Machine learning to predict developmental neurotoxicity with high-throughput data from 2D bio-engineered tissues

Finn Kuusisto[1,*], **Vitor Santos Costa**[2], **Zhonggang Hou**[1,+], **James Thomson**[1,3,4], **David Page**[5], **Ron Stewart**[1]

[1]Morgridge Institute for Research, Regenerative Biology, Madison, WI, USA

[2]University of Porto, Department of Computer Science, Porto, Portugal

[3]University of Wisconsin, Department of Cell and Regenerative Biology, Madison, WI, USA

[4]University of California, Department of Molecular, Cellular, and Developmental Biology, Santa Barbara, CA, USA

[5]Duke University, Department of Biostatistics and Bioinformatics, Durham, NC, USA

Abstract

There is a growing need for fast and accurate methods for testing developmental neurotoxicity across several chemical exposure sources. Current approaches, such as ***in vivo*** animal studies, and assays of animal and human primary cell cultures, suffer from challenges related to time, cost, and applicability to human physiology. Prior work has demonstrated success employing machine learning to predict developmental neurotoxicity using gene expression data collected from human 3D tissue models exposed to various compounds. The 3D model is biologically similar to developing neural structures, but its complexity necessitates extensive expertise and effort to employ. By instead focusing solely on constructing an assay of developmental neurotoxicity, we propose that a simpler 2D tissue model may prove sufficient. We thus compare the accuracy of predictive models trained on data from a 2D tissue model with those trained on data from a 3D tissue model, and find the 2D model to be substantially more accurate. Furthermore, we find the 2D model to be more robust under stringent gene set selection, whereas the 3D model suffers substantial accuracy degradation. While both approaches have advantages and disadvantages, we propose that our described 2D approach could be a valuable tool for decision makers when prioritizing neurotoxicity screening.

Keywords

machine learning; neurotoxicity; tissue model; gene expression

---

[*]fkuusisto@morgridge.org.

[+]Current Address: University of Michigan, Department of Biological Chemistry, Ann Arbor, MI, USA

I. INTRODUCTION

The Toxic Substances Control Act (TSCA) lists 84,000 chemicals, almost all of which have not been tested for developmental neurotoxicity [1]. The developing human brain is especially sensitive to toxic exposures [2], and estimated costs of early developmental neurotoxicity exposure are enormous [3], [4]. Fast, inexpensive, and accurate methods for testing developmental neurotoxicity are thus urgently needed.

Current approaches involve *in vivo* animal studies, and assays of animal and human primary cell and tissue cultures. These approaches suffer from challenges such as time, cost, availability of primary cells and tissues, and poor applicability to human physiology [5], [6], [7]. The result of these difficulties can be observed in part by a decrease in drug approval rates despite increases in research and development spending [8]. Human pluripotent stem cells can help address these challenges by providing a scalable source of applicable human cells at relatively low cost.

In 2012, the National Institutes of Health (NIH) launched the Microphysiological Systems Program [6], a collaboration with the Defense Advanced Research Projects Agency (DARPA) and the U.S. Food and Drug Administration (FDA), to develop human tissue chips containing bio-engineered tissue models that mimic human physiology. The aim of this program is to use these chips to predict the safety and efficacy of candidate drugs. Within this program, Schwartz et al. developed 3D constructs of developing human neural tissue from several cell types and trained machine learning models to predict developmental neurotoxicity with gene expression data gathered from these constructs [9]. The 3D model is capable of accurately identifying compounds that are neurotoxic, and it shows potential for recapitulating relevant biological mechanisms, thus demonstrating potential for further insight. Nevertheless, this 3D model is necessarily complex and requires extensive expertise and manual effort to construct and employ successfully. In cases where the goal is only to construct an assay of developmental neurotoxicity, we propose that it may be sufficient to use a simpler 2D tissue model cultured from a single cell type.

Here, we report results applying off-the-shelf machine learning algorithms to a simpler 2D model of neural tissue. We run several experiments with varying numbers of chemical exposure lengths and feature selection methods. We compare the accuracy of learned models between those trained on data from a 2D tissue model and those trained on data from a 3D tissue model. Importantly, we find that our accuracy in distinguishing known human developmental neurotoxins from non-toxins is substantially higher when using the 2D model than when using the 3D model. We describe our data collection and predictive experiments in the next section, follow with a discussion of results, and finish with conclusions and proposals for future work.

II. MATERIALS AND METHODS

For all of our experiments, we consider a dataset of 45 compounds. Our outcome of interest is a binary prediction of toxic or non-toxic, and of these 45 compounds, 29 are considered toxic and 16 are considered non-toxic. These 45 compounds are a subset of the 70 used in

prior work [9]. For our experiments here, we use the same compound concentrations and the same toxic/non-toxic binary labels assigned to each compound from this prior work (see Section V for more information about the data).

We collect transcriptome-wide gene expression profiles via RNA sequencing (RNA-Seq) from 2D tissue cultures following exposure to these compounds. We use the 45 compound subset of 3D tissue culture data from the prior work [9] as our comparison gene expression data for 3D tissue cultures. Throughout the paper, we refer to these datasets as the 2D and 3D dataset, respectively. Both datasets consist of gene expression measurements in transcripts per million reads (TPM) from 19,084 protein-coding genes for every sample. Each sample thus represents a gene expression profile from either a 2D or 3D tissue sample, a single compound, and a single length of exposure to that compound.

To collect gene expression data for our 2D experiments, we seeded neural progenitor cells (NPC) on Matrigel coated plates and started compound treatment on the same day (day 0). We collected samples at 11 time points of exposure: 1, 2, 3, 4, 6, 8, 10, 15, 21, 27, and 39 days. We then performed RNA-Seq and calculated gene expression values in transcripts per million reads. The dataset contains at most one biological sample for each compound and exposure length. There are missing samples for some exposure lengths of some compounds because of cell death and other experimental factors (see Table I). For more information about our cell culture approach, our sequencing pipeline, and how to access the data and code, see Section V.

The 3D dataset contains samples at two time points of exposure: 2 and 7 days. There are two biological samples for each compound at each time point, with no missing samples.

Overall, the goal of our prediction experiments is then to learn a model that can map from gene expression profiles (19,084 features) to a binary label of toxic or non-toxic. We make no prior assumptions about the common or unique biological effects of the compounds in the expression data. Thus, we make no attempt to separate or explicitly model different types of toxicity or non-toxicity that each individual compound may elicit. Similarly, we make no attempt to explicitly model the effects of different compound exposure lengths and do not include exposure length as an input feature. Instead, we simply allow the machine learning algorithms to observe samples from many different compounds at different exposure lengths and find patterns across the gene expression profiles that are associated with toxicity or non-toxicity. Using this approach, we aim to develop an accurate predictive model that generalizes beyond any single exposure length or type of toxicity. The primary advantage of this approach is thus its generalizability, but the disadvantage is that analyzing the model to understand any one particular toxic effect becomes more difficult because all individual signals are effectively combined.

We first describe and perform three sets of prediction experiments, performing each separately on the 2D and 3D datasets, and compare results. The first experiment evaluates the ability to predict toxicity by training and testing common machine learning models on samples from a single time point of compound exposure. Next, we evaluate the ability to predict toxicity by training and testing models using all available time points of compound

exposure pooled together. Finally, we evaluate the same models once more using all time points of exposure but with the addition of feature selection.

A. Classification experiments

We used four common machine learning algorithms in our classification experiments: support vector machines (SVMs), logistic regression, random forests, and naive Bayes. Traditional support vector machines (SVM) are binary classifiers that construct a maximum-margin decision boundary separating the two classes of training samples [10]. Logistic regression is a binary classification algorithm that models the posterior probability of the response variable as a logistic function applied to a linear combination of the predictor variables and model coefficients [11]. Random forest classifiers produce class predictions by aggregating over an ensemble of decision trees [12], each of which is fit using a bootstrap sample of the training dataset. Naive Bayes is a probabilistic graphical model that makes the strong simplifying assumption that all predictor variables are conditionally independent of one another given the class label [13].

We used the Scikit-learn [14] (v0.18.2) implementations of these algorithms, specifically the SVC, LogisticRegression, RandomForestClassifier, and MultinomialNB classes, respectively. For SVMs, we used a linear kernel, probability estimates, and scaled gene expression values to [0.0, 1.0]. For logistic regression, we standardized the data, used L2 regularization, and used the dual formulation. For both SVMs and logistic regression, we used internal cross-validation to select the C parameter from {0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0}. For random forests, we used the entropy splitting criterion and 100 trees. We used the default settings for all other algorithm parameters.

To evaluate the predictive performance of these algorithms, we used the receiver operating characteristic (ROC) curve and the area under the curve (AUC). We used this standard metric to get an overall sense of predictive performance without having to choose a single classification threshold for each model. An AUC of 1.0 represents a perfect ordering of toxic and non-toxic compounds, whereas an AUC of 0.5 represents random guessing.

We used a standard leave-one-compound-out cross-validation for all of our experiments to avoid overly optimistic estimates of future predictive performance. This means that we performed each experiment in 45 steps, corresponding with the 45 compounds in our dataset. For each step, a single compound was held out of the training set, a model was trained on the remaining 44 compounds, and the model was used to make a prediction for the held out compound. We then aggregated the predictions across all 45 compounds to make the final ROC curve evaluation. Note that in experiments where we pooled multiple samples for each compound (e.g. samples from different exposure lengths), we held out all samples for the held out compound at once and averaged the predicted probabilities across samples to produce the final compound prediction.

For our first classification experiment, we trained and tested our models using only one compound exposure length at a time. The experiment shows there is variation in predictive performance over the extent of exposure lengths. We ran five replicates of each to account

for AUC variations between runs due to randomness in some of the algorithms. Results for the 2D and 3D datasets can be found in Figure 1.

For our second classification experiment, we trained and tested our models by pooling samples from all exposure lengths together into one dataset. Recall that we do not explicitly model the different exposure lengths, instead opting to simply include different exposure lengths as additional samples of the same compounds. Again, we computed probability estimates for each compound by averaging the individual sample estimates. As before, we ran five replicates of each. Figure 2 (top) shows a comparison of AUCs for each of the algorithms between the 2D and 3D tissue culture methods.

Because there are fewer exposure lengths in the 3D dataset, there is a chance that differences in accuracy between the 2D and 3D dataset are a result of having many more exposure lengths in the 2D dataset. To account for this difference, we ran two smaller pooled experiments in the 2D dataset combining only two exposure lengths each. The exposure lengths in the 3D dataset are two days and seven days, but we do not have a seven day exposure length in our 2D dataset. Six and eight days of exposure are the closest analogous lengths in the 2D dataset. Thus, we ran these two day experiments in two ways, once by pooling days two and six, and another by pooling days two and eight. Again, we ran five replicates of each. Figure 2 (bottom) shows the results.

B. Feature selection experiments

We next trained the same models using all days pooled, but also applied feature selection. We ran experiments on 19 different fixed sizes of selected gene sets: 1,000 genes down through 100 with a step size of 100, and through 10 genes with a step size of 10. We chose these gene set sizes to demonstrate a wide range of model performance from large triple digit feature sets all the way down to small double digit sets. We used the same feature selection algorithms for the 2D and 3D datasets, and all algorithms used the same set of selected features for their respective datasets. We performed feature selection for each fold of cross-validation with the same data used for training the predictive models to avoid overly optimistic estimates of AUC.

We used three algorithms for our feature selection experiments: recursive feature elimination, filtering by mutual information, and sparse logistic regression. Recursive feature elimination works by recursively training a linear model and eliminating the least important features in each step, until some stopping condition is met [15] (e.g. a desired feature set size). Mutual information feature selection ranks all features by their mutual information with the class label and then filters to a specified feature set size [16]. Sparse logistic regression works by applying L1, rather than L2, regularization to the model, driving many feature coefficients to 0 and thus leading to a smaller feature set [17].

Again, we used Scikit-learn [14] implementations for these algorithms, specifically the RFE, SelectKBest, and LogisticRegression classes, respectively. For recursive feature elimination, we scaled expression values to [0.0, 1.0], used a linear SVM with C set to 1.0, and used a step size of 1%. For sparse logistic regression selection, we standardized the data, used L1 regularization, and set C to 1.0. We used the default settings for all other parameters.

Note that the RFE and mutual information methods easily lend themselves to selection of exact numbers of genes, whereas spare logistic regression does not. Thus, for sparse logistic regression, we first ran the sparse model and then selected the top K genes by the magnitude of their learned coefficients. We skipped experiments with the sparse logistic regression method when K was larger than the number of nonzero learned coefficients. As before, we ran five replicates of each experiment to account for AUC variations between runs. Figures 3 and 4 show the results from these experiments on the 2D and 3D datasets, respectively.

Because we performed feature selection for each fold of leave-one-compound-out cross-validation, each feature selection method was effectively performed 45 times for each gene set size. To assess the variability of each feature selection method on these datasets, we report the number of unique genes selected across all folds combined for each method at a gene set size of 10 (see Table II). Each feature selection method may thus pick a maximum of 450 genes, and a minimum of 10, across all folds.

III. Results and Discussion

Figure 1 shows the single day-by-day prediction results for the 2D and 3D datasets. Each point along the x-axis shows the predictive performance of a model that is trained and tested on samples at a single length of compound exposure in days. Error bars around the points show a 95% confidence interval from the five replicate runs of each algorithm and exposure length. Recall that 1.0 is a perfect AUC, while 0.5 is equivalent to random guessing. one commonality between the two culture methods is that naive Bayes does not appear to perform as well as the other three algorithms. This is perhaps unsurprising due to the high dimensionality of the dataset, which skews naive Bayes' predictions more toward 0.0 and 1.0. Predictive performance appears to improve overall at longer exposure lengths in the 2D dataset. This same trend is not immediately apparent in the 3D dataset, but trends are difficult to support with only two exposure lengths. The overall result, however, indicates that we can exceed the predictive performance of the 3D tissue model using the simpler 2D tissue model.

Figure 2 (top) shows a side-by-side comparison of the models on the 2D and 3D datasets when all available exposure lengths are pooled for training and testing. Again, error bars around the points show a 95% confidence interval from the five replicate runs of each. All of the models have substantially better AUC on the 2D dataset than on the 3D dataset. Naive Bayes again does not appear to perform as well as the other algorithms, but has substantially better accuracy on the 2D dataset once all of the samples are pooled.

Recall that the 2D dataset has several more compound exposure lengths than the 3D dataset. Figure 2 (bottom) shows results on the 2D dataset when pooling only two days of compound exposure, which is analogous to the two available in the 3D dataset. Again, naive Bayes does not perform as well as the other algorithms. Still, in all cases, the 2D models perform better than the analogous experiments on the 3D dataset. This suggests that the accuracy of the models on the 2D dataset is not simply a result of having a greater number of samples from a larger range of compound exposure lengths.

Finally, Figures 3 and 4 show results from the same pooled models with the addition of three feature selection methods. The horizontal axis gives the AUC at 19 decreasing gene set sizes, from 1,000 genes down to 10 genes. The results show a common trend of accuracy degradation with smaller gene set sizes, but the trend is more pronounced in the 3D dataset. For example, no method has an AUC above 0.8 in the 3D dataset at a gene set size of 10, whereas all but one method has an AUC above 0.9 in the 2D dataset.

In addition to better prediction, Table II shows that all of the feature selection methods generally selected more consistent sets of genes across folds on the 2D dataset, suggesting that the predictive signal may be less distributed in the 2D dataset than in the 3D dataset. Note that there is no overlap between the 2D and 3D selected gene sets for any of the three methods, suggesting that the predictive signal is also distinct between 2D and 3D.

Overall, the 2D tissue model appears to produce more accurate and more consistent predictive models. Certainly some difference between the two could result from differences in sample preparation and the sequencing depth achieved for each, but we think two more obvious potential explanations stand out. First, the 2D tissue model, being composed of only one cell type, is less complex than the 3D model and thus produces a less variable signal. Second, compound diffusion is likely more complete in the 2D tissue model than in 3D. Both of these potential explanations could lead to stronger signals of gene expression perturbation for the machine learning algorithms to detect.

IV. CONCLUSIONS

Here, we present common machine learning models applied to the task of predicting developmental neurotoxicity of several compounds from gene expression data. We compare the AUCs of these models between datasets collected from a 2D tissue culture approach, and a more complex 3D tissue culture approach. We compare results from training models on single lengths of compound exposure, from multiple pooled lengths of exposure, and with the addition of feature selection. Overall, our results show that the models trained on data collected from a simpler 2D tissue model are more accurate than those trained on data from a 3D model. While a 3D tissue model is perhaps more likely to recapitulate relevant biology needed to fully understand toxicity mechanisms, a 2D tissue model is certainly a viable option and easier to produce efficiently [18]. We would thus still recommend a 3D tissue model if the primary goal is to study biological mechanisms in depth, but our results here suggest that the 2D tissue model is an excellent choice for producing a broad toxicity assay.

Furthermore, our results show that models trained on the 2D data experience very little degradation in AUC under stringent feature selection, whereas the models trained on 3D data show extensive degradation. Our results also demonstrate that the genes selected were more consistent across folds on the 2D data than on the 3D data; this is important because it suggests that we may be able to simplify the model even further by reducing the number of genes that need to be quantified to perhaps far fewer than 100 without loss of accuracy. With a much smaller gene set, it may be possible to develop a similar assay using quantification methods that are still faster and cheaper than RNA-Seq. We propose this direction of research for future work.

We further propose evaluating the use of 2D tissue models made from simpler cell types than NPCs to predict developmental neurotoxicity or toxicity in general. NPCs require substantial experience to successfully differentiate and culture, whereas a tissue culture based on a cell type such as dermal fibroblasts may be more approachable. While such a cell type may not be neural in nature, and the pattern of response would surely be different, the cell may still exhibit gene expression perturbations indicative of toxicity sufficient for the purposes of an assay.

Current models for toxicity screening are simply too slow and expensive to comprehensively test all new or less understood chemical exposures. These models are certainly here to stay for the foreseeable future, but high-throughput screening methods, such as we have presented here, show a great deal of potential. We hope and expect that both of these approaches may complement one another and accelerate findings by helping stakeholders choose which exposures to explore further and with what urgency.

V. DATA COLLECTION AND AVAILABILITY

Supplementary material with detailed explanations of our cell culturing and RNA sequencing approaches can be found at https://morgridge.org/research/regenerative-biology/bioinformatics/publications. Our 2D tissue model sequencing data are available through GEO Series accession number GSE126786, and the 3D tissue model sequencing data we use for comparison are available through GEO Series accession number GSE63935. All code and processed expression data can be found on GitHub at https://github.com/finnkuusisto/DevTox2D.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

REFERENCES

[1]. Betts KS, "Growing knowledge: Using stem cells to study developmental neurotoxicity," Environmental Health Perspectives, vol. 118, no. 10, p. A432, 2010. [PubMed: 20884390]

[2]. Rice D and Barone S Jr, "Critical periods of vulnerability for the developing nervous system: Evidence from humans and animal models," Environmental Health Perspectives, vol. 108, no. Suppl 3, p. 511, 2000. [PubMed: 10852851]

[3]. Grandjean P and Landrigan PJ, "Neurobehavioural effects of developmental toxicity," The Lancet Neurology, vol. 13, no. 3, pp. 330–338, 2014. [PubMed: 24556010]

[4]. Trasande L and Liu Y, "Reducing the staggering costs of environmental disease in children, estimated at $76.6 billion in 2008," Health Affairs, vol. 30, no. 5, pp. 863–870, 2011. [PubMed: 21543421]

[5]. Judson R et al., "In vitro and modelling approaches to risk assessment from the us environmental protection agency toxcast programme," Basic & Clinical Pharmacology & Toxicology, vol. 115, no. 1, pp. 69–76, 2014. [PubMed: 24684691]

[6]. Fabre KM, Livingston C, and Tagle DA, "Organs-on-chips (microphysiological systems): Tools to expedite efficacy and toxicity testing in human tissue," Experimental Biology and Medicine, vol. 239, no. 9, pp. 1073–1077, 2014. [PubMed: 24962171]

[7]. Olson H et al., "Concordance of the toxicity of pharmaceuticals in humans and in animals," Regulatory Toxicology and Pharmacology, vol. 32, no. 1, pp. 56–67, 2000. [PubMed: 11029269]

[8]. Hay M, Thomas DW, Craighead JL, Economides C, and Rosenthal J, "Clinical development success rates for investigational drugs," Nature Biotechnology, vol. 32, no. 1, pp. 40–51, 2014.

[9]. Schwartz MP et al., "Human pluripotent stem cell-derived neural constructs for predicting neural toxicity," Proceedings of the National Academy of Sciences, vol. 112, no. 40, pp. 12516–12521, 2015.

[10]. Cortes C and Vapnik V, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[11]. Cox DR, "The regression analysis of binary sequences," Journal of the Royal Statistical Society. Series B (Methodological), vol. 20, no. 2, pp. 215–242, 1958 [Online]. Available: http://www.jstor.org/stable/2983890

[12]. Breiman L, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[13]. Mitchell T, Machine Learning. New York: McGraw-Hill, 1997.

[14]. Pedregosa F et al., "Scikit-learn: Machine learning in python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[15]. Guyon I, Weston J, Barnhill S, and Vapnik V, "Gene selection for cancer classification using support vector machines," Machine learning, vol. 46, no. 1-3, pp. 389–422, 2002.

[16]. Yang Y and Pedersen JO, "A comparative study on feature selection in text categorization," in ICML, vol. 97, 1997, pp. 412–420.

[17]. Friedman J, Hastie T, and Tibshirani R, The elements of statistical learning. New York: Springer series in statistics, 2001.

[18]. Chandrasekaran A et al., "Comparison of 2d and 3d neural induction methods for the generation of neural progenitor cells from human induced pluripotent stem cells," Stem Cell Research, vol. 25, pp. 139–151, 2017. [PubMed: 29128818]
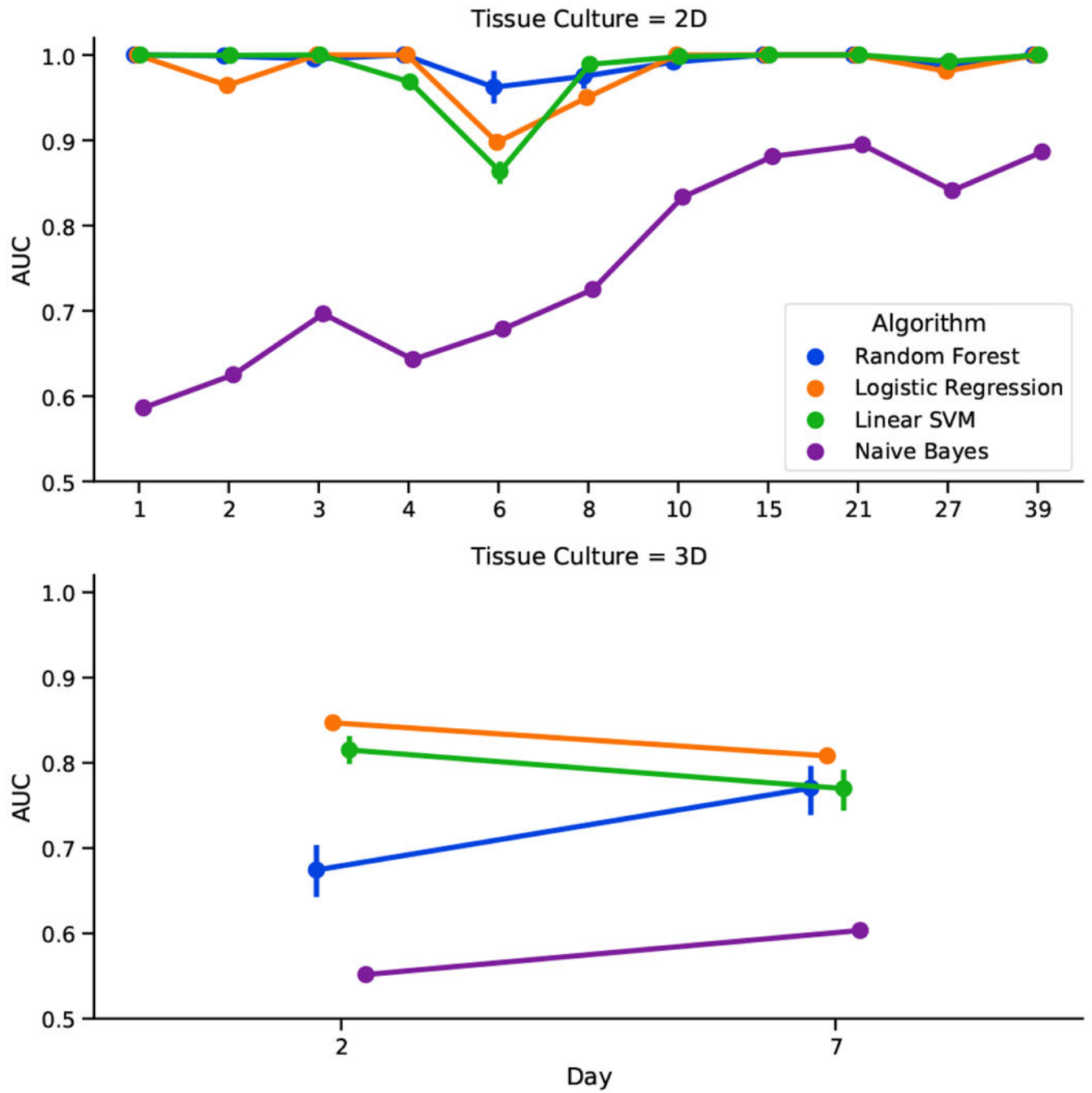
Fig. 1.
AUCs for single day train and test on the 2D (top) and 3D (bottom) tissue culture datasets.
Error bars around points give a 95% confidence interval from five replicate runs of each.
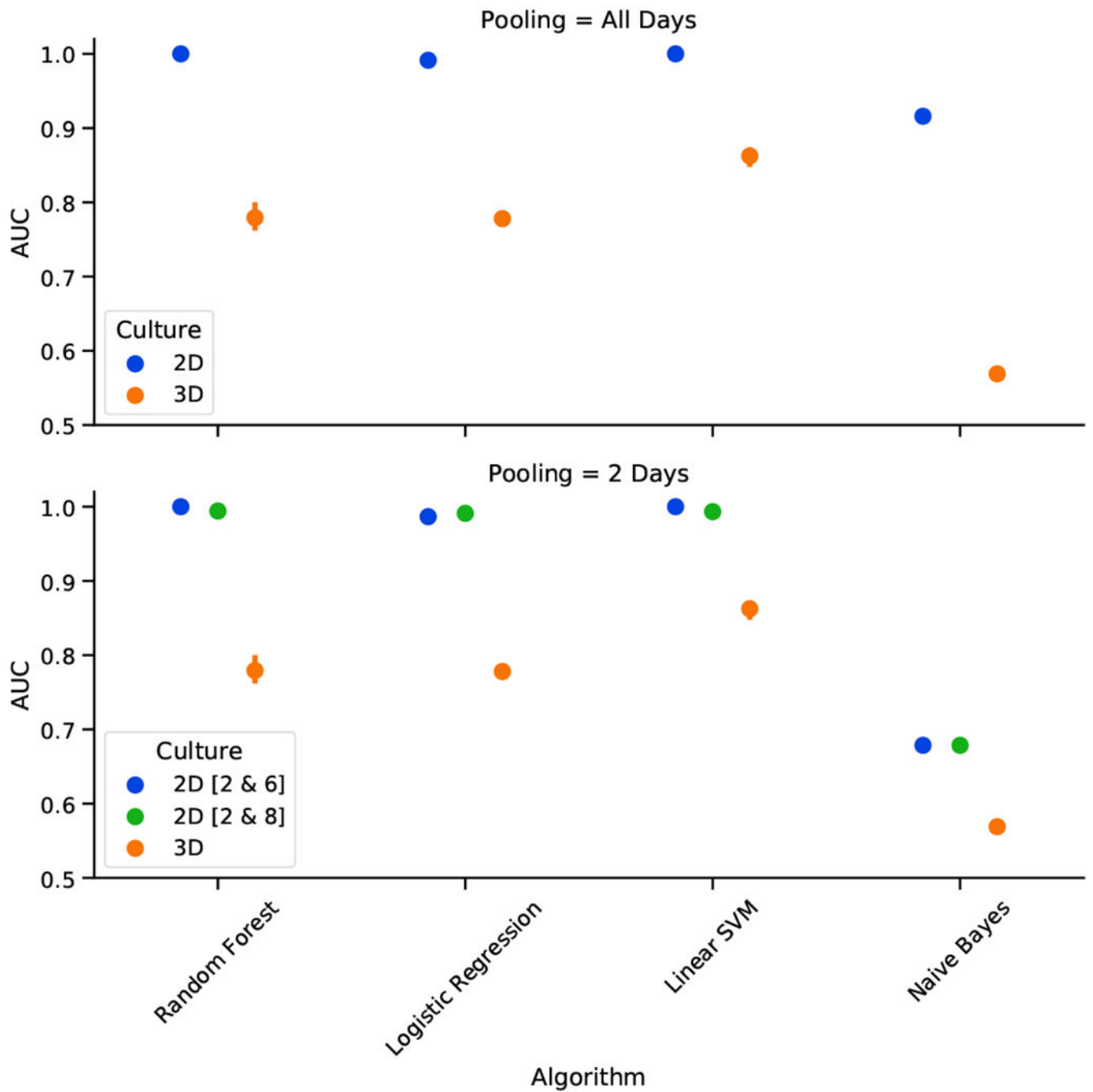
Fig. 2.
Comparison of AUCs between 2D and 3D tissue culture methods when all available
exposure lengths are pooled for train and test (top), and when only two days of exposure are
pooled for the 2D method (bottom). The 3D results are the same in both. Error bars around
points give a 95% confidence interval from five replicate runs of each experiment.
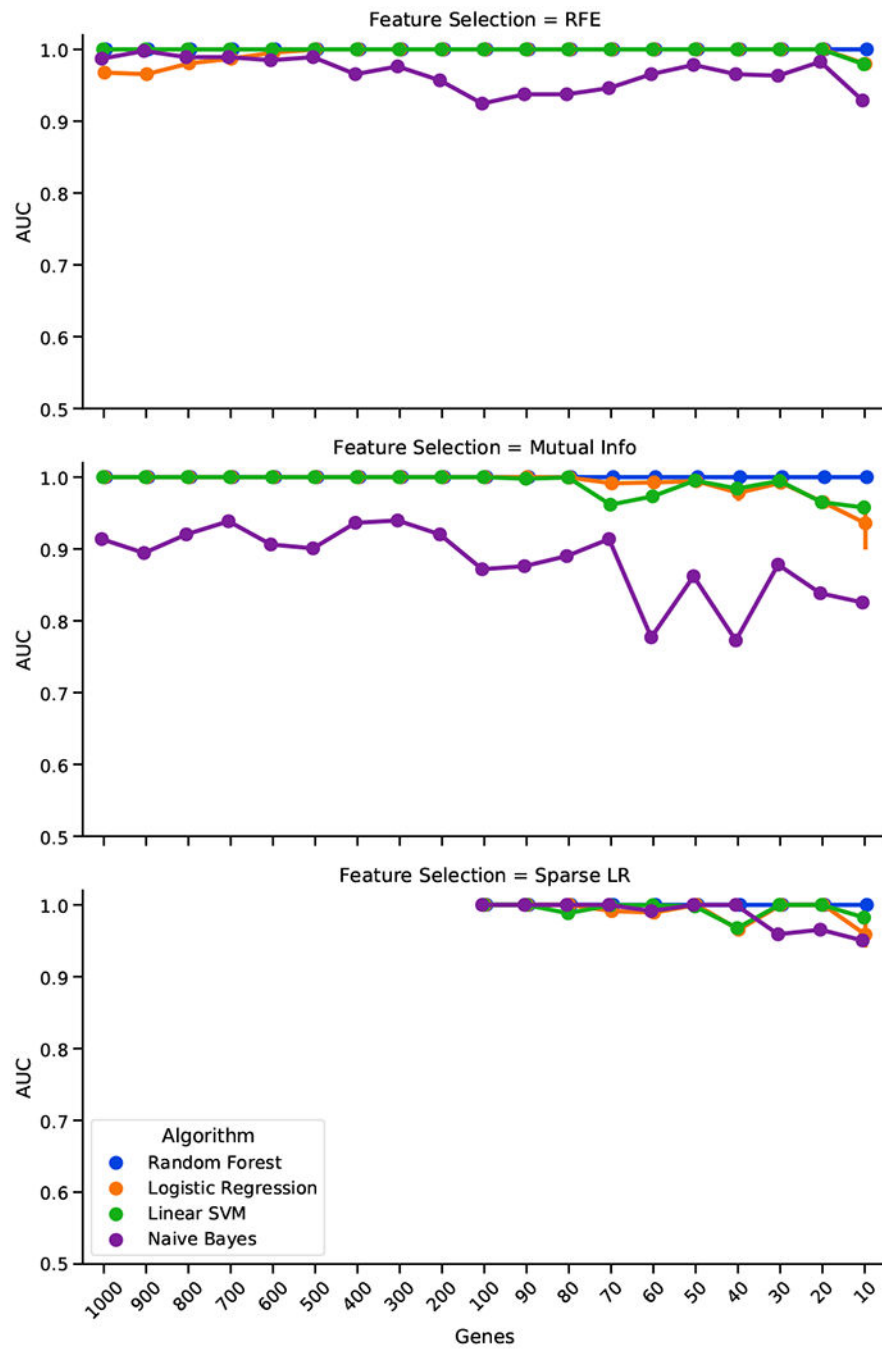
Fig. 3.
AUCs for models trained and tested with feature selection on all exposure lengths pooled for the 2D dataset. The horizontal axis is the selected gene set size. Error bars give a 95% confidence interval from five replicate runs each.

Fig. 4.
AUCs for models trained and tested with feature selection on all exposure lengths pooled for the 3D dataset. The horizontal axis is the selected gene set size. Error bars give a 95% confidence interval from five replicate runs each.
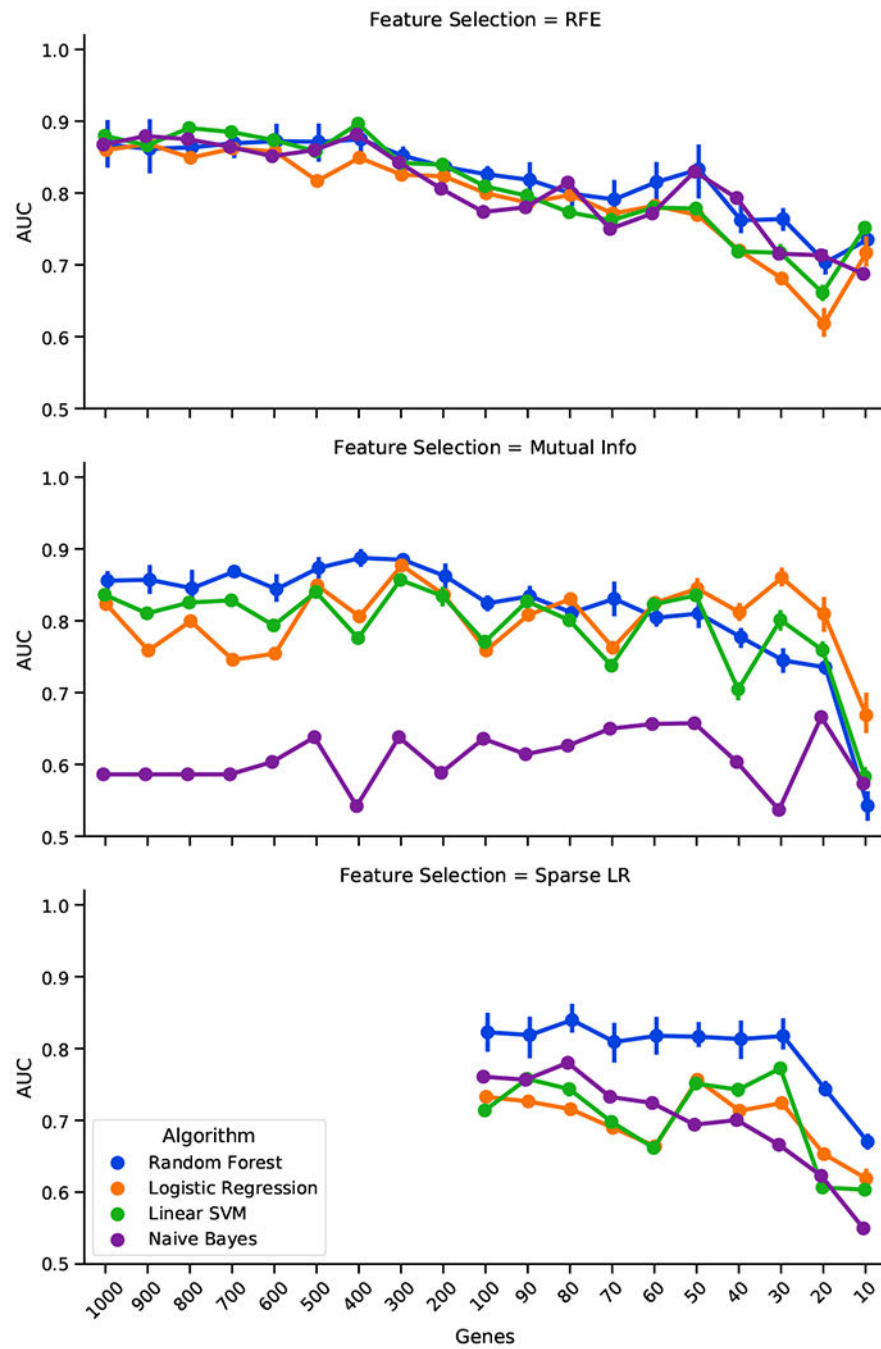
TABLE I

MISSING SAMPLES FROM THE 2D TISSUE CULTURE DATASET.

	Chemical	**Missing Days**
Toxic	Arsenic	8 - 39
	Busulfan	8 - 39
	Cadmium	8 - 39
	Cytosine β-D-arabinofuranoside	8 - 39
	5-Fluorouracil	8 - 21
	2-Imidazolidinethione	21
	Maneb	2 - 39
	Okadaic Acid	8 - 39
	PD166866	8
	U0126	21
	Vincristine	8 - 39
Non-Toxic	Acetaminophen	8, 15
	Aspirin	8 - 39
	Glucosamine	15 - 39
	Glycerol	8 - 27
	Ibuprofen	15
	Naproxen sodium	6
	PEG 3350	15
	Glyphosate	8, 15
	Sorbitol	8 - 21
	Saccharin	8 - 27

TABLE II

Cross-validation unions of selected genes with K of 10.

Culture	RFE	Mutual Info	Sparse LR
2D	99	25	37
3D	153	81	41