STEM Fellowship Big Data Challenge 2020

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

July 4, 2020

Abstract

The emergence of precision oncology approaches have begun to inform clinical decision-making in diagnostic, prognostic, and treatment contexts. High-throughput technology has enabled machine learning algorithms to utilize molecular characteristics of tumours to generate personalized treatments, revealing relationships between genomic alterations and drug sensitivity. However, precision oncology studies have yet to generate a predictive biomarker incorporating gene expression profiles to stratify tumours into similar pan-drug sensitivity profiles. Here we show that a neural network with 10 hidden layers accurately classifies pan-cancer cell lines into two distinct chemotherapeutic response groups with respect to a pan-drug dataset with 89.0% accuracy (AUC = 0.904, weighted F1 score = 0.89). From unsupervised clustering algorithms, we found that a cohort of cell line gene expression data from the Genomics of Drug Sensitivity in Cancer could be clustered into two response groups with significant differences in pan-drug chemotherapeutic sensitivity. We used the Boruta feature selection algorithm to identify genes of the highest predictive value; the 300 selected genes were enriched primarily for the focal adhesion, ECM-receptor and proteoglycan interaction pathways. Using the selected genes, a deep learning neural network was developed to predict response groups. The classification efficacy of the neural network validates our postulate that cell lines with similar gene expression profiles present similar pan-drug chemotherapeutic sensitivity, and it provides evidence for the potential utility of similar combinatorial biomarkers for the selection of potent candidate drugs. Given that not all tumours present with targetable features and are treated with less targeted and consequently more cytotoxic chemotherapies, our computational framework allows for the identification of candidate drugs for tailored treatments that maximize therapeutic response and minimize cytotoxic burden.

Keywords
Clustering, neural network, transcriptomics, chemotherapeutic response, combinatorial biomarker, molecular profile, therapeutic sensitivity, cancer

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]. These approaches attempt to improve both targeted and conventional therapeutics. Existing precision oncology approaches have largely focused on 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 also potentiated 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 and cancer types. However, a pan-cancer and 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 most chemotherapies. 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 traditional clinical criteria (ex: tumour location, TP53 and KRAS mutation status), which fail to capture the dimensionality of available transcriptomic data [11, 12, 13, 14]. Furthermore, advanced deep learning approaches that are capable of handling such dimensionality often fail to allow interpretability and consequently require transcriptomic data that is clinically infeasible [10]. Thus, deep learning approaches should minimize the number of selected transcriptomic features to maximize both the accuracy and interpretability of such predictive biomarkers in a clinical context. This allows for meaningful pathway functional enrichment analyses which provide a deeper understanding of the underlying biological significance of features contributing to the deep learning model.

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. Further, this biomarker may prove to be a useful technique to stratify patients into distinct treatment response groups, allowing chemotherapy regimens to be tailored to the transcriptomic characteristics of a particular cancer.

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 their therapeutic response. This substantiates the existence of substantial chemotherapeutic response heterogeneity between and within these clinical subgroups. Further, this motivates the need for a biomarker capable of accurately classifying tumors into response groups in order to effectively guide clinical management in the context of chemotherapeutic treatment. To create such a predictive marker, we utilize a biologically agnostic feature selection algorithm to iteratively select and identify a subset of 300 relevant genes predictive of chemotherapeutic sensitivity. In this study, we develop a deep learning model from the selected genes, which showcases a strong predictive ability to stratify pan-cancer cell lines into therapeutic response groups.

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 [8]. Following data collection and curation, we utilized unsupervised clustering algorithms to define two groups based on their 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. Neural network architectures were optimized using a grid search and evaluated using 5-fold stratified cross-validation on the training data. Final model evaluation was performed on the testing dataset. 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 lines and associated therapeutic sensitivity data were obtained from the GDSC database. The 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 dataset was used to generate a matrix with cell lines and accompanying therapeutic information. The dataset was then 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. The filtered dataset was analyzed to ensure that it was still representative of a relatively pan-cancer dataset.

2.2 Identification of pan-cancer therapeutic response cohorts

Cell line therapeutic sensitivity matrices were used to evaluate whether conventional clinical crite-

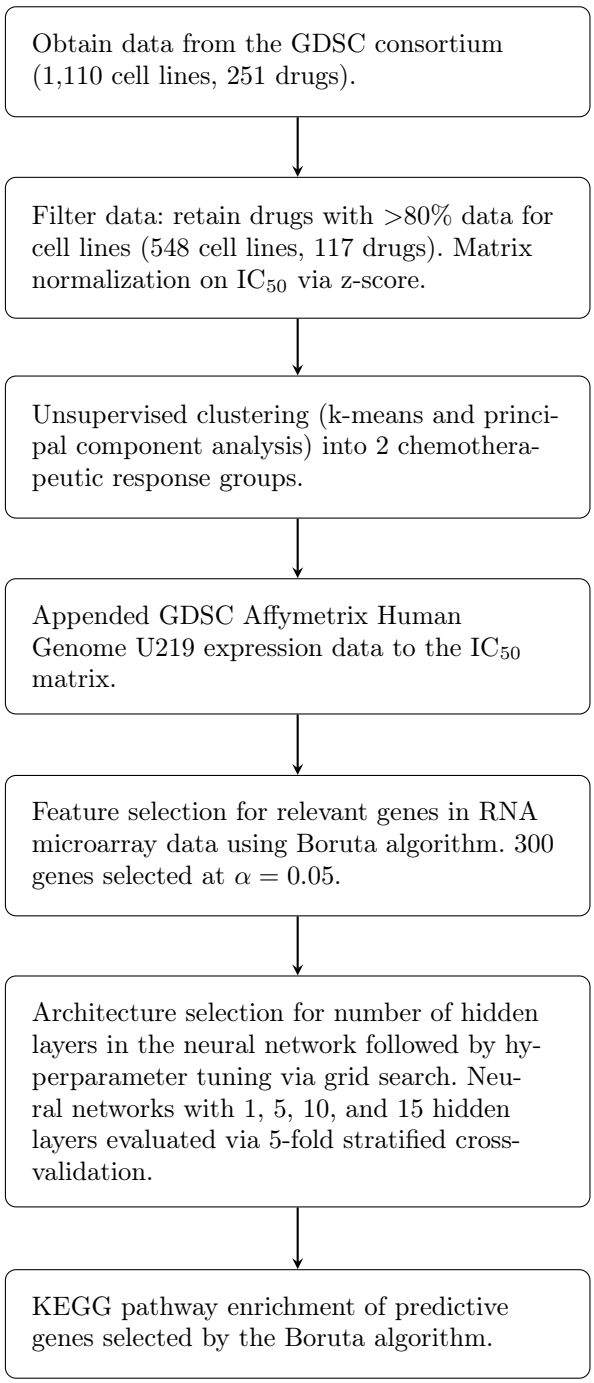

Figure 1: Summary of the data analysis pipeline.

ria 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 [11, 12, 13, 14]. Cell lines were separated into subgroups based on these criteria and visualized to determine whether these criteria effectively clustered response groups.

Following the 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 normalized Euclidean distance matrix for the retained cell lines based on their pan-chemotherapy response. The matrix was 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. From the matrix, k-means clustering was utilized to assign cell line candidates to appropriate therapeutic response cohorts (the elbow method was used as a heuristic for determining the number of clusters). Generalized differences in chemotherapeutic efficacy between cohorts were visualized using a heatmap generated by the pheatmap package [15] in R. The heatmap displayed 2 groups clustered using k-means clustering. The separation between clusters was also visualized using principal component analysis with the plotly package [16] 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 IC_{50} values between the groups. False discovery rate (FDR) correction was utilized to correct for multiple comparisons. To further validate the existence of these distinct chemotherapeutic response clusters, intra-cancer heterogeneity was evaluated. 11 cancer subtypes with sufficient cell line representation ($n > 20$) were identified. Differences in chemotherapeutic efficacy between cell lines assigned to group A and B were visualized using heatmaps.

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 [17] 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 (for example, univariate selection or recursive feature elimination) as well as training and optimizing neural networks. We addressed these issues with the Boruta feature selection algorithm using the BorutaPy package [18] in Python 3. Boruta is a feature selection algorithm based on Random Forest classification which iteratively removes features that are statistically less significant than a shuffled version of the same feature [19].

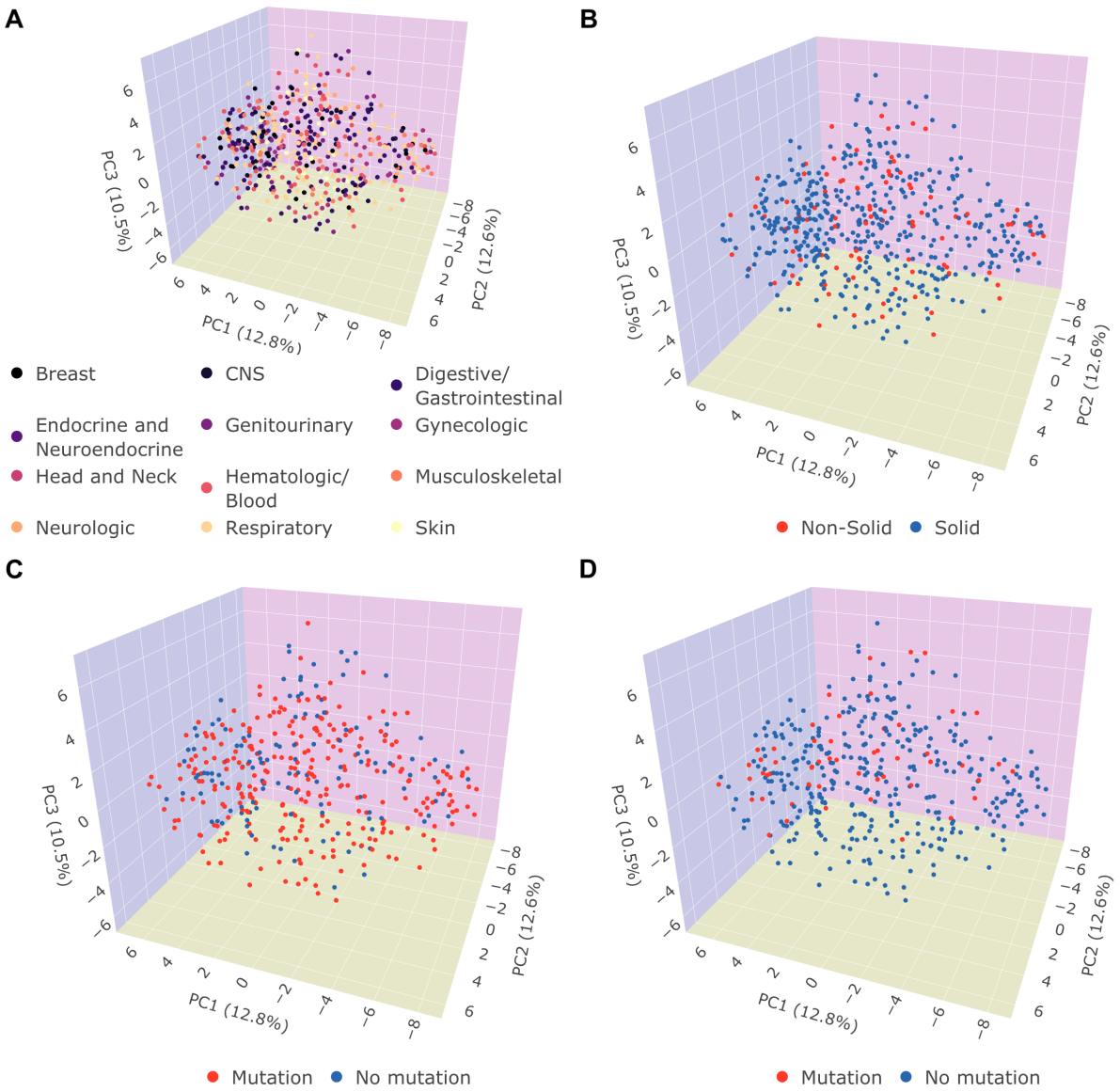

Figure 2: Principal component analysis of pan-cancer cell line therapeutic efficacy generated from IC_{50} values of all available chemotherapeutics.The x-axis shows the first principal component, the y-axis shows the second component, and the z-axis shows the third principal component. Cell lines are 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.

2.4 Classification of cell lines using an optimized neural network

The neural network was constructed using the Tensorflow Keras sequential deep learning API [20] 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 carefully 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 [21]; we reserved 80% of the data (432 training samples, 541 total) for training and the remaining 20% for testing. Model selection was performed by hyperparameter tuning using a grid search followed by 5-fold stratified cross-validation on the training data (Table S1). All grid searches were performed with 3-fold cross-validation on the training data to determine the parameters that minimize the binary cross-entropy loss function. GridSearchCV from the scikit-learn library [22] was used to iterate through combinations of epochs, batch size, neurons per hidden layer, L2 regularization

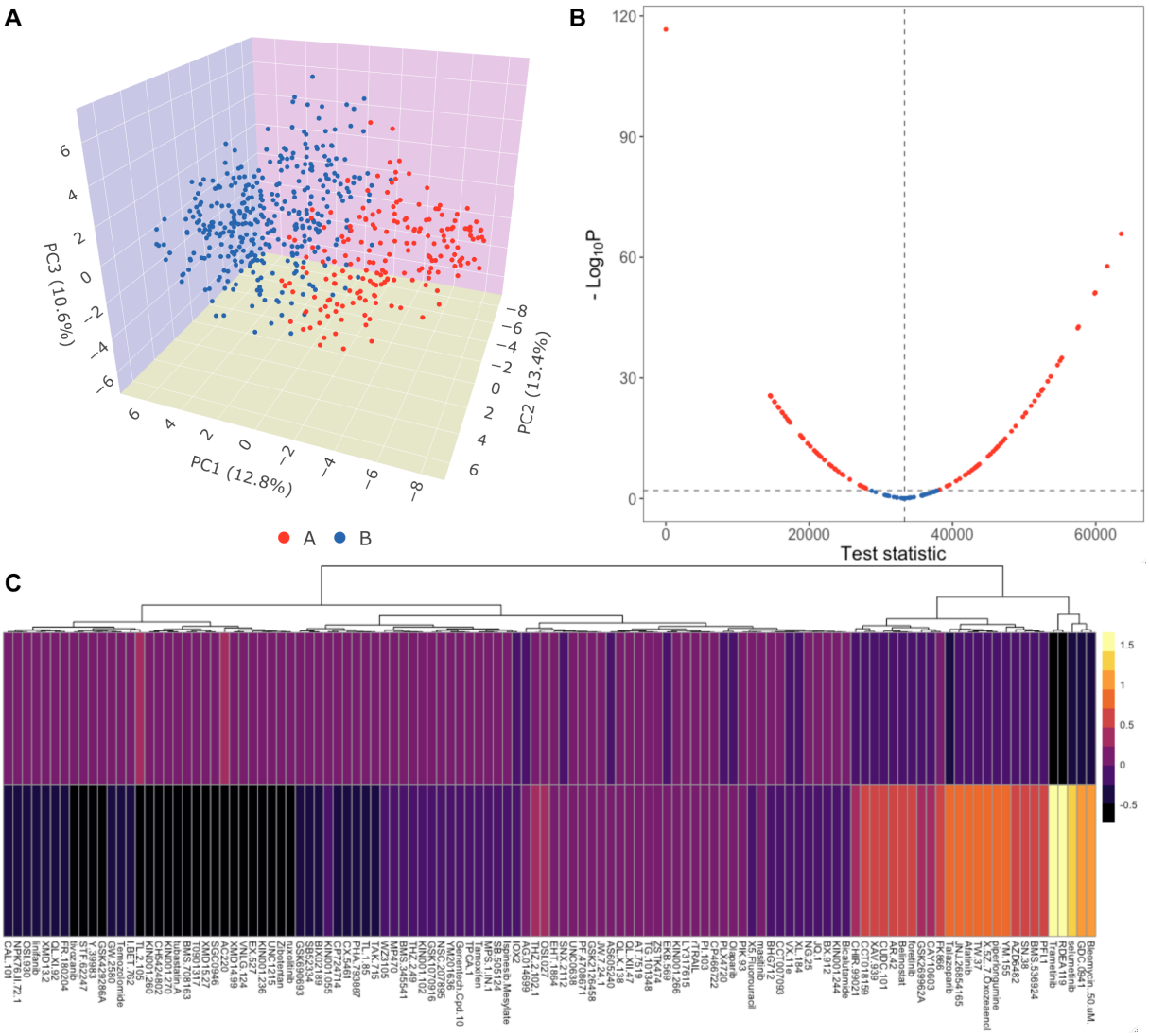

Figure 3: (**A**) Principal component analysis of pan-cancer cell line therapeutic efficacy generated from IC_{50} values of all available chemotherapeutics. The x-axis shows the first principal component, the y-axis shows the second principal component, and the z-axis shows the third principal component. The two identified therapeutic response clusters (A and B) 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.

penalty, optimizer, and kernel initializer to find the optimal model (Table S1). To prevent class imbalance during training, we used the Synthetic Minority Oversampling Technique (SMOTE) from the imblearn package [23] for Python 3 to generate synthetic data on the training and validation datasets only. Each model's performance on the validation dataset was evaluated by the Area Under Curve (AUC) value of the receiver operating characteristic (ROC) curve.

Following grid search and cross-validation, we evaluated the performance of our model on the testing data. Youden's index from the ROC curves corresponding to each neural network's performance on the training dataset was used to determine the optimal threshold to classify cell lines into chemotherapeutic response groups for the testing dataset. Youden's index ranges from 0 to 1 inclusive, and an index of 0 indicates that the binary classifier is no better than that obtained from a fair coin flip. Conversely, a value of 1 indicates that the test is perfect (i.e., no false positives or false negatives). The formula for Youden's index is defined as

$$J \doteq \text{sensitivity} + \text{specificity} - 1$$

All indices were obtained from the ROC curves by an iterative search for the maximum J. That is,

$$J = \text{argmax}(\text{TPR} - \text{FPR})$$

where TPR and FPR (FPR $= 1 -$ specificity) are

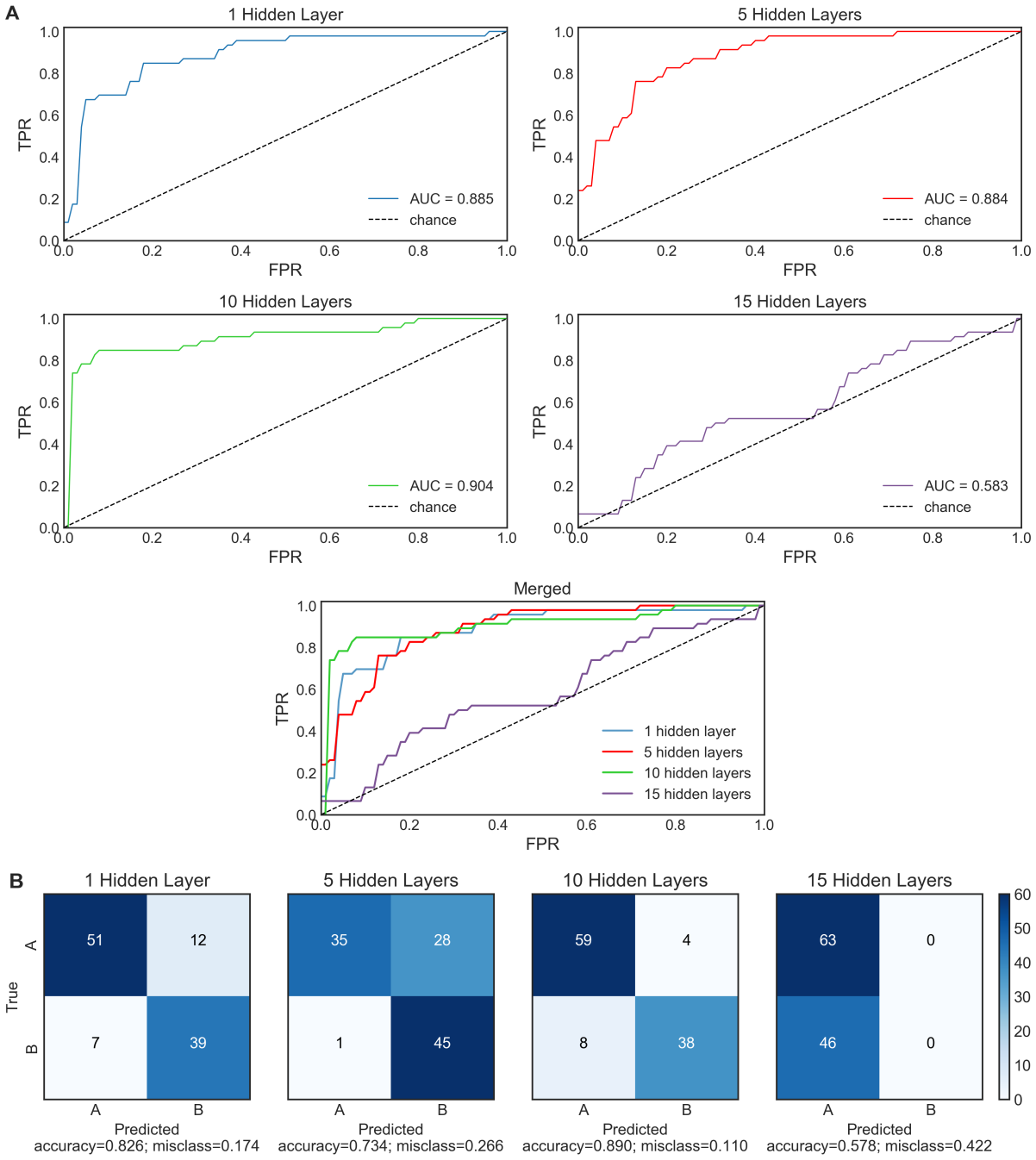

Figure 4: (**A**) ROC curves for neural networks with 1, 5, 10, and 15 hidden layers on the testing dataset. (**B**) Confusion matrices for the 1, 5, 10, and 15 hidden layer neural network models on the testing set. The threshold for classification is determined using Youden's index from the training dataset (thresholds were 0.17, 0.10, 0.49, and 0.45 for neural networks with 1, 5, 10, and 15 hidden layers respectively). The models classify cell line microarray datasets into chemotherapy response cohorts A and B.

arrays containing the true positive and false positive rate corresponding to each point on the ROC curve. The threshold corresponding to J was used to classify cell lines.

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. Further analysis of the filtered dataset suggested that it was still rep-

resentative of a pan-cancer dataset. Next, 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. It is clear that common molecular markers such as TP53 and KRAS mutation status, as well as clinical markers like cancer type or solidity are unable to effectively stratify chemotherapeutic efficacy (Fig. 2).

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 normalized dataset, allowing for visualization of defined therapeutic response groups. This process identified two distinct clusters of therapeutic sensitivity (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 the 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. Comparison of identified chemotherapeutic response cohorts between cancer subtypes shows considerable intra-cancer chemotherapeutic heterogeneity (Fig. S1).

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

3.3 A neural network with 10 hidden layers accurately classifies patients into responder and non-responder cohorts

Unsupervised learning via k-means clustering of the cancer cell line transcriptomes indicated substantially different responses to chemotherapies. Since k-means clustering instantiates randomly-placed centroids, we ran the algorithm several times, and each iteration returned similar results. Using these distinct therapy response cohorts, we developed a

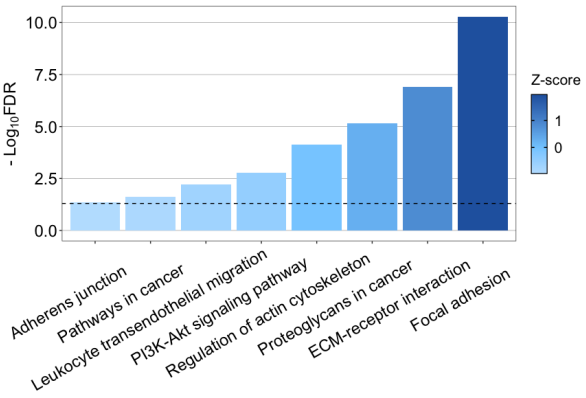

Figure 5: KEGG pathway functional enrichment for feature-selected genes included in the deep learning model. The vertical dotted line indicates the threshold for significance (adjusted $p < 0.05$).

deep learning binary classifier to predict drug response groups based on transcriptomic data. We analyzed four neural network architectures: 1, 5, 10, and 15 hidden layers (Fig. S2). Prior to model evaluation, SMOTE was used to prevent class imbalance in the training dataset; no synthetic samples were added to response group A and 146 synthetic samples were added to response group B (original data had 289 samples in response group A and 143 samples in response group B; 432 total original samples, 578 total samples with SMOTE). SMOTE was not applied to the testing dataset to ensure generalizability to real-world data. During training, hyperparameter optimization via grid search with 3-fold cross-validation returned similar results for each model: 50 epochs, batch size of 32, 50 neurons per hidden layer, L2 kernel regularization with $\lambda = 10^{-3}$, Adam as the optimizer, and a normal kernel initializer. Following this, we validated the various architectures with 5-fold stratified cross-validation.

On the training data, the neural network with 5 hidden layers performed the best on the validation fold with 93.1% accuracy (88.8% for 1 hidden layer, 84.5% for 10 hidden layers, 56.9% for 15 hidden layers). The mean AUC across the 5 folds were 0.965 ± 0.017, 0.957 ± 0.0144, 0.913 ± 0.026, and 0.515 ± 0.115 for the neural networks with 1, 5, 10, and 15 hidden layers respectively (Fig. S3). Here the relative ratio between the model's false positive rate and its true positive rate was averaged between 5 stratified k-folds. The mean ROC curves plotted from the training set showed relatively small variance across all 5 k-folds, as evidenced by the small width of the confidence interval (Fig. S3). The mean AUC for the 5 trials was used to compare the four network architectures. While there were no significant differences between AUC for the 1, 5, and 10 hidden layer models, the model with 15 hidden layers performed significantly worse compared to the rest. Youden's index was determined from the ROC

curves, and the thresholds were 0.17, 0.10, 0.49, and 0.45 for neural networks with 1, 5, 10, and 15 hidden layers respectively. These thresholds were used to classify the cell lines into chemotherapeutic response groups in the final evaluation of neural network performance (i.e., on the testing dataset).

The testing dataset showed slightly different metrics to that from the training dataset (Fig. 4). Similar to the training data, the neural network with 15 hidden layers performed the worst of the models evaluated. The neural network with 10 hidden layers performed the best on the testing dataset with an accuracy of 89.0% (82.6% for 1 hidden layer, 73.4% for 5 hidden layers, and 57.8% for 15 hidden layers). The AUC was 0.885, 0.884, 0.904, and 0.583 for the neural networks with 1, 5, 10, and 15 hidden layers respectively. Of note, the neural network with 15 hidden layers classified all cell lines into the responder group. The weighted F1 scores for the neural networks with 1, 5, 10, and 15 layers were 0.83, 0.73, 0.89, and 0.42 respectively.

4 Discussion

Differences in chemotherapeutic response between and within cancer types and clinical subgroups represents a critical challenge to current clinical management [24, 25]. Heterogeneity in treatment response results in subsets of patients bearing significant side-effect burden with minimal treatment efficacy, substantially limiting quality of life and limiting future treatment course. Although there have been multiple attempts to identify molecular and clinical features predictive of response to targeted chemotherapies, there remains considerable variability within subgroups (ex: anatomical location of the tumour) identified using these factors. In this study, we 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 provide evidence for the utility of a pan-drug predictive biomarker. Given the 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 limit preliminary bias due to multicollinearity in the neural network models. The genes were then fed into neural networks, which were optimized using a grid search (Table S1). Each network was carefully monitored for overfitting during training (loss value on the validation set was 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. To account for imbalanced classes during the cross-fold evaluations, we utilized SMOTE and stratified k-folds on the training dataset. The mean ROC curves plotted from the training set showed relatively small variance across all 5 k-folds, suggesting that the error from each model is relatively stable, and thus generalizable to real-world data.

Evaluation of the final neural network models on the testing dataset showed that the best network architecture utilized 10 hidden layers, demonstrating an 89.0% accuracy in classifying cell lines into chemotherapeutic 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 significantly lower accuracy of the neural network with 15 hidden layers on the testing data (57.8%) as compared to the networks with 1, 5, and 10 hidden layers indicates severe overfitting. Overfitting was also evidenced by both the low AUC and high loss for this model during all stages of the neural network evaluation pipeline. The lower accuracy of the neural networks with 1 (82.6%) and 5 hidden layers (73.4%) compared to that with 10 hidden layers (89.0%) suggests that 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. To this end, a 10 hidden layer model is the ideal neural network depth to analyze the GDSC dataset.

The deep learning transcriptomic model consists of 300 gene inputs, and KEGG pathway enrichment analyses indicated that predictive genes are significantly associated with numerous pathways, most notably the PI3K signalling and focal adhesion pathways (Fig. 5). 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 [26]. Several studies have identified increases in Akt signalling

in cancer cell lines exposed to chemotherapy and radiotherapy [27, 28, 29]. Moreover, significant increases in Akt have been identified in chemoresistant and radioresistant cancer models [30]. 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 [31]. 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 [32].

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 the 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. As such, Boruta is sensitive to hyperparameter values. Our use of Boruta was primarily motivated by its efficacy demonstrated by prior studies of the algorithm as compared to other feature selectors [32, 33]. 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. The Boruta feature selection algorithm was used to select a 300 gene subset that served as inputs to multiple neural networks. We determined that a network with 10 hidden layers was the most accurate model, producing a binary classifier to predict cell line therapy response with 89.0% accuracy. To our knowledge, a pan-cancer, pan-drug chemotherapeutic classification model has not been investigated. Future studies will investigate the efficacy of our model to predict chemotherapy response in various cancer types and treatment contexts.

Acknowledgements

We wish to acknowledge the STEM Fellowship for organizing the 2020 Big Data Challenge, as well as Roche, SAS, Canadian Science Publishing, Digital Science, Altmetric, and Overleaf for their contributions that enabled this competition. We would like to thank our mentor, Dr. Daiva Nielsen for her feedback on our paper. We would also like to acknowledge Matthew Pietrosanu, Stephen Styles, Danyi Liu, and Boya Peng for their critical statistical review of the manuscript draft.

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] Roger Weber, Hans-Jörg Schek, and Stephen Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, page 194–205, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] Carson Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020.

[17] 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.

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

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Theodore Sakellaropoulos, Konstantinos Vougas, Sonali Narang, Filippos Koinis, Athanassios Kotsinas, Alexander Polyzos, 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.

S Supplementary Data

S.1 Figures

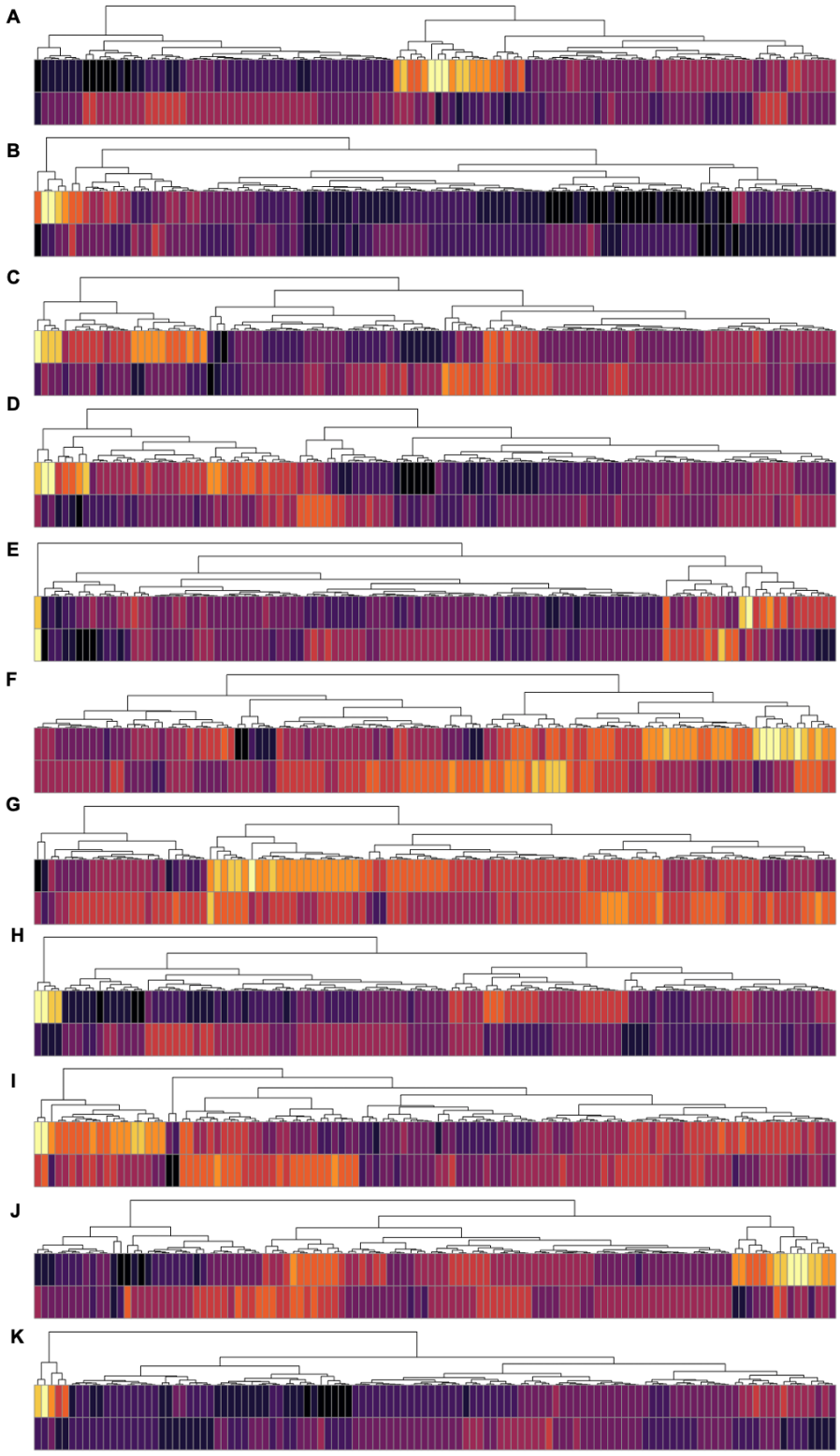

Figure S1: Intra-cancer differences in chemotherapeutic efficacy between identified chemotherapeutic clusters. Clusters are arranged into cluster A (upper group) to cluster B (lower group) for each figure. The cancers represented are (**A**) non-small cell lung cancer, (**B**) breast cancer, (**C**) other digestive tract cancers, (**D**) small-cell lung cancer, (**E**) other nervous system cancers, (**F**) skin cancers, (**G**) urogenital cancers, (**H**) lymphomas, (**I**) colorectal cancers, (**J**) bone cancers, and (**K**) leukemia.

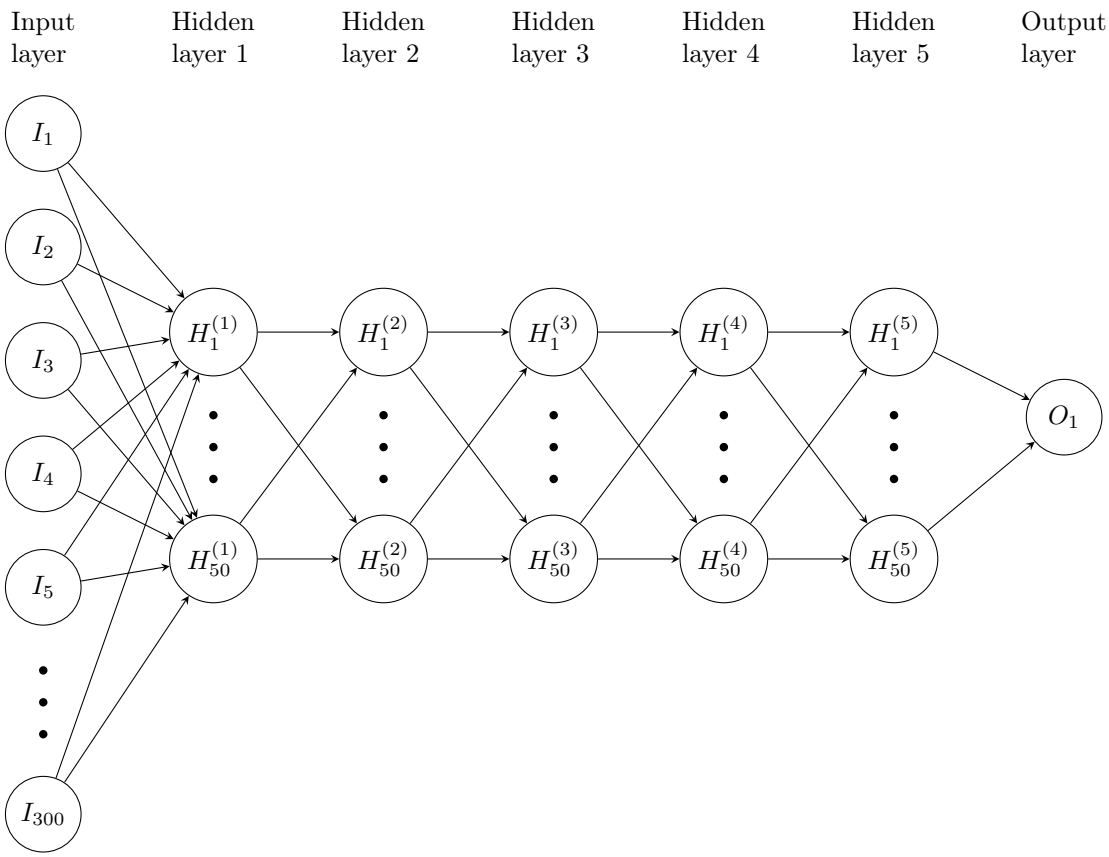

Figure S2: Neural network architecture representation with 5 hidden layers (300 inputs). Inputs include the feature-selected genes from the cell line microarray dataset. Each hidden layer is batch-normalized with an L2 kernel regularization penalty ($\lambda = 10^{-3}$) and a dropout rate of 0.3.

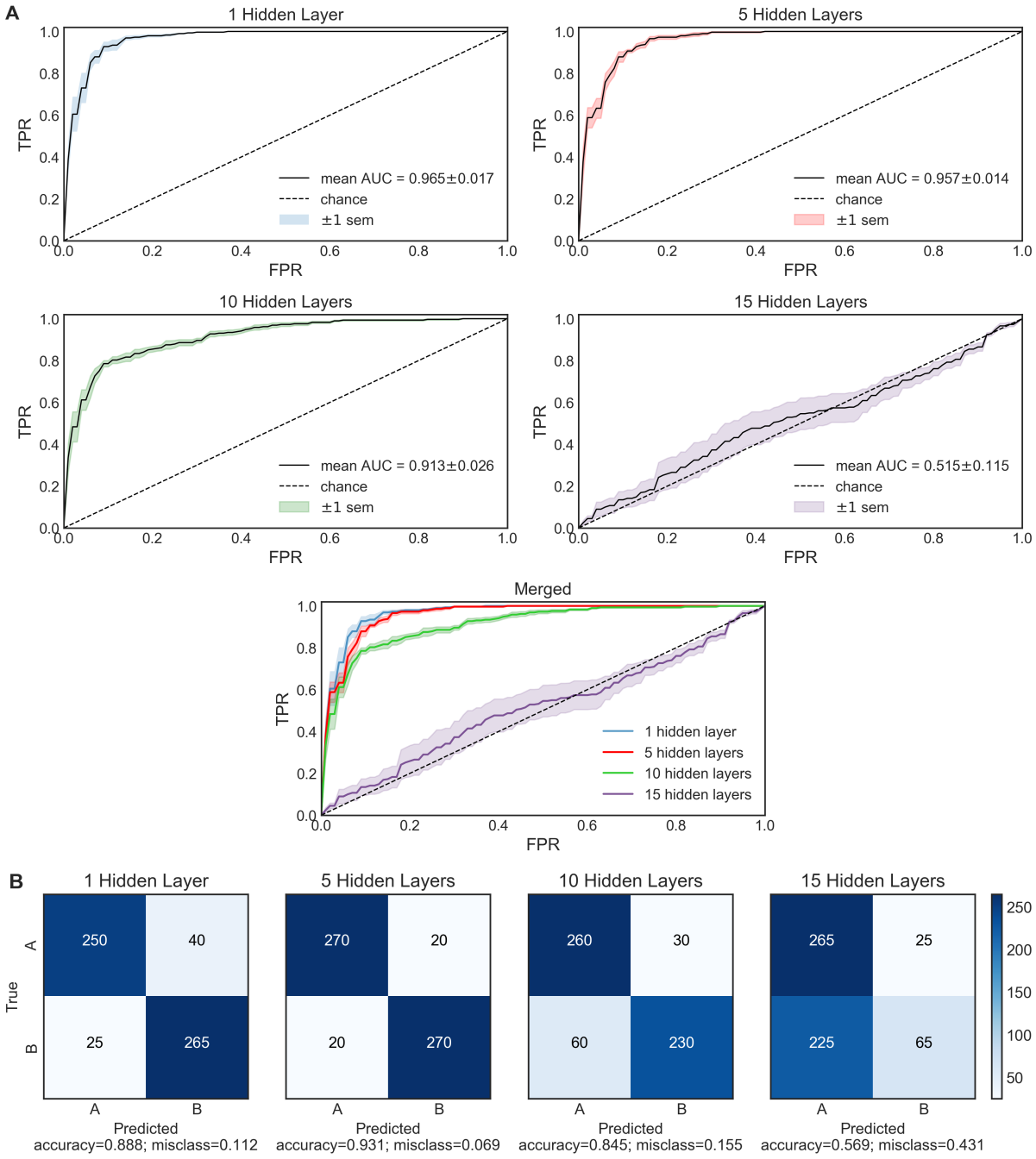

Figure S3: (**A**) ROC curves for 1, 5, 10, and 15 hidden layers neural network models with confidence bands of ±1 standard error of the mean (sem). Each model was subject to 5-fold stratified cross-validation on the training set and the mean ROC curve for the validation set across all trials was plotted. The threshold corresponding to Youden's index was used to classify cell lines into response groups for the testing data (thresholds were 0.17, 0.10, 0.49, and 0.45 for neural networks with 1, 5, 10, and 15 hidden layers respectively; corresponding Youden's indices reported in Table S1). (**B**) Confusion matrices for the 1, 5, 10, and 15 hidden layer neural network models on the validation set. The models classify cell line microarray datasets into chemotherapy response cohorts. The threshold to classify cell lines into response cohorts is 0.5.

S.2 Tables

Table S1: 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: 50 epochs, batch size of 32, 50 neurons, $\lambda = 10^{-3}$, Adam as optimizer, normal kernel initializer.

Epochs	Batches	Neurons	λ (L2 reg)	Optimizer	Kernel initializer
25	16	50	1e-4	Sgd	Normal
50	32	100	1e-3	Adagrad	Uniform
75	64	150	1e-2	Adam	Glorot uniform

Table S2: Classification scores for the 1, 5, 10, and 15 neural networks on the testing data.

Hidden layers	Accuracy	AUC	Weighted F1 score	Youden's index
1	82.6%	0.885	0.83	0.736
5	73.4%	0.885	0.73	0.690
10	89.0%	0.904	0.89	0.780
15	57.8%	0.583	0.42	0.500