of RefDNN and baseline models

The CCLE and GDSC datasets were used to train the RefDNN model and some evaluation metrics, including accuracy, AUCROC, precision, recall, f1 score and AUCPR, were collected during the training process, and compared with other baseline models such as Random Forest and Elastic Net. Note that *RefDNN-test* is an additional experiment that I performed and will be discussed later. Specifically, a 5-fold nested cross-validation strategy was used to train the model, and within each round of the outer cross-validation, Bayesian optimization was used to search for the best set of the hyperparameters (see 3.3) out of 20 different combinations of hyperparameters. For each combination of hyperparameters, an inner loop of 3-fold inner cross-validation was done to further split the training data from the outer loop cross-validation into a training set and a validation set, to evaluate the metrics for given hyperparameters. For each round of the outer cross-validation, the best combination of hyperparameters will be used to obtain the final metrics. All metrics from the 5 rounds of cross-validation were summarized in Figure 1. All the 6 metrics for the RefDNN model are around 0.9 to 0.95 in the CCLE dataset and around 0.8 to 0.9 in the GDSC dataset, which is consistent with the values reported in the original paper (Choi et al., 2020).

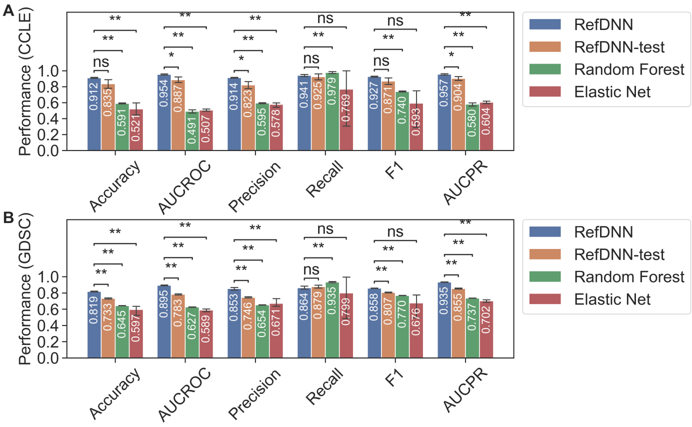

Figure 1: Drug resistance prediction performance of the RefDNN model and comparison with baseline models. The RefRNN, Random Forest and Elastic Net models were trained on the CCLE (A) or GDSC dataset (B), and the accuracy, AUCROC, precision, recall, f1 score, and AUCPR values are shown in bar plots (Mean ± Std, with the Mean values displayed in the bars). RefDNN-test is an additional test on the RefDNN model. Statistical significance was evaluated between RefDNN and other models using the Welch's t-test followed by the Benjamini-Hochberg correction. *: $p < 0.05$; **: $p < 0.01$; ns: not significant.

For most of the performance metrics, the RefDNN model significantly outperformed the Random Forest and Elastic Net baseline models. My results support the *claim 1* and are overall consistent with the original paper's conclusion, although there are some slight difference in the baseline model Random Forest's performance, and larger variation of the Elastic Net model's recall scores, neither of which affects the conclusion.

4.2 Performance on unseen cancer types or drugs

Practically, one of the ultimate goals of drug sensitivity prediction is to make such prediction on new patients or new compound. Thus, it is interesting to evaluate the model's performance on untrained cancer types or drugs here. The leave-one-group-

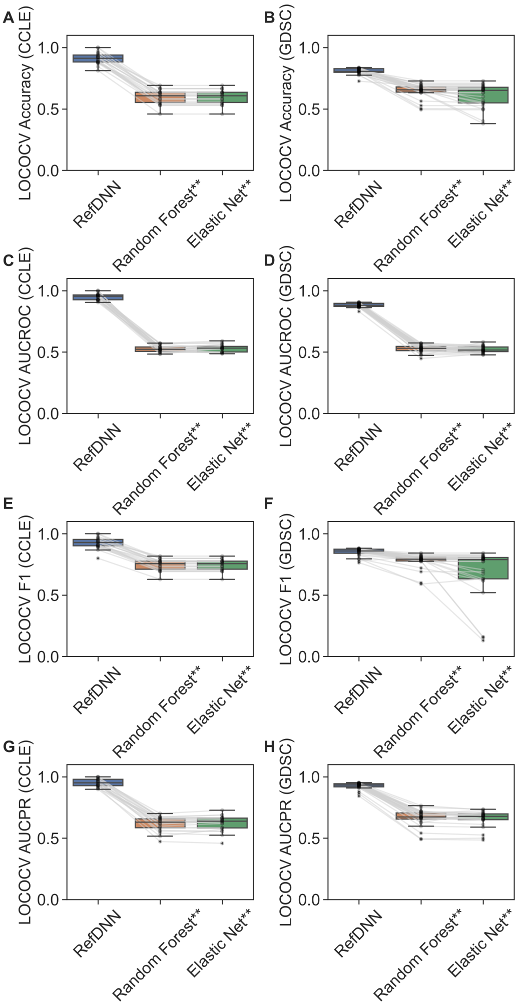

Figure 2: Prediction performance of RefDNN for untrained cancer types. Accuracy (A-B), AUCROC (C-D), f1 score (E-F), and AUCPR (G-H) of the RefRNN, Random Forest and Elastic Net models were evaluated on the CCLE (A, C, E, G) or GDSC (B, D, F, H) dataset using the leave-one-cancer-type-out cross validation (LOCOCV) strategy. Box plots summarize the values of different cancer types and lines link the paired cancer types across the models. Statistical significance was evaluated between RefDNN and other models using the Wilcoxon signed-rank test followed by Bonferroni correction. *: $p < 0.05$; **: $p < 0.01$.

out cross validation strategy was used for cancer types and drugs, respectively.

4.2.1 Leave-One-Cancer-type-Out Cross Validation (LOCOCV)

All of the cell lines within the CCLE or GDSC datasets were categorized into around 20 - 30 cancer types. Then LOCOCV was performed to evaluate the prediction performance for each cancer type by using the model trained on all the data except for the corresponding cancer type. As shown in Figure 2, in both CCLE and GDSC datasets, the RefDNN model has significantly better accuracy, AUCROC, f1 score and AUCPR than the baseline models. These performance metrics have relatively small variations and are comparable to the values obtained in Figure 1, indicating that the model trained in some cancer types can very well generalize to unseen cancer types.

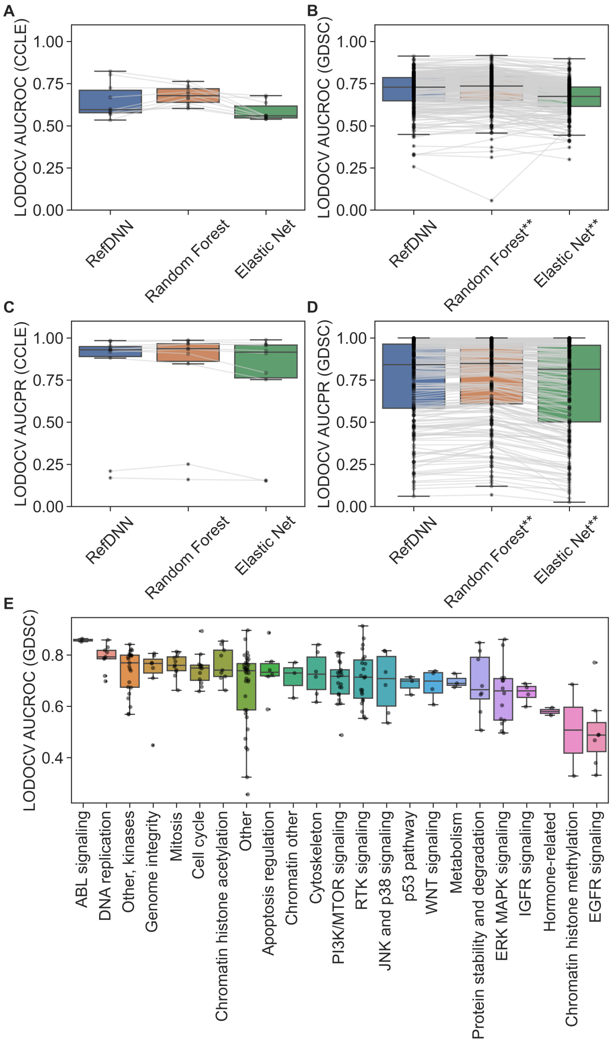

Figure 3: Prediction performance of RefDNN for untrained drugs. (A-D): AUCROC (A-B) and AUCPR (C-D) of the RefRNN, Random Forest and Elastic Net models were evaluated on the CCLE (A, C) or GDSC (B, D) dataset using the leave-one-drug-out cross validation (LODOCV) strategy. Box plots summarize the values of different drugs and lines link the paired drugs across the models. Statistical significance was evaluated between RefDNN and other models using the Wilcoxon signed-rank test followed by Bonferroni correction. *: $p < 0.05$; **: $p < 0.01$. (E): the LODOCV AUCROC values of the RefDNN model on the GDSC dataset for each drug were visualized in box plots by drug categories and ranked according to the median value.

4.2.2 Leave-One-Drug-Out Cross Validation (LODOCV)

LODOCV was performed similarly to evaluate the model performance on unseen drugs (Figure 3).

In the CCLE dataset there are only 12 drugs and there are not significant difference of the prediction AUCROC and AUCPR among the 3 models (Figure 3A and C). On the other hand, in the GDSC dataset (Figure 3B and D) which has 222 drugs, the values of the performance metrics vary a lot across the drugs, although the difference between the RefDNN model and the baseline models was statistically significant. In addition, different models seem to have consistent performance for the same drug, where the drugs that have better performance in one model tend to also have good performance in another model. This indicates that it is more of the drug's instinct nature rather than the model that dominants its predictability. I thus investigated the prediction performance (specifically, the AUCROC score here) of the drugs by their categories (Figure 3E). Interestingly, drugs of some categories, such as the ones that target ABL signaling, DNA replication, Cell cycle are on the higher performance end, while the resistance to the drugs of hormone-related, chromatin histone methylation, and EGFR signaling appear harder to predict.

Overall, these results support the *claim 2* and are consistent with the original paper's conclusion, although there are some slight difference in the baseline model Random Forest's performance again, which does not affect the conclusion. In addition, I provide more data here. The original paper doesn't show CCLE data and some metrics such as accuracy and f1 score, and it also didn't visualize the performance by drug category as I show in Figure 3E, which provides interesting insights.

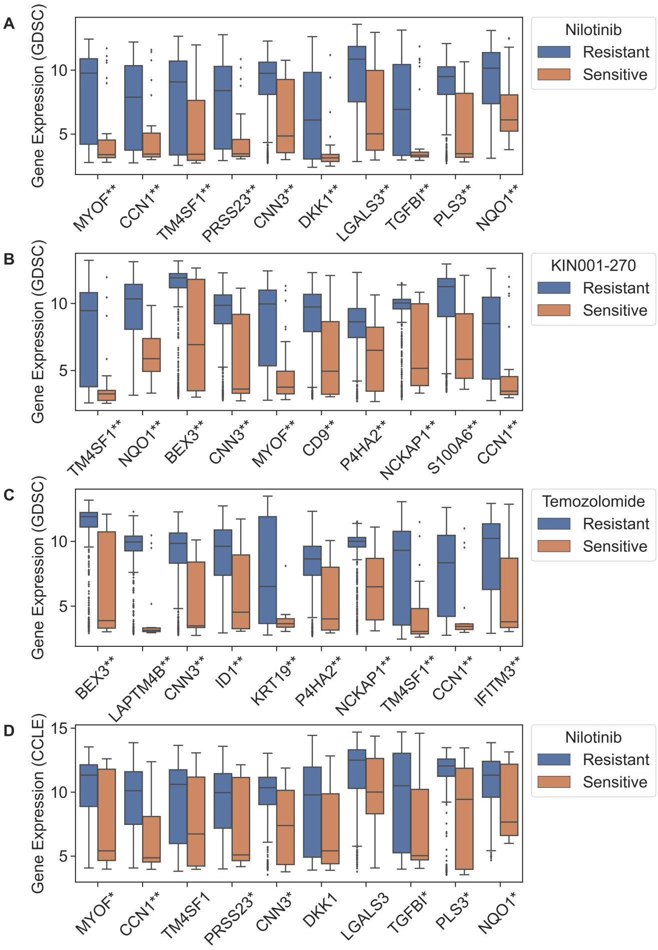

Figure 4: Identification and validation of drug resistance related biomarkers from the RefDNN model. (A-C): The expression level of the top 10 genes associated with the resistance to Nilotinib (A), KIN001-270 (B) or Temozolomide (C) from the RefDNN model trained in the GDSC dataset were summarized in box plots for the resistant verses sensitive cell lines. (D): the expression levels of the Nilotinib biomarker genes in (A) were plotted for the Nilotinib-resistant verses sensitive cell lines in the CCLE dataset. Statistical significance was evaluated between resistant and sensitive cell lines using the Mann-Whitney U test followed by the Benjamini-Hochberg correction. *: $p < 0.05$; **: $p < 0.01$

4.3 Drug resistance related biomarkers

The initial layer of data transformation of the gene expression of each cell line is an Elastic Net, making it plausible to extract the weights of the genes for a certain drug and potentially use that as biomarkers for the resistance of the drug. As mentioned above, some categories of drugs seem to have high predictability, so we chose to investigate the potential biomarkers of those drugs. As shown in Figure 4A-C, the expression levels of the top 10 genes with the highest absolute weights for the drug Nilotinib (an ABL signaling inhibitor), KIN001-270 (a cell cycle inhibitor), and Temozolomide (a DNA replication inhibitor) by the RefDNN model generated from the GDSC dataset were compared between the cell lines that are resistant or sensitive to that drug. All of these genes indeed have significantly different expression according to the drug sensitivity, supporting their potential usage as biomarkers. I also notice that some common genes show up across these drugs, such as CCN1, TM4SF1, and CNN3, indicating they might have some critical function for cancer cell survival and drug sensitivity. For example, the CCN1 gene and Cysteine-rich protein 61 that it encodes, were recently found to regulate the chemosensitivity of leukemia (Song et al., 2019).

To validate these resistance-related biomarkers, the expression levels of the 10 genes associated with Nilotinib were further analyzed in the CCLE dataset (Figure 4D). Consistently, most of these genes were also differentially expressed between the Nilotinib-resistant and -sensitive cell lines, although some of them did not reach statistical significance.

Overall, these results support the *claim 3* and are consistent with the original paper's conclusion, although the detailed genes are not exactly the same because there are certainly some randomness here, which does not affect the conclusion. Some of the genes I found actually overlap with the original paper. I show more drugs as example here, and instead of the IC50 values used in the original paper for the validation part, I still divide the cell lines to resistant or sensitive, because I found that the IC50 values in the dataset have some issues and the authors might have somehow further processed them, making it hard to reproduce exactly the same result.

4.4 Additional Experiment

As mentioned above, the RefDNN model uses two optimizers and adds up the losses from both an initial cell line representation layer and the final DNN output and use the total loss as the learning objective. I performed an additional ablation experiment to further explore the RefDNN model, where I removed the loss calculation and the optimizer of the initial cell line representation layer and investigated how it affects the performance. Now the *RefDNN-test* model merely uses the DNN output to calculate the loss and optimize the parameters. As shown in Figure 1, it significantly decreased the overall performance as compared with the full RefDNN model, by 5-10% for most of the metrics in both datasets, with only a few exceptions (such as recall scores).

5 Discussion

Overall, the original paper is quite reproducible. I could achieve comparable results for most of the experiments. The additional ablation experiment that I performed suggests that the complex design of using two optimizers and considering the loss of the initial representation step of the cell lines does contribute to the prediction performance. In addition, it is also very likely that this design increases the interpretability of the model, such as the biomarker identification, which is an important aspect for the healthcare and biomedical field. Of course, more thorough investigation is still needed to make more solid conclusion on this point.

Here, I would like to provide some more in-depth insights about the prediction behavior in the leave-one-group-out cross validation since it was not thoroughly discussed in the original paper. The LOCOCV (Figure 2) had much better performance than the LODOCV (Figure 3). Cancer type does not seem to be critical for the prediction and the reason could be that although different cancer types may react to drugs differently but that is mainly driven by their molecular features, and the gene expression profile is already provided to the model, so in this case, the cancer type does not provide additional information to make a prediction. However, for drugs it is a totally different story, and we can see that the prediction performance varies a lot from drug to drug. The basic assumption of the RefDNN model is that drugs with similar structure (or similar fingerprint representation here) would have similar pharmacological characteristics. This is sometimes true but not always a valid statement. For example, kinases are critical regulator of many key pathways and also the targets of many small molecules. There could be different mechanisms of kinase inhibitors, such as ATP-competitive inhibitor which usually occupy the kinase pocket, and allosteric inhibitors which usually bind to a different site. They can have the same target but totally different structure.

5.1 What was easy

It was easy to set up the computational environment and re-use some of the author's scripts after the authors updated their repo upon my request. Also, the authors provided most of the data in their repo so that saved me some time to prepare the data. The figures in the original paper are very clearly explained (including which statistical method they used) so it was easy to follow.

5.2 What was difficult

At the beginning, the authors' repo was outdated so it was hard to set up the environment. I got many errors or warnings that I had to deal with. But that was solved after the authors updated their repo.

In addition, it took a lot of time and computational resources to run all of these experiments (see Table 2 for details).

The most time consuming part for me was to organize all of the output data of the different models/datasets/cross-validation methods and perform data analysis (including statistical analysis) and data visualization. Because the authors did not provide any of code for this, I had to spend a lot of time on this from scratch.

5.3 Recommendations for reproducibility

I think the most critical aspect of improving reproducibility is to provide exactly the same data and all of the code that the authors used to generate all of the figures and tables in the manuscript. In addition, well-organized documentation (in the original paper and in a repository) is the key. If the authors organize the data, code, and documentation with a goal for other people to easily reproduce the work, it will be very helpful. In reality, I found that most of the time, the authors only provide partial code they used.

References

Delora Baptista, Pedro G Ferreira, and Miguel Rocha. 2021. Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, 22(1):360–379.

Jonghwan Choi, Sanghyun Park, and Jaegyoon Ahn. 2020. Refdnn: a reference drug based neural network for more accurate prediction of anticancer drug resistance. *Scientific reports*, 10(1):1–11.

Yanfang Song, Qing Lin, Zhaolian Cai, Taisen Hao, Yaohan Zhang, and Xianjin Zhu. 2019. Cysteine-rich protein 61 regulates the chemosensitivity of chronic myeloid leukemia to imatinib mesylate through the nuclear factor kappa b/bcl-2 pathway. *Cancer Science*, 110(8):2421–2430.