STEM Fellowship Big Data Challenge 2020

Deep learning transcriptomic model for prediction of pan-cancer chemotherapeutic sensitivity

Eddie Guo[1], Mehul Gupta[2], Pouria Torabi[1], and Sunand Kannappan[2]

[1]University of Alberta
[2]University of Calgary

May 31, 2020

Abstract

[INSERT HERE]

Keywords
neural network, pan-cancer, clustering

1 Introduction

It is well established that tumour sensitivity to chemotherapy is highly heterogeneous between, but also within cancer types. Despite this, traditional treatment courses are largely standardized within specific cancer types. As such, a subset of cancer patients fails to respond to cancer type-specific chemotherapy while also carrying a large side effect burden from treatment. Given the significant impact of chemotherapy failure and associated toxicities, it is crucial to improve methods to select chemotherapies tailored for individual cancer patients.

Past precision oncology strategies have largely focused on utilizing tumour-specific characteristics, including genetic aberrations and expression profiles, to generate targeted therapies. Emerging precision oncology approaches have begun to utilize molecular profiling of tumours to better select existing chemotherapies predicted to have high potency. While most chemotherapies are non-targeted and have broad activity, there are molecular determinants of chemotherapeutic efficacy, including gene expression. It is well known that gene expression may mediate chemotherapeutic activity pathways. Cell line drug panels provide a novel data source to inform these studies, providing in-vitro cell line sensitivity with associated molecular profiling. These data sources allow for the identification of genetic predictors of drug sensitivity.

We postulate that a gene signature can predict chemotherapeutic response across cancer types. This study effectively clustered cell lines into chemotherapeutic response profiles, followed by the development of a neural net model that could stratify cell lines into these clusters utilizing mRNA expression molecular profiling.

2 Methods

Pan-cancer therapeutic response cohorts

To better understand the impact and predictive ability of transcriptomic dysregulation in chemotherapeutic response, a pan-cancer cohort of cell-line and associated therapeutic efficacy data were obtained from the Genomics of Drug Sensitivity in Cancer (GDSC). This data includes 1,110 cell lines from various different tumour types, and is thought to represent a relatively comprehensive pan-cancer dataset. In addition, the acquired dataset contained therapeutic efficacy information in the form of half-maximal inhibitory concentration (IC_{50}) values for 251 chemotherapies. These values correspond to the minimal concentration of therapeutic required to induce cell death in 50% of the cells cultured, with lower values being associated with improved drug efficacy. This data was used to generate a matrix with cell-line and accompanying therapeutic information. This dataset was filtered to exclude therapies with less than 80% of data for all cell-lines, followed by the exclusion of cell lines lacking response data for the drugs retained in the first step. This resulted in the inclusion of 548 cell-lines and 117 therapeutics for clustering based analysis.

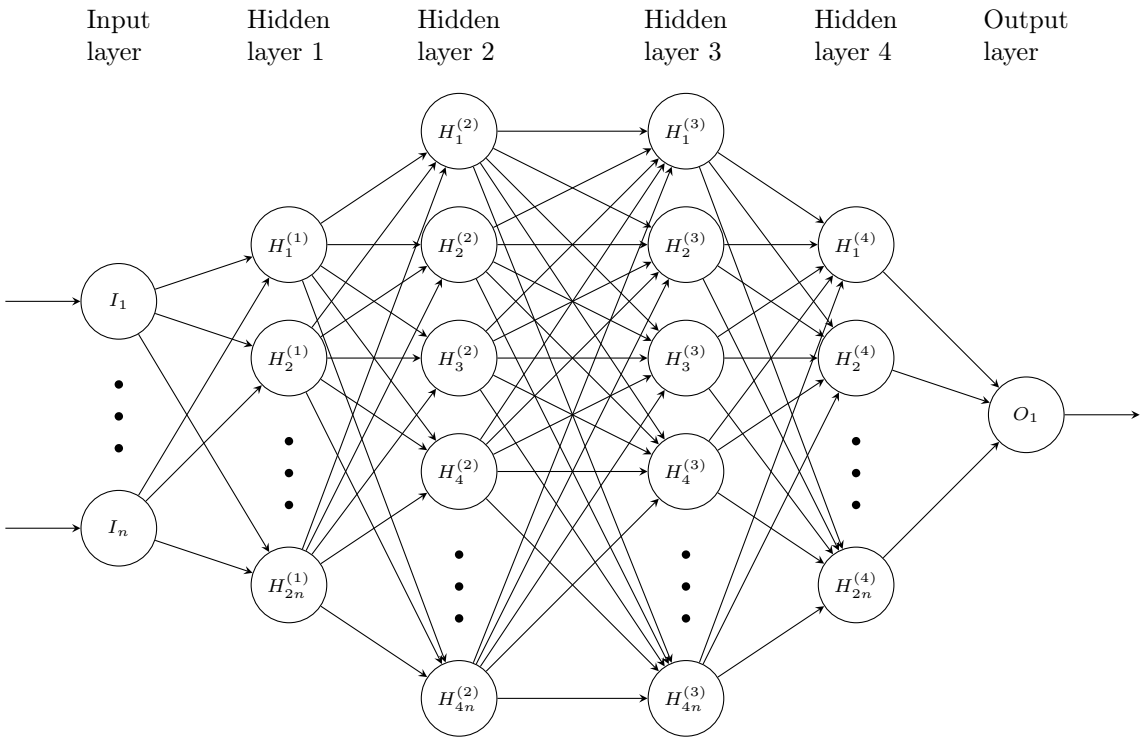


Figure 1: Neural network architecture representation with four hidden layers ($n = 300$). Inputs include the feature-selected genes from the RNA-seq dataset. Each hidden layer has a dropout rate of 0.3 and is subject to batch normalization.

Identification of pan-cancer therapeutic response cohorts

Cell line therapeutic response matrices obtained from the GDSC consortium were used to bifurcate candidate cell lines into defined response cohorts. We developed a Euclidean distance matrix for the retained cell lines based upon their pan-chemotherapy response. This matrix was then used to identify an optimal number of clusters capable of representing the therapeutic heterogeneity identified across the cancer cell lines. K-means clustering was then utilized to assign cell line candidates to appropriate therapeutic response cohorts. Generalized differences in chemotherapeutic efficacy between cohorts were visualized using a heatmap generated by the Pheatmap package in R. Separation between clusters was also visualized using principal component analysis with the factoextra package in R. Following the identification of defined clusters, differences in therapeutic efficacy between the identified cohorts were evaluated. Mann-Whitney U tests were utilized to compare the half-maximal inhibitory concentration (IC50) values between the groups. False discovery rate (FDR) correction was utilized to correct for multiple comparisons.

Feature Selection

Feature reduction was the first of multiple steps taken to avoid overfitting. The number of genes within the RNA-seq dataset were reduced using the BorutaPy package in Python 3. The Boruta algorithm is a feature selection wrapper algorithm based on Random Forest classification, and it iteratively removes features that are statistically less significant than a shuffled version of the same feature. The algorithm identified 300 relevant genes from the original set of 16,382 genes at $\alpha = 0.05$ with a maximum tree depth of 5. Given the feature variance in RNA expression, prior to fitting the model, expression values were scaled by z-score.

Classification using an optimized neural network

The neural net was constructed using the Tensorflow Keras sequential deep learning API in Python 3. The model underwent multiple instances of optimization, starting with the manipulation of the overall hidden layer architecture. The classifier's predictive accuracy and misclassification rate were monitored to determine the optimal number of dense hidden layers (Figure - confusion matrix) in addition to iterative manipulation of the number of neurons in each hidden

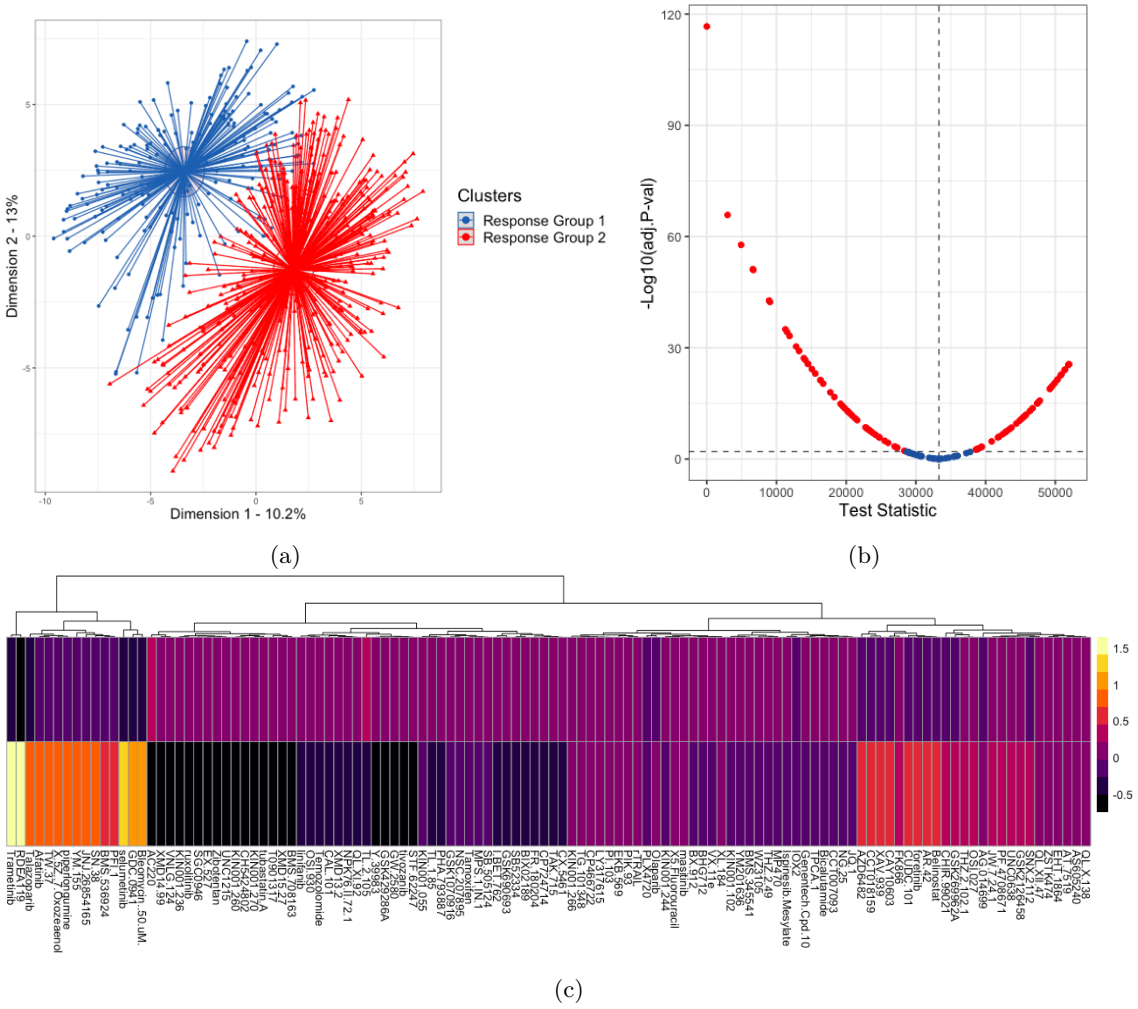
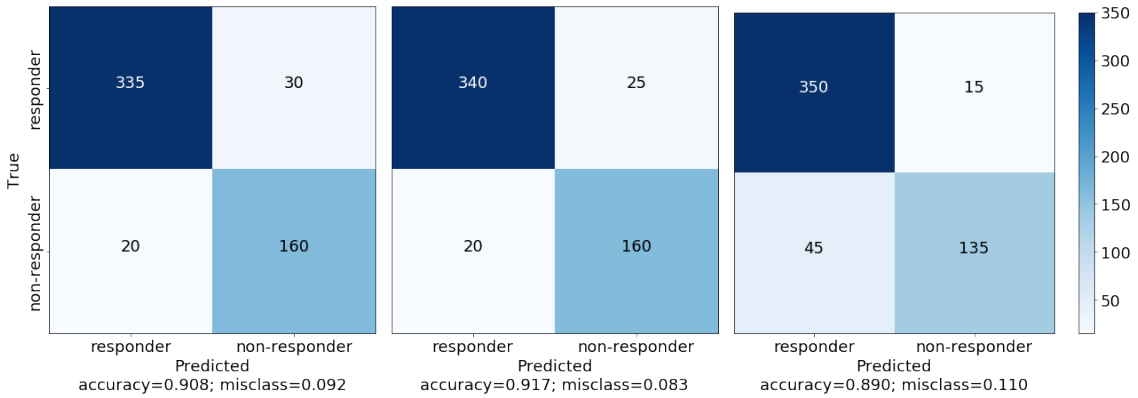


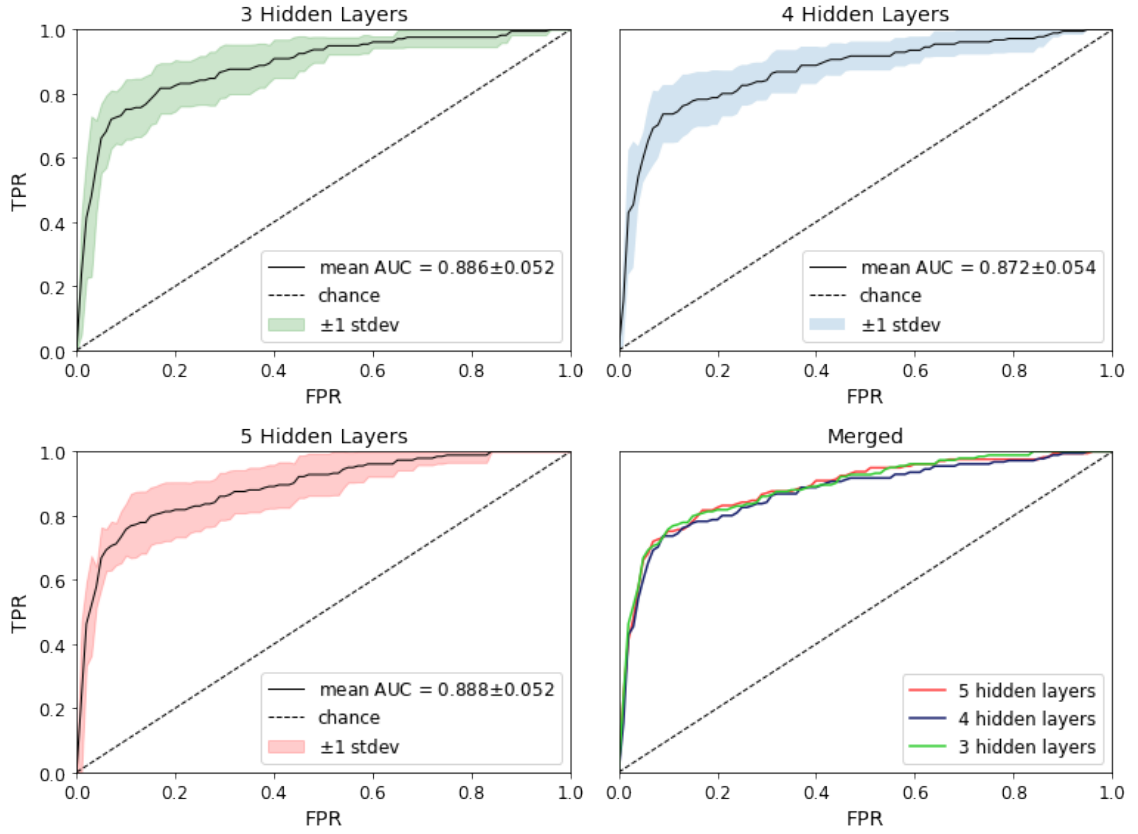
Figure 2: (2a) Principal component analysis of cell line therapeutic efficacy. The two identified therapeutic response clusters are indicated in blue and red respectively. (2b) Volcano plot identifying chemotherapeutics with significantly different IC50 values between therapeutic response clusters. Drugs identified in red meet the criteria for significance (FDR adjusted $p < 0.05$). (2c) Heatmap of therapeutic IC_{50} for the two identified therapeutic response clusters. Columns represent individual chemotherapies, and are clustered according to Euclidean distance. Colours range from yellow to purple, with a shift toward the latter indicating increased efficacy of the corresponding chemotherapeutic.

layer. The rectified linear unit (ReLu) was chosen as the neuronal activation function for all the layers except the output layer which used a sigmoid activation as a means of classifying instances into binary classes. All optimization procedures were conducted on 80% of the data, while the other 20% was saved for model validation.

The model was rigorously monitored for and protected against overfitting. The dataset was randomly segregated into 80% training and 20% validation blocks. Additionally, dropout layers with a 0.3 dropout rate and batch normalization layers were employed to improve the generalizability of the model. Model selection was performed by tuning the model's hyperparameters via a grid search and 5-fold cross-validation (Table 1). We performed a grid search with 3-fold cross-validation on the training data (80% of the dataset; 432 training samples, 541 overall) to determine the parameters under which to minimize the binary cross-entropy loss function. GridSearchCV from the scikit-learn library was used as a means of iterating through multiple possibilities of epochs, batch size, optimizer, and kernel initializer to find the optimal model. To prevent class imbalance during training, we used the Synthetic Minority Oversampling Technique (SMOTE) from the imblearn package for Python 3. Each model's performance was evaluated by the area under the receiver operating characteristic (ROC) curve (AUC). Performance



(a)



(b)

Figure 3: (3a) From left to right: confusion matrices for the 3, 4, and 5 hidden layer neural network models evaluating the true positive, false positive, true negative, and false negative rate. The models classify patient RNA-seq datasets into chemotherapy response cohorts. (3b) ROC curves for 3, 4, and 5 hidden layers neural network models with confidence bands of ±1 standard deviation. The models classify patient RNA-seq datasets into chemotherapy response cohorts. Each model was subject to 5-fold cross-validation and the mean across all trials was plotted.

evaluation of the final model was performed with the testing set.

3 Results

Clustering of pan-cancer cell lines identifies two distinct therapeutic response cohorts

To identify defined cohorts of pan-cancer cell-lines with similar trends in therapeutic efficacy,

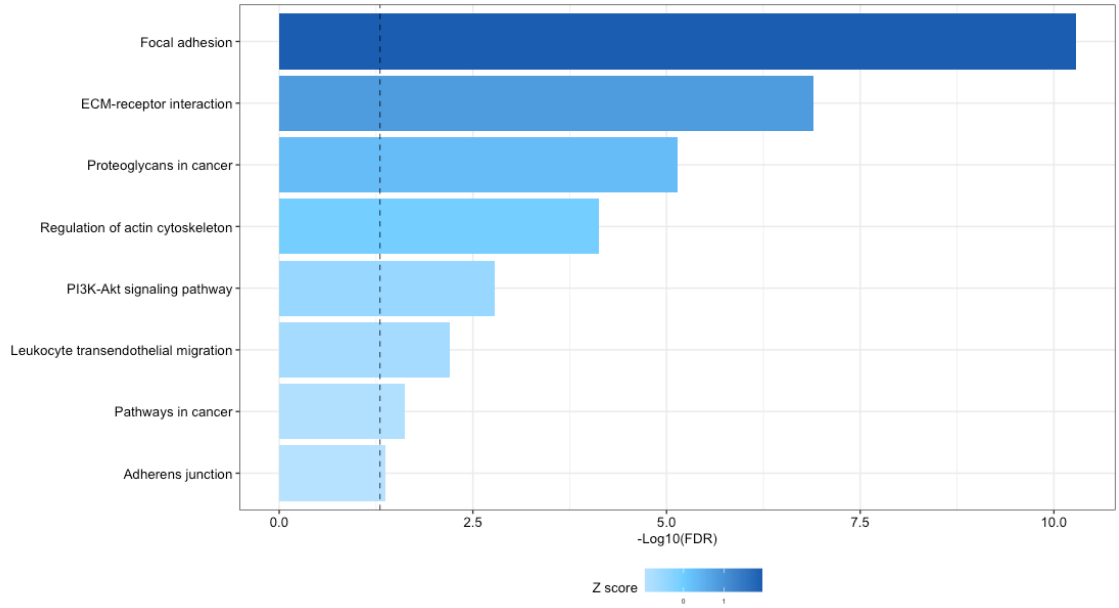


Figure 4: KEGG pathway functional enrichment for predictive genes included in the deep learning model conducted using g:Profiler.

we employed clustering of retained cell-lines. This process identified two distinct clusters of therapeutic efficacy (Fig. 2a), 362 cell lines identified in response group 1, and 186 cell lines identified in response group 2. Further analysis of these clusters demonstrate that a subset of therapeutics performs substantially differently between the different cohorts (Fig. 2c). To quantify differences in therapeutic response between clusters, IC50 values were compared between candidate cell lines (Fig. 2b). Of the 117 therapies included, 95 had significant differences in efficacy between the two cohorts identified. This suggests that these cohorts represent groups of cell lines with vastly different therapeutic responses. Therefore the ability to accurately stratify into these cohorts may be a valuable tool for stratification prior to chemotherapeutic treatment.

A neural network with four hiddeen layers accurately classifies patients into responder and non-respondeer cohorts

Unsupervised learning in the form of the K-Means Clustering of the cancer cell line transcriptomes indicated substantially different responses to chemotherapies. Using these distinct therapy response cohorts, we developed a deep learning binary classification to predict drug response based on transcriptome data. We initially analyzed five neural network architectures, each corresponding to 1-5 hidden layers (Fig. 1). Hyperparameter optimization via grid search returned similar results for each model: 50 epochs, batch size of 32, Adagrad as the optimizer, and a normal kernel initializer. Neural network architectures containing 3-5 hidden layers performed similarly with approximately 90% accuracy. The architectures with 1 and 2 hidden layers performed less optimally with 80% accuracy. We proceeded to validate the architectures with 3-5 hidden layers using 5-fold cross-validation. Of note, the model with 4 hidden layers had the lowest false positive rate (FPR = 3.67%) and a false negative rate (FNR = 4.59%). The model with five hidden layers had the highest FPR of the models evaluated (8.26%).

A receiver operating characteristic (ROC) curve was plotted for each of the neural network variants as an alternative evaluative method under uneven class sizes (Fig. 3b). The relative ratio between the model's false positive classification rate and its true positive rate was averaged between 5 K-Folds. The mean Area Under Curve (AUC) for the 5 trials was used to compare the three network architectures. However, the differences between the various architectures for AUC was not significatly different. Consequently, confusion matrices (Fig. 3a) and associated misclassification rates were used to pick the optimal model. The neural net with 4 hidden layers boasted the best performance overall with a 91.7% accuracy and an 8.3% misclassification rate.

4 Discussion

In this study we developed an accurate deep learning-based model for classifying cancer patients based on RNA expression data into therapeutic response cohorts on a relatively small cell line therapeutic response dataset. Using unsupervised clustering techniques, we segregated cancer patients into two defined therapeutic response groups with significantly different responses to a multitude to standard chemotherapies. [EXPAND HERE]

The relatively low accuracy of the neural networks with 1 and 2 hidden layers (82.6% and 70.8% respectively) suggests that the therapeutic response cohorts cannot be separated by a linear classifier. Furthermore, the high FNR of the network with 5 hidden layers as compared to the 3 and 4 hidden layer networks indicates overfitting. To this end, either a 3 or 4 hidden layer network is the ideal architecture for analyzing our data. Interestingly, there is no significant difference between AUC for these models (Fig. 3b). [EXPAND HERE]

A major limitation of our study was the availability of large datasets to train our model. Here we faced a $p \gg n$ problem as machine learning models expect that the number of features p is much larger than the number of observations n. To minimize this bias, we applied the Boruta algorithm to reduce our 16,382 genes by 541 cell lines dataset to a 301 by 541 matrix. The algorithm has been shown in various journals to be an effective feature selector method in high dimensional omics datasets [1]. To prevent overfitting of the reduced matrix, we applied batch normalization and dropout layers immediately preceding each hidden layer. Of note, a neural network architecture where the initial hidden layers diverge and the latter hidden layers converge provides the most accurate classifications of the RNA-seq data.

Future investigations will look to validate the efficacy of the model in prediction of chemotherapy response in various forms of cancer using the current transcriptomic signature. It is likely that the model accuracy will vary between therapy targets, as such, further studies can make use of our project pipeline to create a stratified model whereby the drug class and target are additional inputs.

Conclusions

Using transcriptomics data from the cancer cell lines, two chemotherapeutic response clusters were identified via unsupervised learning in the form of K-means Clustering. A feature selection algorithm was used to select a 300 gene signature which served as inputs to multiple neural networks. We determined that the network with 4 hidden layers was the most accurate model, producing a binary classifier to predict patient therapy response with 91.7% accuracy.

Acknowledgements

INSERT HERE

References

[1] Frauke Degenhardt, Stephan Seifert, and Silke Szymczak. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform*, 20(2):492–503, March 2019.

Supplementary Data

Figures

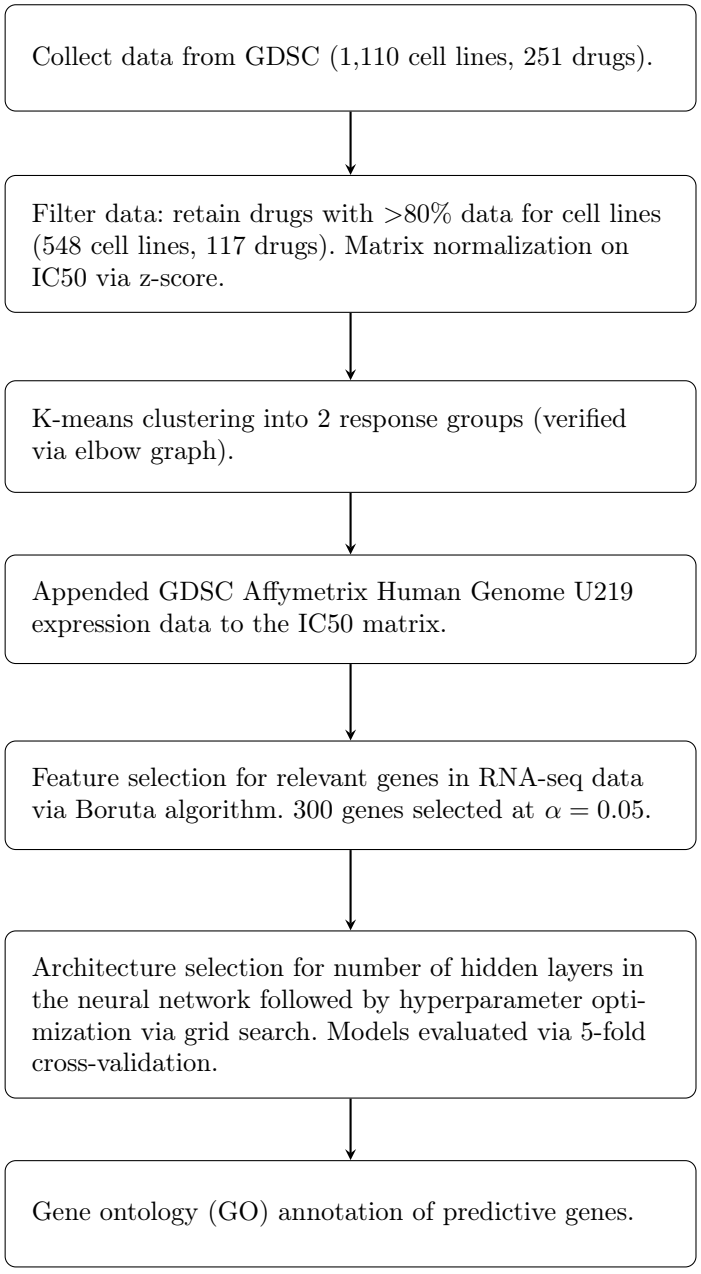


Figure 5: Summary of the data analysis pipeline.

Tables

Table 1: Grid search parameters to optimize all neural network architectures presented in this paper (3, 4, and 5 hidden layers). Each grid search underwent 3-fold cross-validation on the training data.

Epochs	Batches	Optimizer	Kernel initializer
25	15	Stochastic gradient descent	Normal
50	32	Adagrad	Uniform
75	64	Adam	Glorot uniform