SYSTEMATIC ASSESSMENT OF ANALYTICAL METHODS FOR DRUG SENSITIVITY PREDICTION FROM CANCER CELL LINE DATA*

IN SOCK JANG[1], ELIAS CHAIBUB NETO, JUSTIN GUINNEY, STEPHEN H. FRIEND, ADAM A. MARGOLIN[1]
Sage Bionetworks
1100 Fairview Ave. N Seattle, WA 98109, USA
Email: in.sock.jang@sagebase.org
Email: elias.chaibub.neto@sagebase.org
Email: justin.guinney@sagebase.org
Email: friend@sagebase.org
Email: margolin@sagebase.org

Large-scale pharmacogenomic screens of cancer cell lines have emerged as an attractive pre-clinical system for identifying tumor genetic subtypes with selective sensitivity to targeted therapeutic strategies. Application of modern machine learning approaches to pharmacogenomic datasets have demonstrated the ability to infer genomic predictors of compound sensitivity. Such modeling approaches entail many analytical design choices; however, a systematic study evaluating the relative performance attributable to each design choice is not yet available. In this work, we evaluated over 110,000 different models, based on a multifactorial experimental design testing systematic combinations of modeling factors within several categories of modeling choices, including: type of algorithm, type of molecular feature data, compound being predicted, method of summarizing compound sensitivity values, and whether predictions are based on discretized or continuous response values. Our results suggest that model input data (type of molecular features and choice of compound) are the primary factors explaining model performance, followed by choice of algorithm. Our results also provide a statistically principled set of recommended modeling guidelines, including: using elastic net or ridge regression with input features from all genomic profiling platforms, most importantly, gene expression features, to predict continuous-valued sensitivity scores summarized using the area under the dose response curve, with pathway targeted compounds most likely to yield the most accurate predictors. In addition, our study provides a publicly available resource of all modeling results, an open source code base, and experimental design for researchers throughout the community to build on our results and assess novel methodologies or applications in related predictive modeling problems.

Keywords: Cancer cell lines, pharmacogenomics, machine learning, predictive modeling.

1. Introduction

Molecular analysis of cancer has revealed that tumor subtypes differ in pathway activity, progression, and chemotherapeutic response, leading to the development of therapeutic approaches with demonstrated efficacy in molecularly defined cancer subtypes [1-4]. Human cancer cell lines represent an attractive pre-clinical system for identifying molecular characteristics of tumors predictive of therapeutic response.

Recently, two ambitious initiatives, named the cancer cell line encyclopedia [5, 6] and the genomics of drug sensitivity projects [7] have performed large-scale small molecule screens on

* Work supported by grant U54CA149237 from the Integrative Cancer Biology Program of the National Cancer Institute.

[1] corresponding authors

panels of hundreds of molecularly characterized cancer cell lines. Both studies also demonstrated that employing modern machine learning algorithms to develop predictors of drug response based on molecular profiling measurements of each tumor could effectively identify known pharmacogenomic predictive biomarkers. These proof-of-concept studies have established cell line-based screens as a viable pre-clinical system for identifying functional biomarkers underlying drug sensitivity or resistance and for suggesting patient selection strategies for clinical trial design.

As computational approaches for modeling therapeutic response become increasingly common in research and translational applications, a study is warranted to systematically assess different modeling approaches, and recommend best practices for future applications. To address this question, we defined important categories of modeling choices, such as the predictive algorithm and genomic features for model inclusion (among others), and performed a large multifactorial experiment with crossed factors, where the modeling choices represent the experimental factors, and the predictive performance measures (derived from model fits, and spanning all possible combinations of modeling choices) represent the response data. This experimental design allows for formal statistical testing and quantification of the relative importance of the modeling choices.

Our results provide statistically principled, data-driven guidelines for best-in-class modeling practices. Our findings suggest the use of elastic net or ridge regression applied to continuous valued response data, summarized using the area under the fitted dose response curve, and using all molecular features (in particular, gene expression data). Moreover, our results suggest that pathway targeted compounds lead to more accurate predictors than classical broadly cytotoxic chemotherapies. In addition, we performed detailed analysis comparing models based on continuous versus discretized response measurements, suggesting that discretizing data (e.g. into sensitive and resistant calls) causes decreased model accuracy. Finally, we report a discordance in reported values across the 2 datasets for the same compounds and suggest that raw dose-response data should be made publicly available to facilitate comparison of the 2 datasets based on the same procedures for processing and summarizing dose-response values.

Our study provides a publicly available interactive resource of modeling results and an open source analysis package. The results for all >110,000 models are available at (https://www.synapse.org/#!Synapse:syn2009053), providing a resource for other researchers to interactively browse the results of all models and perform additional downstream analyses. Moreover, we are releasing the open source "predictiveModeling" R package (https://github.com/Sage-Bionetworks/PredictiveModel_pipeline and https://github.com/Sage-Bionetworks/predictiveModeling), containing all code used to infer models in this study, and providing a modular API that may be extended by the community and used to conduct similar research studies.

2. Material and Methods

2.1 Data Sets

The CCLE and Sanger datasets contain compound screening data performed on large panels of molecularly characterized cancer cell lines. Both datasets contain genome-wide gene expression and copy number profiling, as well as sequencing data on a subset of genes (described in the next section). Gene expression, copy number, and mutation data were summarized to gene-level features. The Sanger panel is composed of 30,672 genomic features and 138 compounds profiled on 714 cell lines (535 cell lines contain all measurement types). The CCLE panel is composed of 41,814 genomic features and 24 compounds profiled on 504 cell lines (411 cell lines contain all measurement types). All data was normalized as described in the original papers [5-9]. Mutation data was summarized to binary gene-level variables represented as 0 (wild type) and 1 (mutation). We also annotated each cell line with a representative "tumor type" label, derived by manually curating the provided meta-data annotations. Each tumor type was then included as a binary feature variable.

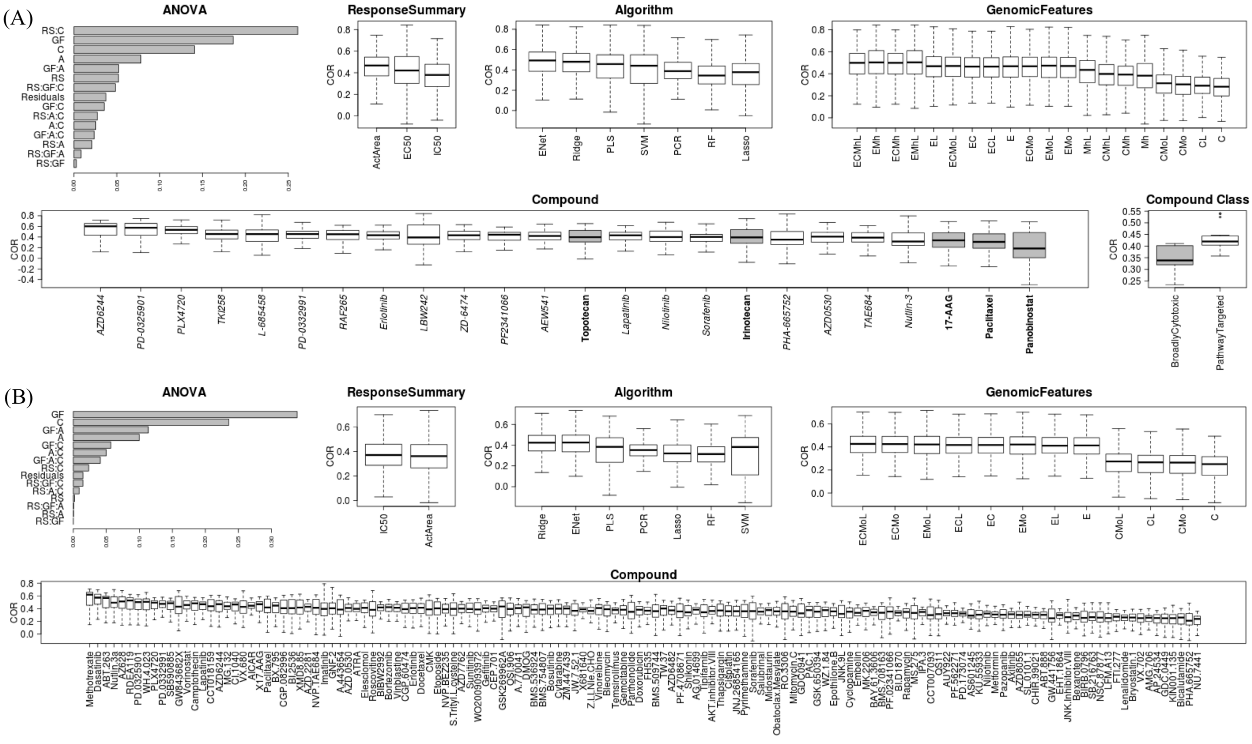

Figure 1 – Summary of evaluation of regression models. (A) Results for CCLE. (B) Results for Sanger. The left panel displays the percent variance of predictive accuracy (COR) explained by each category of modeling choice after fitting our 4-way ANOVA model. The panels labeled *Compound*, *ResponseSummary*, *Algorithm*, and *GenomicFeatures* correspond to each of our tested categories of modeling choices, and display the distribution of predictive performance (COR) scores for each modeling choice (factor levels) within the category. For the CCLE *Compound* panel, compounds classified as "BroadlyCytotoxic" are displayed as shaded boxes and bold text, and compounds classified as "PathwayTargeted" are displayed as white boxes and non-bold text. The panel titled *Compound Class* displays the distribution of predictive performance scores for the BroadlyCytoxic vs. PathwayTargeted compound classes.

Both studies provide multiple statistics used to summarize dose-response curves to compound sensitivity values for each cell line (described in the next section). We used the summarized sensitivity values reported in each dataset, as raw dose-response values were not available to process both datasets using the same procedures.

2.2 Definition of modeling choices

Our goal was to systematically assess the effect of modeling choices on predictive performance given a *drug response vector* and a *molecular feature matrix*. We enumerated the following 5 categories of modeling choices, as well as the possible choices of modeling factors within each category

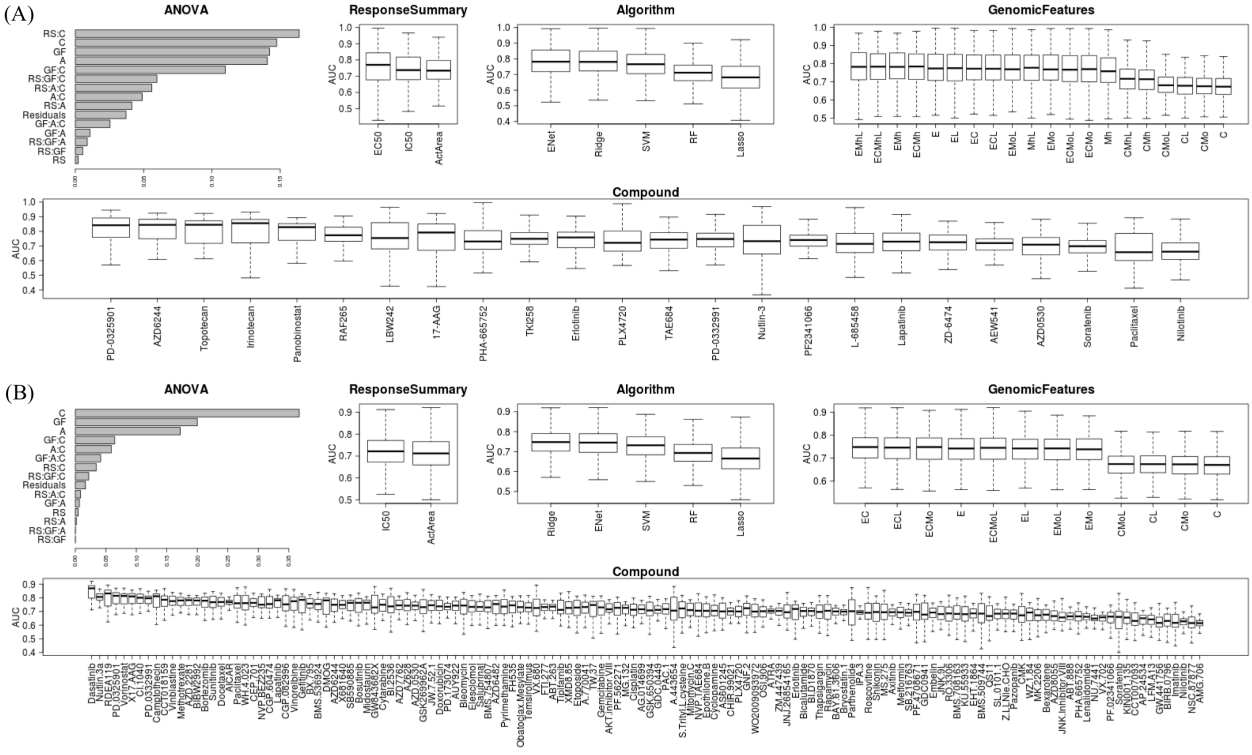

Figure 2 – Summary of evaluation of classification methods. (A) Results for CCLE. (B) Results for Sanger. Results are presented as described in **Figure 1**, based on evaluation of classification models using the AUC predictive performance statistic.

GenomicFeatures: Represent the distinct data types used as features in the predictive algorithms. In Sanger we have 4 distinct types: gene expression measurements (E) on 12,024 genes; copy number variation measurements (C) on 18,601 genes; cell line tumor type classifications (L) according to 93 distinct tumor lineages; and mutation profiling (Mo) on 47 genes. We tested 12 distinct data type combinations as shown in the *GenomicFeatures* panels in Figure 1B and Figure 2B (specifically, we tested all combinations other than those corresponding to small feature sets, such as L+Mo). For the CCLE panel we have 5 distinct data types: gene expression measurements (E) on 18,897 genes; copy number measurements (C) on 21,217 genes; cell line tumor type classifications (L) of 97 tumor lineages; mutation profiling (Mo) on 33 genes using the oncomap 3.0 platform [10]; and mutation profiling of 1,667 genes using hybrid capture sequencing (Mh). We tested 20 distinct data type combinations shown in the *GenomicFeatures* panels in Figure 1A and Figure 2A.

Compound: Represents the anti-cancer compounds screened by the cell line projects. There are 138 compounds in Sanger and 24 in CCLE.

ResponseSummary: Represents the statistic used to summarize the dose response curves to a single number, corresponding to the degree of sensitivity of a given cell line to a given compound. For Sanger, the choices are: AUC – the area under the fitted dose response curve; IC50 – the concentration at which the compound reaches 50% reduction in cell viability. For CCLE, the choices are: ActArea – the area above the fitted dose response curve (inverse measure of AUC in Sanger); IC50 – the same as in Sanger; EC50 – the concentration at which the compound reaches 50% of its maximum reduction in cell viability. We note that although they use the same terminology, both studies used different procedures for fitting dose response curves and generating summary statistics.

Continuous vs. categorical models: Whether predictions are made based on continuous or discretized *ResponseSummary* measurements. We tested multiple discretization schemes, including: mean and median based deviation statistics; Gaussian mixture models; and upper/lower third quartile thresholds. We report results based on upper/lower third quartile thresholds, which was the discretization scheme that achieved the highest average classification accuracy (AUC).

Algorithm: Represents the predictive algorithms compared in this study. In the analysis of continuous response variables, we compared: principal component regression (PCR); partial least square regression (PLS); least squares support vector machine regression with linear kernels (SVM); random forests (RF); least absolute shrinkage and selection operator (LASSO); ridge regression (RIDGE); and elastic net regression (ENet) [11-19, 27]. For the analysis of binary response variables, we considered: least squares support vector machine classification with linear kernels (SVM); random forests (RF); binomial least absolute shrinkage and selection operator (LASSO); ridge binomial regression (RIDGE); and elastic-net binomial regression (ENet) [8, 11, 12, 14, 15, 20].

2.3 Model fitting procedures

We employed a multifactorial experimental design and tested all combinations of modeling choices (e.g. the cross product of all choices of *ResponseSummary* × *Compound* × *GenomicFeatures* × *Algorithm* × *Discretization*, excluding application PCR and PLS to discrete data). This resulted in testing a total of 114,048 models.

For Sanger and CCLE the input dataset was divided into five non-overlapping sample groups, used as cross-validation folds for training and testing data. For each cross-validation fold, each model was trained on 4/5$^{\text{ths}}$ of the samples, and used to make predictions of sensitivity for the held out 1/5$^{\text{th}}$ of samples. Within each training step, a separate 5-fold cross-validation procedure was employed for parameter tuning of each model.

Predicted vs. observed response vectors were compared to assess the performance of each algorithm. The predicted response vector was computed by concatenating the prediction vectors for each cross-validation fold. For continuous models we computed the Pearson correlation coefficient (COR). For discrete models we computed area under the receiver operating characteristics curves (AUC).

2.4 Statistical Analysis

We evaluated the effect of modeling choices on predictive performance using multiway-ANOVA with crossed factors. For instance, in the analysis of continuous models in the CCLE panel, we adopted COR as the response variable, and performed ANOVA using 4 factors:

GenomicFeatures, composed of 20 levels representing distinct data type combinations; *Compound*, composed of 24 levels, each representing one of the anti-cancer compounds tested in the CCLE panel; *ResponseSummary*, represented by levels ActArea, EC50, and IC50; and *Algorithm* represented by levels ENet, RIDGE, PLS, SVM, PCR, LASSO, and RF. For each one of the possible 20 × 24 × 3 × 7 = 10,080 modeling choice combinations, we fit a predictive model and recorded the correlation between the observed and predicted outcome as the response variable. Since we only have a single observation per modeling choice combination, our design corresponds to a multiway-ANOVA with 4 crossed factors and a single observation per cell. Hence, we cannot fit a complete model (i.e., with all interaction terms up to order 4) and we restrict our analysis to interactions of order up to 3. In addition to the analysis described above, we also performed analogous ANOVA analyses for the evaluation of continuous models in Sanger, discrete models in CCLE, and discrete models in Sanger.

3. Results

Modeling factors influencing predictive performance

In order to assess the individual contributions of each category of modeling choices (and their interactions) to explaining the total variability of the predictive performance statistic (COR or AUC), we examined the decomposition of the total sum of squares of the predictive performance variable into residual sum of squares plus sum of squares terms for each one of the factors and factor interactions in our 4way-ANOVAs, including all possible interactions of order up to 3. We first describe results for continuous models. The left panels in Figure 1A and B present barplots in which each bar represents the sum of squares of the respective term divided by the total sum of squares.

For both the CCLE and Sanger datasets, most of the variance of predictive accuracy is explained by the modeling factors considered in our study, as indicated by the small percent of variance attributable to residuals. For both CCLE and Sanger the modeling factors explaining the highest percent variance are: 1) the type of molecular features used to build the model; and 2) the compound being predicted by the model. The third most important modeling factor is the type of algorithm, although this factor is considerably less important than the first 2. This result is consistent with previous studies [21], suggesting that input data is the dominant factor related to model performance, whereas the specific modeling strategies are of secondary importance.

The CCLE dataset contains a strong interaction term between *Compound* and *ResponseSummary*, suggesting that model performance depends both on the compound being modeled, and the ability to summarize the compound's dose response measurements. By contrast, *ResponseSummary* has negligible effect in the Sanger dataset. We point out that, although Sanger and CCLE both report response data in terms of IC50 and AUC (referred to as ActArea in CCLE) summarizations, the 2 studies use quite different procedures for fitting dose response curve and summarizing them to IC50 of AUC statistics. The discordant importance of the *ResponseSummary* factor between the 2 studies, compared with the highly concordant importance of all other factors, suggests that the procedures for summarizing dose response curves to summary statistics may be inconsistent between the 2 studies. Indeed, comparison of IC50 and AUC values for compounds profiled in both datasets suggests a relatively high degree of inconsistency (Figure 3). Unfortunately, raw dose response data used for curve fitting it not available in either study, limiting our ability to investigate this issue further. This result highlights the importance of making raw forms of data publicly available, in addition to computed summary statistics, such that the community may more transparently analyze and improve the value of the data resource.

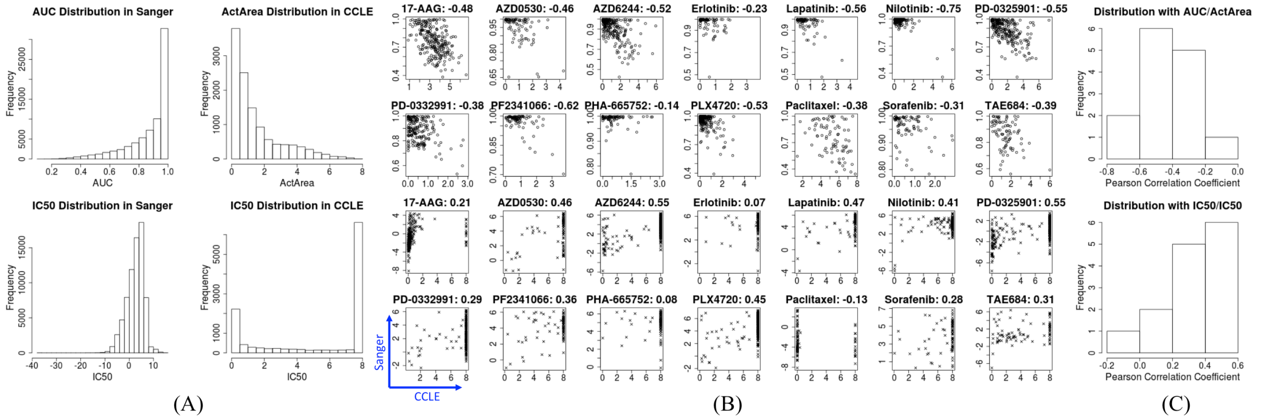

Figure 3 – Comparison of IC50 and AUC summary statistics for 14 compounds and 283 cell lines in common between the Sanger and CCLE datasets. (A) Distribution of IC50 and AUC/ActArea values in Sanger and CCLE. Note that the AUC value reported in Sanger corresponds to the area under the dose response curve in which values of 0 correspond to complete reduction in cell viability and values of 1 correspond to no reduction in cell viability. The ActArea value reported in CCLE corresponds to the area over the dose response curve in which values of -100 correspond to complete reduction in cell viability and values of 0 correspond to no reduction in cell viability. Therefore a negative correlation is expected between AUC and ActArea values. (B) Scatter plots comparing AUC/ActArea values (top) and IC50 values (bottom) across the 2 studies. (C) Histograms of the distribution of correlations across the 2 studies for the 14 common compounds based on ActArea/AUC (top) and IC50 (bottom).

*Assessment of best performing modeling strategie*s

The ANOVA analysis detected highly significant interaction and main effects in explaining predictive performance, indicating the importance of some modeling choices over others. Figure 1 and Figure 2 depict boxplot panels for each one of the modeling choice factors in our analyses, showing the distribution of predictive performance as a function of the modeling factor levels. For both datasets, expression data was the most informative molecular feature type, as all of the best performing models included use of expression data. Models using other molecular features types in addition to expression data performed slightly better than using expression data alone, although performance improvements were modest. For both datasets, elastic net and ridge regression were the top performing algorithms. For the CCLE dataset, summarizing dose response values based on ActArea achieved the highest performance. For Sanger, response summarization had little effect on model performance, warranting closer investigation starting from raw dose response data.

For both datasets, some compounds were easier to predict than others, as clearly shown by the *Compound* panels in Figure 1. Inspection of predictability scores for CCLE compounds suggested a general trend. Compounds with low predictability scores tended to be more classical chemotherapeutics that disrupt broad cellular processes (e.g. topoisomerase inhibitors). Compounds with high predictability scores tended to target proteins in specific pathways, primarily related to mitogen signaling (e.g. MEK inhibitors). To test this hypothesis, we manually annotated each compound in one of these 2 classes, which we called “BroadlyCytotoxic (BC)” and

“PathwayTargeted (PT)”. Indeed, PT compounds displayed significantly higher predictability scores compared to BC (*P=0.003529* by Wilcoxon rank sum test, as shown in the top-right panel of Figure 1).

Assessment of categorical models

An alternative strategy to modeling the drug response as a continuous-valued variable is to discretize the response vector into a binarized “sensitive” and “resistant” vector. To evaluate this strategy, we implemented the categorical analogues of lasso, ridge, elastic net, random forests, and support vector machines, and discretized each response summarization (IC50, EC50, AUC or ActArea) base on the upper and lower third quartiles.

Results from this analysis were highly consistent with results from our continuous models (Figure 2). For both CCLE and Sanger, the relative importance of model factors was consistent with results for continuous models (e.g. *GenomicFeatures* and *Compound* being most important, followed by *Algorithm*). The relative performance of modeling choices was also consistent between categorical and continuous models (e.g. the order of predictive performance of algorithms is fully consistent).

One advantage of categorical models is the ability to interpret AUC values as the probability of correctly classifying a new sample as sensitive or resistant. For example, analysis of the distribution of AUC scores suggests that sensitive vs. resistant samples can be classified with >70% accuracy for 22 of 24 (91.7%) compounds in CCLE and 83 of 138 (60.1%) compounds in Sanger. More specific analysis of the AUC curves can be used to determine the expected trade-offs between false positives and false negatives. We suggest that such analysis may be useful in assessing the potential clinical utility of a predictive model, for example, by applying criteria such as requiring less than a 5% false positive rate (e.g. correctly prescribing a drug to 95% of patients who might benefit) at the expense of a less than 20% false negative rate (e.g. failing to prescribe the drug to 20% of the patients who will benefit from it). Of course, such statistics derived from cell line studies are unlikely to directly translate in a clinical context, but may be useful to identify predictive models that should be prioritized for further clinical studies.

Comparison of continuous vs. categorical models

In order to directly compare the performance of continuous vs. categorical models, we computed the AUC scores of the rank-ordered predictions in comparison to the discretized response data. That is, we calculated the sensitivity and specificity at each threshold of the rank-ordered predictions in order to compute an ROC curve for each model. We based our comparison on the best performing regression and classification methods, which was elastic net in both cases (results were similar for other methods). In general, regression models, trained using continuous *ResponseSummary* values, outperformed classification models, trained using discretized *ResponseSummary* values (P<< 2.2e-16 for Sanger, based on AUC; P<< 2.2e-16 for Sanger, based on IC50; P<< 2.2e-16 for CCLE, based on ActArea; P=0.1587 for CCLE, based on IC50. See Figure 4. Classification methods outperformed regression methods only when using the CCLE IC50 values, as explained by the fact that these values are inherently discretized. Sanger IC50 values utilized extrapolations of the curve fits beyond the tested concentration range. By contrast, out of 11,670 IC50 values reported in CCLE (426 excluding NA values), 6,499 (55.69%) were set to a value of 8, corresponding to the maximum tested compound dose of 8μM (Figure 3A).

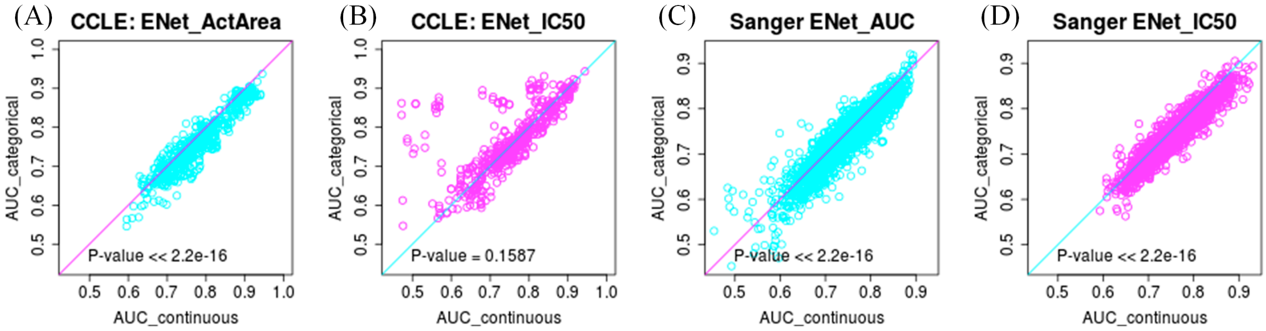

Figure 4 – Comparison of predictive performance of continuous (regression) vs. categorical (classification) models. Results were compared for the continuous and categorical versions of elastic net, which were the best performing continuous and categorical models. (A) CCLE data with ActArea, (B) CCLE data with IC50, (C) Sanger data with AUC, and (D) Sanger data with IC50.

4. Discussion

As large-scale complex genomic resources become increasingly available, there is a pressing need to develop community standards and robust assessment methods to determine the best performing approaches for analyzing such data. Pharmacogenomic screens performed on genomically characterized cancer cell lines provide rich data resources, and application of machine learning methodologies to such data have demonstrated evidence of uncovering genomic mechanisms underlying drug response.

From an analytical perspective, such pharmacogenomic data resources are particularly well suited to application of statistical learning methods by representing genomic and compound sensitivity data, respectively, as predictive features and response variables in a supervised learning scheme. In this study, we performed a controlled analysis of many modeling choices that may be used in this application. We believe this work contributes to the community in 3 ways: 1) by providing a set of recommended best practices for inferring pharmacogenomic predictive models, and a study on the relative importance of each; 2) by establishing a resource of over 110,000 modeling results, providing a baseline set of scores that researchers may use in future studies to demonstrate improved performance of novel methodologies; 3) by providing an experimental design template, and open source modeling package, that can be extended for use in other predictive modeling applications.

Our study suggests a statistically principled set of recommended best modeling practices: using **elastic net or ridge regression** with input features from all genomic profiling platforms, most importantly, **gene expression features**, to predict **continuous-valued** sensitivity scores summarized using the **area under/over the dose response curve**, with **pathway targeted** compounds will most likely yield the most accurate predictors.

The use of elastic net regression is consistent with modeling choices reported in previous studies [14][16], and is a particularly attractive option due to the ability to perform feature selection based on inferred feature weights. We investigated several methods that have previously been shown to achieve superior predictive accuracy, but lead to less interpretable models, such as support vector machines, random forests, and principal components regression [22, 23].

Nonetheless elastic net regression achieves the highest predictive accuracy without requiring a trade-off of model interpretability. Moreover, elastic net is designed to seek the optimal trade-off of model complexity penalties imposed by lasso and ridge regression. While the sparse feature selection encouraged by lasso indeed leads to inferior predictive performance, elastic net performs as well as ridge regression based on predictive accuracy, suggesting that elastic net effectively balances the strengths of the two methods by encouraging sparser models without compromising predictive accuracy. We note that although we employed standard and well-accepted cross-validation schemes for parameter tuning of all models, it is possible that alternative methods could improve the performance of some models.

The observation that gene expression features provide the most informative predictors might be explained by the increased "information content" of gene expression data. In particular, copy number values are highly correlated with each other and the mutation data profiles only a small subset of genes. Although gene expression data provides advantages in predictive accuracy, genomic (e.g. somatic mutation and copy number) data possess advantages in potential translation to clinical biomarkers. From a technical standpoint, the increased molecular stability of DNA compared to RNA facilitates easier development of clinical assays, even from archival samples. Perhaps more importantly, features derived from genomic data are more likely to correspond to functional driver events related to drug sensitivity, whereas features derived from gene expression may be correlative, rather than causal, biomarkers. Thus genomic features are more likely amenable to functional validation experiments, such as testing if knockdown or overexpression of predicted functional biomarkers confers the predicted suppression or enhancement of sensitivity. By extension, genomic predictors of drug resistance may suggest targets for combination therapies [24].

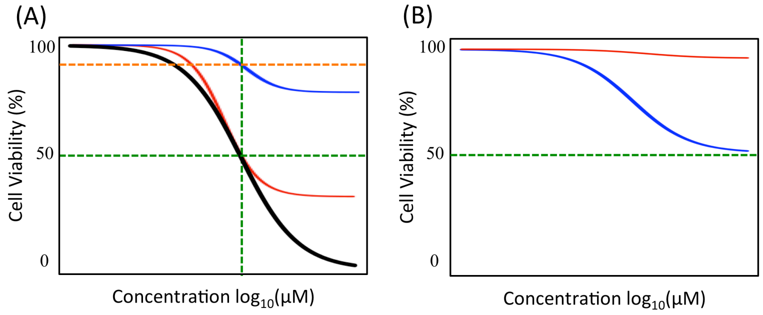

Figure 5 – Illustration of differences in dose response curves not captured by IC50 or EC50 statistics. (A) The red curve and black curve achieve 50% reduction in cell viability at the same compound concentration, but the black curve achieves increased reduction in cell viability at higher compound concentrations. Both curves correspond to the same IC50 value (vertical dotted green line), while the area under the dose response curve (AUC) captures the increased sensitivity shown in the black curve. The blue curve illustrates a sample with limited maximal reduction in cell viability at high compound concentrations. The EC50 statistic would be the same for the blue and black curves (vertical dotted green line), while the AUC statistic captures the increased response of the black curve. (B) The red and blue curves fail to reach 50% reduction in viability within the tested concentration range. The IC50 statistic would be set to the maximum tested concentration in both cases (or extrapolated outside the tested range), while the AUC statistic naturally captures the increased sensitivity displayed in the blue curve.

We also investigated alternative methods of assigning a summary statistic representing the sensitivity of a given cell line to a given compound. Predictive accuracy was improved by computing the area under/over the dose response curve (AUC/ActArea), as opposed to the more traditional metric of IC50. Following the theme described above, we suggest that AUC/ActArea captures more information from the experiment than IC50. Specifically, IC50 assumes a canonical sigmoidal shape of dose response curves, with zero growth inhibition in the absence of compound and 100% growth inhibition at high compound doses. This assumption fails to differentiate samples that achieve 50% growth inhibition at the same dose, even if one of the samples achieves far higher growth inhibition at higher doses (Figure 5A). An alternative statistic, EC50, is designed to account for this situation by computing the concentration at which a sample achieves 50% of its maximal growth inhibition; but this statistic suffers from additional degeneracies. Moreover, many samples do not achieve 50% growth inhibition within the tested dose range (Figure 5B). Therefore, IC50 calculations must set all such cases to a single threshold value (e.g. the highest tested dose, as reported for CCLE), or attempt to extrapolate based on fitted curves (as reported for Sanger). By contrast, the AUC/ActArea statistic is able to discriminate the examples listed above, and captures additional information contained in the dose response curves related to differential sensitivity (see Figure 5).

Our observation that continuous regression models, in general, outperform discrete classification models also follows the general theme of using data with the maximal amount of information as model inputs. Discretization of sensitivity data reduces the amount of information contained in the continuous valued data. Such a trade-off may be desirable if discretization reduces noise in the data (e.g. by only modeling the tails of the data, which are more likely to correspond to true differences in sensitivity and resistance, while ignoring the noisy intermediate values). Although this argument may apply in selective cases, it is highly dependent on choosing an accurate discretization scheme. We investigated several alternatives, including mixture models and mean and median-based deviation statistics (not shown). We observed that each scheme worked in some cases but not others; e.g. deviation-based statistics may classify no samples as sensitive or resistant for some compounds, while quartile-based statistics do not capture variable numbers of samples that may be sensitive to different compounds.

In addition to assessing the performance of modeling choices within our evaluated categories, we also assessed the relative importance of the categories themselves. Consistent with previous studies [21], our general conclusion is that the choice of input data (which molecular features are used and which compound is being predicted) dominates in explaining the high or low accuracy of a model. The choice of modeling algorithm also matters, but far less than the input data. While this conclusion may be sobering for data analysts (such as ourselves) in pursuit of the next great algorithm, we point out that our study was limited to machine learning methods designed to operate on specified feature and response data. Thus we suggest that optimization of methodologies in this context are unlikely to achieve dramatic improvements over current state-of-the-art methods; however, methodologies that incorporate additional information sources, such as other large-scale genomic datasets or information from pathway databases, were not tested in our study and may yield such improvements. This intuition is consistent with our observation that the quality and information content of input data dominates predictive performance, as such strategies augment the amount of information used to build a predictor. Indeed, in a recent community-based assessment of genomic predictors of breast cancer survival, the best performing method integrated information from all of TCGA in addition to the dataset directly used to build predictors [25, 26].

We note that our study does not assess all possible modeling choices. For example, we utilized the normalized genomic data provided by the CCLE and Sanger resources and did not assess the impact of alternative normalization or data processing procedures. We invite researchers throughout the community to build on and improve our work to investigate the myriad of additional approaches. Indeed, we hope the resource released by our study serves as initial input to a community effort promoting critical assessment of modeling methodologies. Innovative approaches developed by any researcher may be assessed in comparison to our results, thus providing a pre-defined set of performance criteria and baseline model scores against which novel approaches may objectively demonstrate their value.

REFERENCES

1. Ferte, C., et al., *Nature Reviews Clinical Oncology* **7**, 367-380 (2010).
2. Roche-Lestienne, C., et al., *New England Journal of Medicine* **348**, 2265-2266 (2003).
3. Peggs, K. and S. Mackinnon, *New England Journal of Medicine* **348**, 1048-1050 (2003).
4. Savage, D.G. and K.H. Antman, *New England Journal of Medicine* **346**, 683-693 (2002).
5. Barretina, J., et al., *Nature* **483**, 603-607 (2012).
6. Marum, L., *Pharmacogenomics* **13**, 740-741 (2012).
7. Garnett, M.J., et al., *Nature* **483**, 570-U87 (2012).
8. Forbes, S.A., et al., *Nucleic Acids Research* **39**, D945-D950 (2011).
9. Bindal, N., et al., *Genome Biology* **12**, 5-5 (2011.).
10. MacConaill, L.E., et al., *Plos One*, **4**, 11 (2009).
11. Jolliffe, I.T., *Journal of the Royal Statistical Society Series C* **31**, 300-303 (1982).
12. Wold, S., et al., *Chemometrics and Intelligent Laboratory Systems* **58,** 109-130 (2001).
13. Balabin, R.M. and E.I. Lomakina, *Analyst* **136**, 1703-1712(2011).
14. Wu, Y.F. and S. Krishnan, *Journal of Experimental & Theoretical Artificial Intelligence* **23**, 63-77 (2011).
15. Breiman, L., *Machine Learning* **45**, 5-32 (2001).
16. Tibshirani, R., *Journal of the Royal Statistical Society Series B* **58**, 267-288 (1996).
17. Tibshirani, R., *Journal of the Royal Statistical Society Series B* **73,** 273-282 (2011).
18. Friedman, J., et al., *Journal of Statistical Software* **33**, 1-22(2010).
19. Hoerl, A.E. and R.W. Kennard, *Technometrics* **42**, 80-86 (2000).
20. *CCLE data portal*. Available from: http://www.broadinstitute.org/ccle/home.
21. Shi, L.M., et al., *Nature Biotechnology* **28**, 827-U109 (2010).
22. Xu, C.J., et al., *Plos One* **7**, 8 (2012).
23. Niculescu-Mizil, R.C.A., *ICML2006*, 161-168 (2006).
24. Wei, G., et al., *Cancer Cell* **21**, 547-562 (2012).
25. Margolin, A.A., et al., *Science Translational Medicine*, **5**, 181 (2013).
26. Cheng, W.Y., et al, *Science Translational Medicine* **5**, 181 (2013).
27. Statnikov, A., et al, BMC Bioinformatics **9**, 319 (2008)