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

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

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

June 25, 2020

Abstract

Emerging precision oncology studies have yet to generate a predictive biomarker that utilizes gene expression profiles to stratify tumours into similar pan-drug sensitivity profiles. This development would allow for the identification of candidate drugs for treatments that maximize therapeutic response and minimize cytotoxic burden. As such, this study utilized cell line sensitivity and molecular profiling data to generate a combinatorial gene expression predictive biomarker via feature selection and a deep learning model. A cohort of cell line gene expression data from Genomics of Drug Sensitivity in Cancer (GDSC) was clustered into two response groups. Cell line response groups showed a significant difference in pan-drug chemotherapeutic sensitivity. Due to the high dimensional nature of the microarray data, the Boruta feature selection algorithm was used to identify genes with the highest predictive value. The feature space was reduced to 300 genes; functional profiling was enriched primarily for the focal adhesion, ECM-receptor and proteoglycan interaction pathways. Using the selected genes, a deep learning neural network architecture was developed to predict response groups. It was determined that a 10 hidden layer neural network was optimal for the dataset; following hyperparameter tuning and 5-fold cross-validation, the model showed a predictive accuracy of 95.4% (AUC = 0.913 ± 0.049). This validates the postulate that cell lines with similar gene expression profiles present similar pan-drug chemotherapeutic sensitivity, and it suggests the potential utility of similar combinatorial biomarkers for the selection of potent candidate drugs.

1 Introduction

With the advent of high-throughput sequencing technology, precision oncology approaches have utilized molecular characteristics of tumours to inform clinical decision-making, including choice of chemotherapeutic regimen [1, 2]. The major focus of these approaches has been the development of targeted therapeutics, which are selective for specific genetic aberrations and expression profiles. Although these approaches may be successful for inducing tumour response, tumours are more likely to gain resistance to therapies with specific targets [3]. Moreover, not all tumours present with targetable features [3].

Emerging precision oncology approaches have begun to utilize high-throughput technology to potentiate the usage of conventional and less-targeted chemotherapy. Given that many of these less-targeted and consequently more cytotoxic chemotherapies have broad activity, the primary determinants of chemotherapeutic selection include cancer type and certain molecular markers [4]. Nevertheless, it is well established that tumour sensitivity to chemotherapy is heterogeneous both between and within cancer types, which results in a subset of patients that fail to respond to conventional chemotherapy regimens while being subjected to significant side effect burden [5]. Given that evidence suggests that gene expression can mediate drug response, recent advances have utilized individual and combinatorial gene expression biomarkers to develop predictors of tumour sensitivity to chemotherapeutic compounds [6].

While previous studies have developed predictive biomarkers for specific drugs, the utility of these biomarkers is limited to particular patients and clinical contexts [7]. That is, these studies are

limited in terms of clinical generalizability to different chemotherapeutic regimens. However, a pan-drug predictive biomarker may provide significant clinical utility in the selection of candidate therapies for particular patients. Such a biomarker could be developed if tumours with similar expression have similar drug responses across all chemotherapies, with few exceptions. The availability of pan-cancer cell line databases with *in vitro* drug sensitivity analyses along with accompanying gene expression profiling provides an ideal model for such analyses [8]. However, the majority of previous drug sensitivity predictive biomarkers built on cell line databases have utilized classical machine learning combinatorial techniques, which fail to capture the dimensionality of available transcriptomic data. Furthermore, advanced deep learning algorithmic approaches that are capable of handling such dimensionality often fail to allow interpretability, and consequently require transcriptomic data that is clinically infeasible [9]. Thus, deep learning approaches should minimize the number of transcriptomic features selected to maximize both the accuracy and interpretability of such predictive biomarkers.

Here we set out to generate a deep learning transcriptomic model for the prediction of pan-drug chemotherapeutic sensitivity across cell lines of all cancer types. If successful, this would demonstrate that gene expression influences chemotherapeutic response across most drugs and further motivate future studies into the development of clinically applicable predictors of candidate chemotherapeutics for tumours of a specific gene expression profile.

Following unsupervised clustering of cell lines into therapeutic response groups with similar pan-drug sensitivity, we show that conventional clinical criteria fail to stratify cell lines by therapeutic response. We utilize a biologically agnostic feature selection algorithm to iteratively select and identify a subset of 300 relevant genes predictive of chemotherapeutic sensitivity. We then develop a deep learning model from the selected genes utilizing neural networks, which showcases a strong predictive ability across pan-cancer cell lines.

2 Methods

We developed a deep learning model to accurately classify cancer cell lines into therapeutic response groups using data from the Genomics of Drug Sensitivity in Cancer (GDSC) consortium. Following data collection and curation, we utilized unsupervised clustering algorithms to define two groups based on chemotherapeutic response. Next, we employed a biologically agnostic feature selection algorithm, Boruta, to select statistically relevant genes for our neural network. These transcriptomic features were fed into a neural network that classifies patients into therapeutic response groups as defined by the clustering algorithms. The neural network

architecture was optimized using a grid search and evaluated using 5-fold cross-validation See Fig. 1 for an overview of the data analysis pipeline.

2.1 Determining 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 sensitivity data were obtained from the GDSC database. This database includes 1,110 cell lines from various 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 cultured cells, with lower values being associated with improved drug efficacy. The 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.

2.2 Identification of pan-cancer therapeutic response cohorts

Cell line therapeutic sensitivity matrices were used to evaluate whether conventional clinical criteria could separate patients into previously-defined chemotherapeutic response groups. These conventional clinical criteria included anatomic location and solid versus non-solid tumour status, as well as broadly applicable molecular markers – TP53 and KRAS mutation status [10, 11, 12, 13]. Cell lines were separated into subgroups based on these criteria and visualized to determine whether these criteria effectively clustered response groups.

Following evaluation of existing classifiers, we attempted to create defined cell line clusters using observed chemotherapeutic response of the pan-cancer cell line samples. We developed a Euclidean distance matrix for the retained cell lines based on their pan-chemotherapy response. This matrix was then used to identify the minimum number of clusters capable of representing the therapeutic heterogeneity identified across the cancer cell lines while maintaining significant inter-cluster distance. Following this, k-means clustering was 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 [14] in R. Separation between clusters was also visualized using principal component analysis with the factoextra package [15] in R.
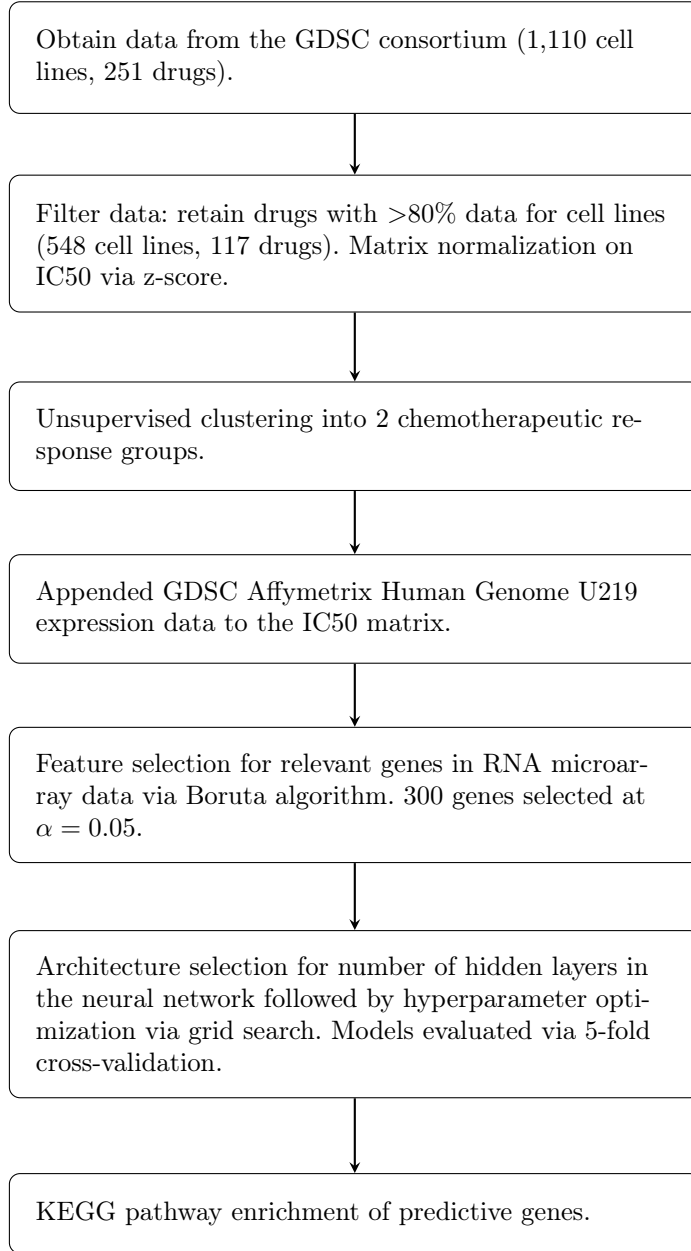
Figure 1: Summary of the data analysis pipeline.

Following the identification of defined clusters, differences in therapeutic efficacy between the identified cohorts were evaluated. Mann-Whitney U tests were utilized to compare IC_{50} values between the groups. False discovery rate (FDR) correction was utilized to correct for multiple comparisons.

2.3 Feature Selection with Boruta

In order to develop a transcriptomic model predictive of therapeutic response clusters, we retrieved expression data quantified by the GDSC consortium using the Affymetrix U219 microarray for each candidate cell line. Here, minimally processed CEL files were obtained from ArrayExpress (ascension number E-MTAB-3610) and processed using the affy package [16] in R. The resulting normalized expression matrix for candidate cell lines was then merged with the existing dataset. This addition resulted in the loss of 7 cell lines (2 from cluster A and 5 from cluster B), resulting in the inclusion of 541 cell lines in model generation. The microarray dataset quantified the expression of 16,382 genes; a model developed from this large feature space is likely to exhibit high multicollinearity and subsequently overfit, compromising the generalizability of the model. In addition, the large feature space limits the efficiency of various feature selection algorithms[1] as well as training and optimizing neural networks. We addressed these issues with the Boruta feature selection algorithm using the BorutaPy package [17] in Python 3. Boruta is a feature selection algorithm based on Random Forest classification which iteratively removes features that are statistically less

---

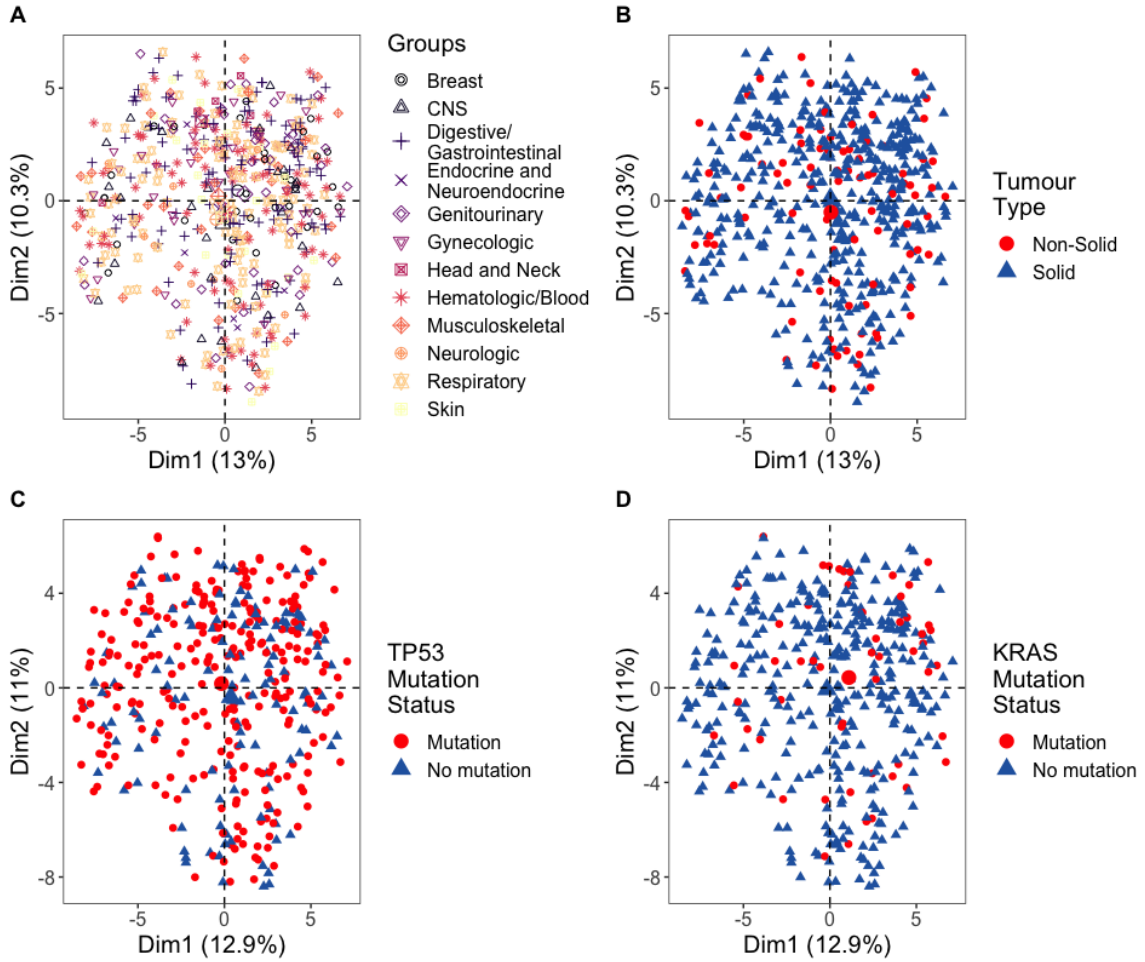[1]For example, univariate selection or recursive feature elimination.

Figure 2: Principal component analysis of pan-cancer cell line therapeutic efficacy, generated from IC_{50} values of all available chemotherapeutics. The horizontal axis shows the first principal component, the vertical axis the second component. Cell lines are visualized based on major cancer type classifications, including (**A**) body system of tumour and (**B**) solid vs. non-solid tumour status. Cell lines were also visualized on major molecular markers, including (**C**) TP53 mutation status, and (**D**) KRAS mutation status.

significant than a shuffled version of the same feature [18]

2.4 Classification of cell lines using an optimized neural network

The neural network was constructed using the Tensorflow Keras sequential deep learning API [19] in Python 3. The model underwent multiple instances of optimization, beginning with the depth of the network (1, 5, 10, and 15 hidden layers were evaluated). The rectified linear unit (ReLu) was chosen as the neuronal activation function for input and hidden layers. The output layer used a sigmoid activation for binary classification of inputs (i.e., chemotherapeutic response groups). The neural network was rigorously monitored for overfitting on the training dataset. To minimize overfitting, we applied L2 kernel regularization ($\lambda = 10^{-3}$), batch normalization, and dropout layers with a 0.3 dropout rate to each hidden layer.

The dataset was randomly segregated using the Pareto principle [20] where we reserved 80% of the data for training and the remaining 20% for validation. Model selection was performed by hyperparameter tuning using a grid search followed by 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 which minimize the binary cross-entropy loss function. GridSearchCV from the scikit-learn library [21] was used as a means of iterating through multiple possibilities of epochs, batch size, neurons per hidden layer, L2 regularization penalty, optimizer, and kernel initializer to find the optimal model (Table 1). To prevent class imbalance during training, we used the Synthetic Minority Oversampling Technique (SMOTE) from the imblearn package [22] for Python 3. Each model's performance on the testing dataset was evaluated by the Area Under Curve (AUC) value of the receiver operating characteristic (ROC) curve.
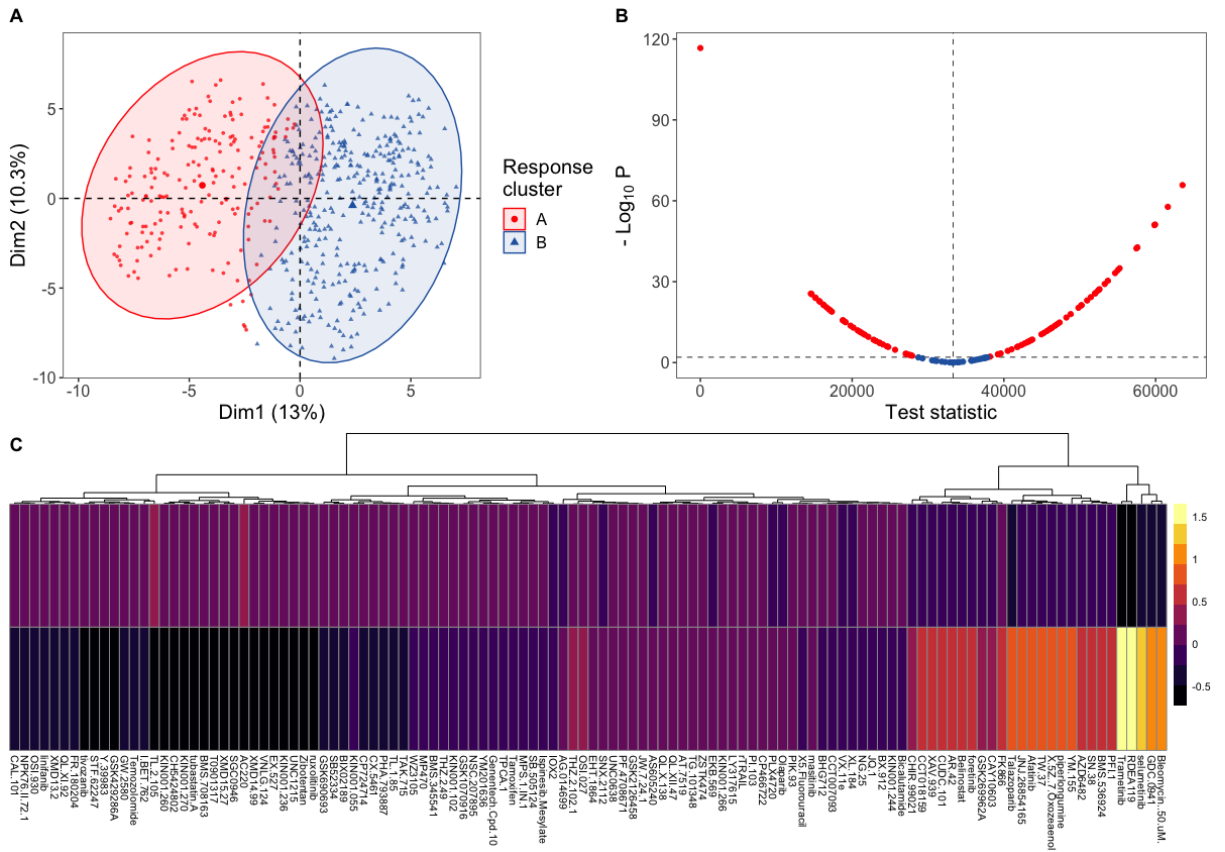
4

Figure 3: (**A**) Principal component analysis of pan-cancer cell line therapeutic efficacy, generated from IC_{50} values of all available chemotherapeutics. The horizontal axis shows the first principal component, the vertical axis the second component. The two identified therapeutic response clusters are indicated in red and blue respectively. (**B**) Volcano plot identifying chemotherapeutics with significantly different IC_{50} values between therapeutic response clusters. Drugs identified in red meet the criteria for significance (FDR adjusted $p < 0.05$). (**C**) 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 black, with a shift toward the latter indicating increased efficacy of the corresponding chemotherapeutic.

3 Results

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

From the GDSC consortium, we included 548 cell lines (49.4% of the original cell lines) and 117 (46.6% of the original drugs) chemotherapeutics for response group clustering. We assessed the ability of common molecular and clinical characteristics using principal component analysis (PCA) to stratify cell lines into groups with similar chemotherapeutic performance by subgrouping cell lines based upon these criteria.

To identify defined cohorts of pan-cancer cell lines with similar trends in therapeutic sensitivity, we employed unsupervised clustering of retained cell lines. PCA was used to reduce the dimensionality of the dataset, allowing for visualization of defined therapeutic response groups. This process identified two distinct clusters of therapeutic sensi-

tivity (Fig. 3), 362 cell lines identified in response group A, and 186 cell lines identified in response group B. That is, the cohorts perform substantially differently in a variety of therapeutics (Fig. 3). To quantify differences in therapeutic response between these clusters, IC_50 values were compared between candidate cell lines (Fig. 3). Of the 117 therapies included, 95 had significant differences in efficacy between the identified cohorts, suggesting that each represents groups of cell lines with vastly different therapeutic responses.

3.2 Boruta selects 300 genes from the 16,382 gene dataset

The Boruta algorithm was used to select genes that are estimated to have the greatest predictive contribution. The feature selection algorithm identified 300 relevant genes from the original set of 16,382 genes at $\alpha = 0.05$ with a maximum tree depth of 5. To better understand the transcriptomic heterogeneity underlying the therapeutic response cohorts, a KEGG pathway enrichment analysis was
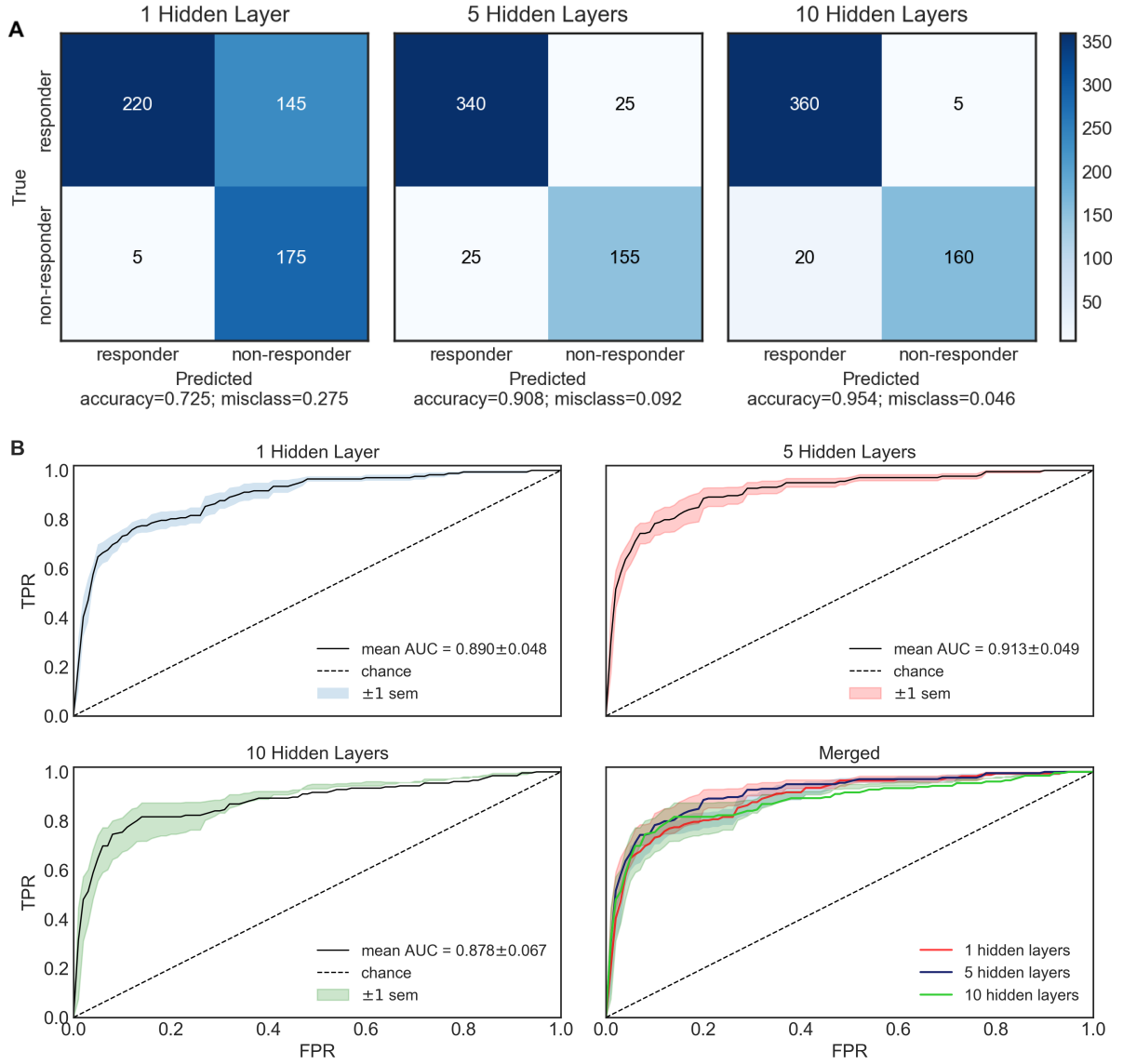
Figure 4: (**A**) Confusion matrices for the 1, 5, and 10 hidden layer neural network models (the 15 hidden layer neural network is not shown). The models classify cell line microarray datasets into chemotherapy response cohorts. (**B**) ROC curves for 1, 5, and 10 hidden layers neural network models with confidence bands of ± 1 standard error of the mean (sem). Each model was subject to 5-fold cross-validation, and the mean ROC curve across all trials was plotted.

performed on feature-selected genes used in the deep learning model. This analysis identified enrichment of gene sets associated with focal adhesion and PI3K signalling among others (Fig. 5).

3.3 A neural network with 10 hidden layers accurately classifies patients into responder and non-responder 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 classifier to predict drug response groups based on transcriptome data. We initially analyzed four neural network architectures, with 1, 5, 10, and 15

hidden layers (Fig. 6). Hyperparameter optimization via grid search returned similar results for each model: 25 epochs, batch size of 16, 100 neurons per hidden layer, L2 kernel regularization with $\lambda = 10^{-3}$, stochastic gradient descent as the optimizer, and a normal kernel initializer. Following this, we validated the various architectures with 5-fold cross-validation.

The neural network with 1 hidden layer performed poorly with 72.5% accuracy. However, the 1 hidden layer model had the highest precision 97.8% of the models (93.1% for 5 hidden layers, 94.7% for 10 hidden layers, 88.0% for 15 hidden layers). The poor accuracy for the 1 hidden layer model was accounted for with a low negative predictive value (54.7%; 86.1% for 5 hidden layers, 97.0% for
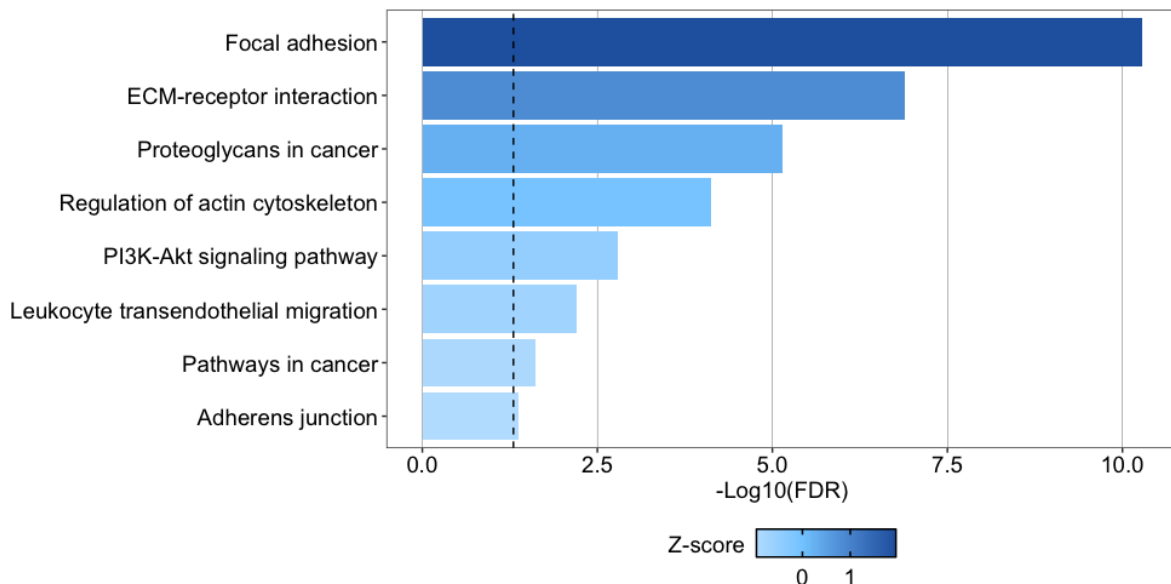
Figure 5: KEGG pathway functional enrichment for feature-selected genes included in the deep learning model. The vertical dotted line indicates the threshold for significance (adjusted $p < 0.05$).

10 hidden layers, 100% for 15 hidden layers). It was determined that the most accurate model was the neural network with 10 hidden layers (95.4% accuracy; 72.5% for 1 hidden layer, 90.8% for 5 hidden layers, 90.8% for 15 hidden layers).

An ROC curve was plotted for each of the neural network variants as an alternative evaluative method (Fig. 4). The relative ratio between the model's false positive rate and its true positive rate were averaged between 5 k-folds. The mean AUC for the 5 trials was used to compare the four network architectures. However, the differences between the various architectures for AUC was not significantly different. Consequently, confusion matrices (Fig. 4) and associated misclassification rates were used to pick the optimal model. The neural net with 10 hidden layers demonstrated the best performance overall with a 95.4% accuracy.

4 Discussion

It is well known that there is substantial heterogeneity concerning chemotherapeutic response between and among cancer types [23, 24]. Although there have been multiple attempts to identify molecular and clinical features predictive of response to particular targeted therapies, there remains considerable variability within subgroups identified using these factors. In this study, we attempt to accurately cluster cancer cell lines into defined groups based on response to a large range of chemotherapeutics, and to create a deep learning transcriptomic model capable of accurately categorizing samples into these defined groups.

Using cell line chemotherapeutic efficacy obtained from the GDSC consortium, we employed unsupervised clustering techniques to identify two defined therapeutic response groups with significantly different responses to a multitude of standard chemotherapies. We demonstrate that commonly-used clinical criteria to predict chemotherapeutic performance – such as anatomical location and morphologic subtype of cell lines, as well as TP53 and KRAS mutation status – failed to identify defined clusters of cells with similar therapeutic responses (Fig. 2). The consistent poor separability using these criteria and the relatively homogeneous distribution of cell responses to chemotherapeutics across major anatomical regions provides evidence for the utility of a pan-drug predictive biomarker. Given separability and the large feature space of the microarray data, a neural network was developed to classify cell lines into therapeutic response groups. This model may inform clinicians and researchers of the predicted therapeutic response of their patients to various chemotherapies.

Following the determination of two defined chemotherapeutic groups based on the GDSC dataset, we used a biologically agnostic feature selection algorithm, Boruta, to reduce the original set of 16,382 genes to a subset of 300 genes and to limiting preliminary bias due to multicollinearity in the model. The genes were then fed into neural networks, which were optimized using a grid search (Table 1). Each network was carefully monitored for overfitting during training (loss value observed to decrease during training across all models). To minimize overfitting, we applied L2 kernel regularization ($\lambda = 10^{-3}$), batch normalization, and dropout layers with a 0.3 dropout rate to each hidden layer. We determined from confusion matrices and ROC curve analysis that the best network architecture utilized 10 hidden layers, and demonstrated a 95.4% accuracy in classifying response groups (Fig. 4). This

validates our postulate that gene expression is a prime determinant of chemotherapeutic response, and that cell lines of similar gene expression profile respond similarly to most chemotherapies.

The comparatively lower accuracy of the neural network with 1 hidden layer (72.5%) suggests that the therapeutic response cohorts cannot be separated by a linear classifier and that classical machine learning techniques are insufficient to capture the complexity of the dataset. Furthermore, the decrease in accuracy from 10 hidden layers (95.4%) to 15 hidden layers (90.8%) indicates overfitting. To this end, a 10 hidden layer model is the ideal depth to analyze our data. Interestingly, there is no significant difference between the mean AUC for these models (Fig. 4). Given that there were no significant differences between the mean AUC for these models, the confusion matrices were used to select the 10 hidden layer deep learning architecture as our model of choice.

The deep learning transcriptomic model consists of 300 genes, with KEGG pathway enrichment suggesting that predictive genes are associated with numerous pathways, most notably PI3K signalling and focal adhesion. Interestingly, there is a growing body of literature that suggests that PI3K/Akt pathway dysregulation may be associated with chemotherapeutic resistance in numerous different cancer and treatment contexts [25]. Several studies have identified increases in Akt signalling in cancer cell lines exposed to chemotherapy and radiotherapy [26, 27, 28]. Moreover, significant increases in Akt have been identified in chemoresistant and radioresistant cancer models [29]. Similarly, several studies have identified focal adhesion as a potential protective mechanism for various cancer cells. In fact, inhibition of particular integrin isoforms has been shown to increase the susceptibility of various cancer cell lines to conventional chemo/radiotherapies [30]. Our results provide further evidence that dysregulation of PI3K signalling and focal adhesion may play a role in chemotherapy resistance in a pan-cancer context.

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 observations n will be much larger than the number of features p. To minimize this bias, we applied the Boruta algorithm to reduce our 16,382 genes by 541 cell lines dataset to a 300 by 541 matrix. The algorithm has been shown in various studies to be an effective feature selection method in high dimensional omics datasets [31]. To prevent overfitting of the reduced matrix, we applied L2 kernel regularization ($\lambda = 10^{-3}$), batch normalization, and dropout layers (dropout rate of 0.3).

Future investigations will look to validate the predictive ability of the model to categorize chemotherapeutic response in various cancer types using the current transcriptomic signature. It is likely that the model accuracy will vary between therapy targets, and 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. In addition, biochemical features of gene targets may provide additional insights, especially for interpretability of the neural network output. It may also be interesting to select relevant genes using a different feature selector given that Boruta specifically operates on patterns of statistical relationships rather than biological relationships. Our use of Boruta was motivated by its efficacy demonstrated by prior studies of the algorithm as compared to other feature selectors [31, 32]. It is possible that a feature selection method informed by gene function and linkage disequilibrium could yield a different set of relevant genes.

5 Conclusions

Using transcriptomic data from pan-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 subset 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 cell line therapy response with 91.7% accuracy. Future studies will investigate the efficacy of our model to predict chemotherapy response in various cancer types and treatment contexts.

Acknowledgements

References

[1] Sean Ekins, Ana C. Puhl, Kimberley M. Zorn, Thomas R. Lane, Daniel P. Russo, Jennifer J. Klein, Anthony J. Hickey, and Alex M. Clark5. Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater*, 18(5):435–441, May 2019.

[2] Cai Huang, Evan A. Clayton, Lilya V. Matyunina, L. DeEtte McDonald, Benedict B. Benigno, Fredrik Vannberg, and John F. McDonald. Machine learning predicts individual cancer patient responses to therapeutic drugs with

high accuracy. *Sci Rep*, 8:16444, November 2018.

[3] Akshat Pathak, Sanskriti Tanwar, Vivek Kumar, and Basu Dev Banarjeec. Present and future prospect of small molecule & related targeted therapy against human cancer. *Vivechan Int J Resp*, 9(1):36–49, March 2018.

[4] Brian A Baldo and Nghia H Pham. Adverse reactions to targeted and non-targeted chemotherapeutic drugs with emphasis on hypersensitivity responses and the invasive metastatic switch. *Cancer Metastasis Rev*, 32(3-4):723–761, December 2013.

[5] Joseph A. Sparano, Robert J. Gray, Della F. Makower, Kathleen I. Pritchard, Kathy S. Albain, Daniel F. Hayes, Charles E. Geyer Jr, Elizabeth C. Dees, Matthew P. Goetz, Jr. John A. Olson, Tracy Lively, Sunil S. Badve, Thomas J. Saphner, Lynne I. Wagner, Timothy J. Whelan, Matthew J. Ellis, Soonmyung Paik, William C. Wood, Peter M. Ravdin, Maccon M. Keane, Henry L. Gomez Moreno, Pavan S. Reddy, Timothy F. Goggins, Ingrid A. Mayer, Adam M. Brufsky, Deborah L. Toppmeyer, Virginia G. Kaklamani, Jeffrey L. Berenberg, Jeffrey Abrams, and George W. Sledge. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N Engl J Med*, 379(2):111–121, July 2018.

[6] Kevin Shee, Jason D. Wells, Amanda Jiang, and Todd W. Miller. Integrated pan-cancer gene expression and drug sensitivity analysis reveals slfn11 mrna as a solid tumor biomarker predictive of sensitivity to dna-damaging chemotherapy. *PLoS One*, 14(11):e0224267, November 2019.

[7] Xuewei Wang, Zhifu Sun, Michael T Zimmermann, Andrej Bugrim, and Jean-Pierre Kocher. Predict drug sensitivity of cancer cells with pathway activity inference. *BMC Med Genomics*, 12(15), January 2019.

[8] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, Sridhar Ramaswamy, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Cyril Benes, Ultan McDermott, and Mathew J Garnett. Genomics of drug sensitivity in cancer (gdsc): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*, 41:D955–61, November 2012.

[9] Mehreen Ali and Tero Aittokallio. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev*, 11(1):31–39, February 2019.

[10] Stacey Shiovitz and William M Grady. Molecular markers predictive of chemotherapy response in colorectal cancer. *Curr Gastroenterol Rep*, 17(2):431, February 2015.

[11] Vincenzo Catalano, Anna Maria Baldelli, Paolo Giordani, and Stefano Cascinu. Molecular markers predictive of response to chemotherapy in gastrointestinal tumors. *Crit Rev Oncol Hemat*, 38(2):93–104, May 2001.

[12] Ikuo Sekine, John D Minna, Kazuto Nishio, Tomohide Tamura, and Nagahiro Saijo. A literature review of molecular markers predictive of clinical response to cytotoxic chemotherapy in patients with lung cancer. *J Thorac Oncol*, 1(1):31–37, January 2006.

[13] I F Faneyte, J G Schrama, J L Peterse, P L Remijnse, S Rodenhuis, and M J van de Vijver. Breast cancer response to neoadjuvant chemotherapy: predictive markers and relation with outcome. *Br J Cancer*, 86:406–412, February 2003.

[14] Raivo Kolde. *pheatmap: Pretty Heatmaps*, 2019. R package version 1.0.12.

[15] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020. R package version 1.0.7.

[16] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[17] Huan Liu. *Feature Selection*, pages 402–406. Springer US, Boston, MA, 2010.

[18] Rudnicki W. Kursa, M. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11), 2010.

[19] Francois Chollet et al. Keras, 2015.

[20] Shie-Yui Liong, Soon-Thiam Khu, and Weng-Tat Chan. Derivation of pareto front with genetic algorithm and neural network. *Journal of Hydraulic Engineering*, 6(1), January 1998.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[23] Ariel Pribluda, Cecile C. de la Cruz, and Erica L. Jackson. Intratumoral heterogeneity: From diversity comes resistance. *Clin Cancer Res*, 21(13):2916–2923, July 2015.

[24] Corbin E Meacham and Sean J Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501:7467, 328-337 2013.

[25] Wei-Chien Huang and Mien-Chie Hung. Induction of akt activity by chemotherapy confers acquired resistance. *Journal of the Formosan Medical Association*, 108(3):180 – 194, 2009.

[26] J M Nelson and D W Fry. Akt, mapk (erk1/2), and p38 act in concert to promote apoptosis in response to erbb receptor family inhibition. *J Biol Chem*, 276:14842–14847, 2001.

[27] S S Ng, M S Tsao, and T Nicklee. Wortmannin inhibits pkb/akt phosphorylation and promotes gemcitabine anti-tumor activity in orthotopic human pancreatic cancer xenografts in immunodeficient mice. *Clin Cancer Res*, 7:3269–3275, 2001.

[28] S S W Ng, M S Tsao, and S Chow. Inhibition of phos-phatidylinositide 3-kinase enhances gemcitabine-induced apoptosis in human pancreatic cancer cells. *Cancer Res*, 60:5451–5455, 2000.

[29] K. Yokoi, A. Kobayashi, H. Motoyama, M. Kitazawa, A. Shimizu, T. Notake, T. Yokoyama, T. Matsumura, M. Takeoka, and S. I. Miyagawa. Survival pathway of cholangiocarcinoma via akt/mtor signaling to escape raf/mek/erk pathway inhibition by sorafenib. *Oncology reports*, 39(2):843–850, Feb 2018. LR: 20181202; JID: 9422756; 0 (FOXO1 protein, human); 0 (Forkhead Box Protein O1); 0 (Phenylurea Compounds); 0 (Protein Kinase Inhibitors); 25X51I8RD4 (Niacinamide); 9HW64Q8G6G (Everolimus); 9ZOQ3TZI87 (Sorafenib); EC 2.7.1.1 (MTOR protein, human); EC 2.7.1.1 (TOR Serine-Threonine Kinases); EC 2.7.11.1 (Proto-Oncogene Proteins c-akt); 2017/07/12 00:00 [received]; 2017/12/07 00:00 [accepted]; 2017/12/19 06:00 [entrez]; 2017/12/19 06:00 [pubmed]; 2018/08/24 06:00 [medline]; ppublish.

[30] Iris Eke and Nils Cordes. Focal adhesion signaling and therapy resistance in cancer. *Seminars in Cancer Biology*, 31:65 – 75, 2015. Intracellular Signaling and Response to Anti-Cancer Therapy.

[31] 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.

[32] Theodore Sakellaropoulos, Konstantinos Vougas, Sonali Narang, Filippos Koinis, Athanassios Kotsinas, AlexanderPolyzos, Tyler J Moss, Sarina Piha-Paul, Hua Zhou, Eleni Kardala, Eleni Damianidou, Leonidas G Alexopoulos, Iannis Aifantis, Paul A Townsend, Mihalis I Panayiotidis, Petros Sfikakis, Jiri Bartek, Rebecca C Fitzgerald, Dimitris Thanos, Kenna R Mills Shaw, Russell Petty, Aristotelis Tsirigos, and Vassilis G Gorgoulis. A deep learning framework for predicting response to therapy in cancer. *Cell Reports*, 29(11):3367–3373, December 2019.
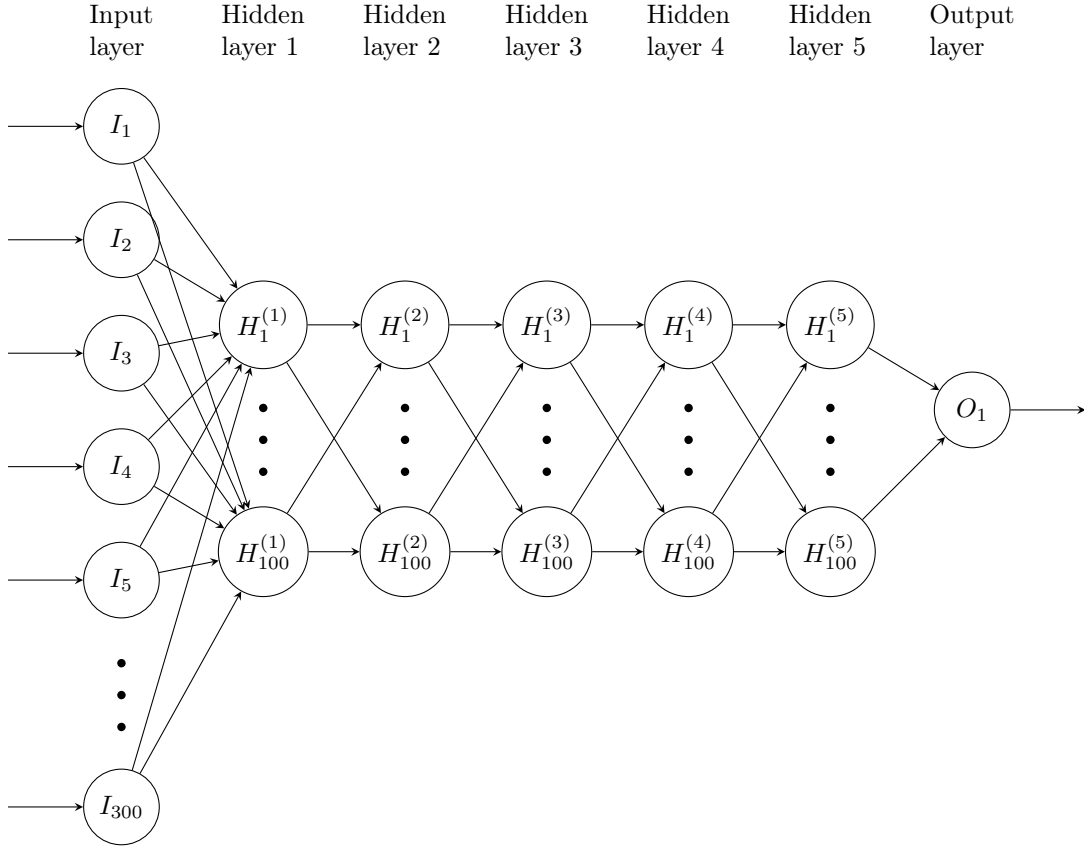
Supplementary Data

Figures

Figure 6: Neural network architecture representation with 5 hidden layers (300 inputs). Inputs include the feature-selected genes from the cell line microarray dataset. Each hidden layer has an L2 kernel regularization ($\lambda = 10^{-3}$) and a dropout rate of 0.3. All hidden layers are subject to batch normalization.

Tables

Table 1: Grid search parameters to optimize all implemented neural network architectures (1, 5, 10, and 15 hidden layers). Each grid search underwent 3-fold cross-validation on the training data. Optimal parameters: 25 epochs, batch size of 16, 100 neurons, $\lambda = 10^{-3}$, stochastic gradient descent (sgd) as optimizer, normal kernel initializer.

Epochs	Batches	Neurons	λ (L2 reg)	Optimizer	Kernel initializer
25	16	100	1e-4	Sgd	Normal
50	32	200	1e-3	Adagrad	Uniform
75	64	300	1e-2	Adam	Glorot uniform