JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

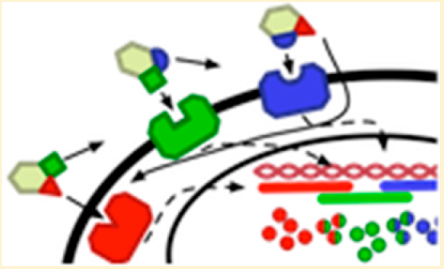
Using Molecular Features of Xenobiotics to Predict Hepatic Gene Expression Response

Guy Haskin Fernald[†] and Russ B. Altman*[,‡]

[†]Biomedical Informatics Training Program, Stanford University School of Medicine and [‡]Departments of Bioengineering and Genetics, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: Despite recent advances in molecular medicine and rational drug design, many drugs still fail because toxic effects arise at the cellular and tissue level. In order to better understand these effects, cellular assays can generate high-throughput measurements of gene expression changes induced by small molecules. However, our understanding of how the chemical features of small molecules influence gene expression is very limited. Therefore, we investigated the extent to which chemical features of small molecules can reliably be associated with significant changes in gene expression. Specifically, we analyzed the gene expression response of rat liver cells to 170 different drugs and searched for genes whose expression could be related to chemical features alone. Surprisingly, we can predict the up-regulation of 87 genes (increased expression of at least 1.5 times compared to controls). We show an average cross-validation predictive area under the receiver operating characteristic curve (AUROC) of 0.7 or greater for each of these 87 genes. We applied our method to an external data set of rat liver gene expression response to a novel drug and achieved an AUROC of 0.7. We also validated our approach by predicting up-regulation of Cytochrome P450 1A2 (CYP1A2) in three drugs known to induce CYP1A2 that were not in our data set. Finally, a detailed analysis of the CYP1A2 predictor allowed us to identify which fragments made significant contributions to the predictive scores.



INTRODUCTION

The liver response to a drug is critical in determining the ultimate effect the drug will have on the body. It is well-known that the first-pass effect of the cytochrome P450s, metabolizing enzymes, and transporters can greatly reduce the bioavailability of a drug or transform a prodrug into its active form.[1] Subsequent metabolic processes often eliminate drugs from the body either exclusively by the liver or in conjunction with the kidney. Because the liver performs these critical roles in processing xenobiotics, the liver response must be considered when determining drug doses or drug combinations to ensure that toxic levels of chemical species do not accumulate in the body and lead to adverse drug reactions.[2] Gene expression response is a well-known way to measure and quantify the liver's response to xenobiotic stimulus. However, the mechanism by which a drug will lead to a change in gene expression is not fully understood. In this work we were interested in determining which genes have their expression predictably changed in the liver directly in response to small molecules. In particular, we used publicly available data sets of drug and liver response data to seek genes whose expression was greatly affected by specific chemical features of small molecules.

Molecular fingerprints provide an efficient method to characterize a chemical as a set of molecular features represented by unique identifiers.[3] There are many varieties of fingerprinting methods used for similarity searching or virtual screening of large chemical libraries.[4] Extended connectivity fingerprints (ECFP) are based on chemical bond topology and capture features relevant to molecular activity.[5] They have been successfully applied for predicting chemical activities, even among structurally diverse compounds.[6,7] ECFP4 fingerprints generate unique identifiers for topological fragments that contain up to four bonds and are among the highest performing fingerprints for identifying similar molecules with known activities.[8]

Gene expression microarrays enable the simultaneous measurement of tens of thousands RNA expression probes in a tissue and have been used to detect significant differences between healthy and diseased tissues.[9] Gene expression experiments have also been used to detect significant RNA expression responses in tissues that have been treated with drugs.[10] For example, the Connectivity Map data set contains gene expression measurements on 1309 compounds and has been used in drug repositioning[11,12] and for elucidating the mechanism of action of drugs.[13] The DrugMatrix database contains RNA expression data from approximately 600 different compounds given in vivo to rats at different doses and time points and then measured on seven different tissues. These DrugMatrix data have been used to study drug toxicology and liver response profiles.[14−17] Others have connected drug structure to gene expression but focused on a large database

Received: December 10, 2012
Published: September 6, 2013

   dx.doi.org/10.1021/ci3005868 | J. Chem. Inf. Model. 2013, 53, 2765−2773

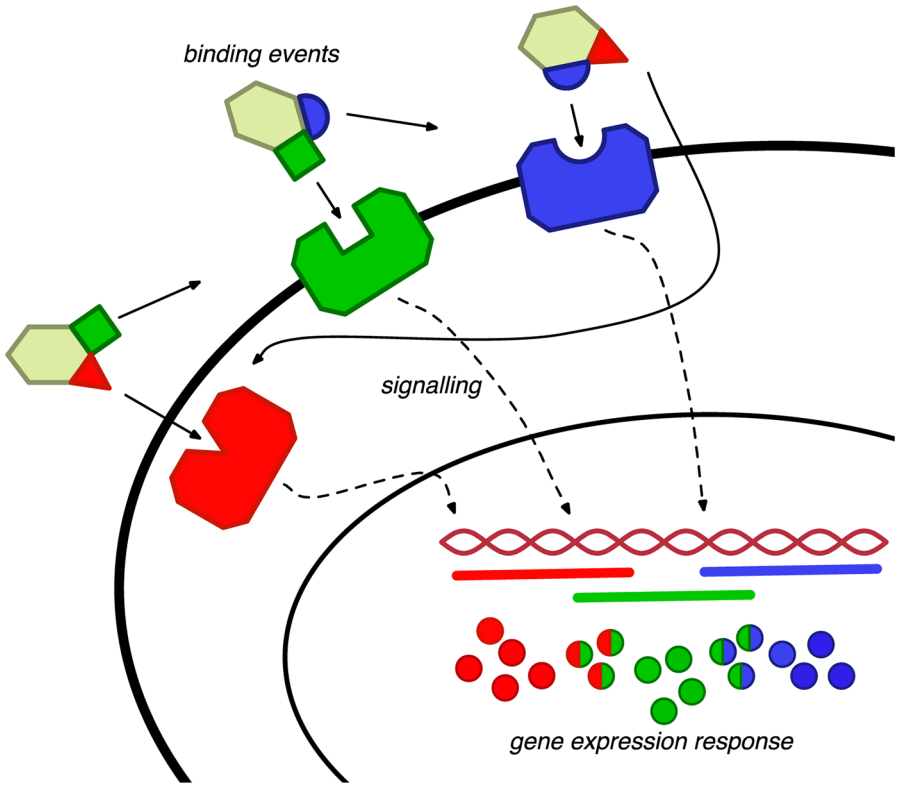


Figure 1. Drugs are shown schematically on the outside of the cell with different chemical features (green square, red triangle, blue half-circle). These features may lead to binding at one or more receptors that begin a signaling cascade ultimately leading to gene expression changes. Gene transcripts as shown as circles, where each circle is drawn in the color (or colors) associated with the chemical features associated with its expression. The work presented here focuses on finding the association of chemical features to the genes whose expression is directly predictable from a subset of these chemical features alone.

of predefined chemical structures and gene expression in cancer cell lines.[18]

In this study, we sought to connect the molecular level information contained in chemical fingerprints to the cellular level measures in the drug-induced liver gene expression data from DrugMatrix. In particular, we were interested in finding genes in the liver that are predictably up-regulated in response to small molecules, described using chemical fingerprints. A model by which a drug elicits changes in gene expression is illustrated in Figure 1. In this model a drug binds to a receptor, such as an allosteric site on a cell surface or a cytosolic protein, which triggers a series of signaling events. The signals reach the nucleus and trigger nuclear responses, which lead to transcriptional changes. The specific interactions and pathways that lead to these changes in transcription are largely unknown, and the elucidation of these pathways is an active area of research.[19−21] Therefore, any attempt to predict expression changes using existing pathways is necessarily based on incomplete information. Our approach circumvents this problem by finding direct relationships between chemical fingerprint features and gene expression changes, as also illustrated in Figure 1.

The method presented here has three parts: (1) we generate a reduced feature space to describe the drugs; (2) we use machine learning techniques to create and validate classifiers which predict if a gene will be up-regulated by a drug; and (3) we analyze the features of the classifiers to describe the drugs which are predicted to up-regulate the genes. To make our reduced feature space, we generated an matrix of fingerprints from a compendium of drugs in DrugBank[22] and then transformed the feature vectors into a reduced space using the twenty largest principal components of the matrix. We then applied machine learning algorithms to determine if any of the genes which were up-regulated at least 1.5 fold times in the liver could be predicted based on these features. Surprisingly, we found 87 genes that could be reliably predicted using only fingerprint information. We validated these predictions internally by cross validation and in an independent data set measuring the response to pregnenolone 16alpha-carbonitrile, a steroid hormone that was not included in any of our data sets. We also validated our predictive model for cytochrome P450 1A2 (CYP1A2) in three known CYP1A2 inducers that were not present in our data set. Finally, we analyzed the CYP1A2 model to highlight those fragments that are most informative for determining the predictive scores.

METHODS

Gene Expression Data Source. We downloaded the normalized Affymetrix whole genome 230 2.0 rat GeneChip array data liver gene expression from the DrugMatrix database[14] and filtered the experiments for probes which had a significant change in expression ($p < 0.05$) and at least a 1.5 fold increase in expression for at least 20 different drugs at any dose or time period and a 1.1 fold increase or lower for at least 20 different drugs. We empirically chose a cutoff of 1.5 for up-regulated genes because we wanted to make predictions on genes with strong signals, but we also needed to keep a sufficient number of genes in the data set. We chose a cutoff of 1.1 for down regulated genes to reduce the possibility of using false negatives in our training data. The resulting data set contains 170 distinct drugs and 3830 distinct probes. We converted this data set into

dx.doi.org/10.1021/ci3005868 | J. Chem. Inf. Model. 2013, 53, 2765−2773

a 170 × 3830 matrix where each entry has a value of 1 for probes having greater than 1.5 fold expression, a value of −1 for probes with less than 1.1 fold expression, and a value of 0 otherwise.

Molecular Feature Space of Chemicals. We matched 4069 drugs from DrugBank to molecules in ChEMBL[23] using drug names and synonyms and downloaded the canonical SMILES strings from ChEMBL. We then used these SMILES strings to generate extended connectivity fingerprint identifiers with a diameter of four atoms (ECFP4) using JChem Base 5.11.5, 2013, ChemAxon (www.chemaxon.com). Two-dimensional ECFP4 fingerprints are effective in recalling compound activities and are frequently used in QSAR and QSPR models.[24] Each identifier represents a topological substructure feature of a molecule. This resulted in a total of 19 810 distinct structure identifiers used to describe 4069 DrugBank drugs. Supporting Information Figure 3 shows a histogram of the drug counts per ECFP4. The histogram shows that there is an exponential drop off in the number of drugs per ECFP4 feature, with a few features that are nearly ubiquitous and many more that occur in only a handful of drugs. We created a binary feature matrix with 4069 rows (one for each drug) and 19 810 columns (one for each ECFP4 identifier). An entry *i*, *j* in the binary feature matrix was set to 1 if and only if drug *i* contained feature *j*; otherwise, the entry was 0. We then performed principal components analysis of this 4069 × 19 810 matrix using the prcomp function in the stats package in R.[25] We used the first 20 principal component loadings as the feature representation for chemicals.

Feature Representation of DrugMatrix Chemicals. We matched 170 drug names from the DrugMatrix data set to molecules in ChEMBL by matching drug names with molecule names and synonyms, downloaded their canonical SMILES strings, and generated ECFP4 fingerprints for each of these chemicals, resulting in 2372 distinct ECFP4 identifiers. We then used these 2372 identifiers to generate a 170 × 19 810 binary feature matrix where an entry *i*, *j* is was set to 1 if and only if the DrugMatrix drug *i* contained the ECFP4 identifier *j*, using only the DrugBank identifiers. If an ECFP4 identifier was present in the set of identifiers from DrugMatrix but not in the set from DrugBank, it was not used. We then projected this 170 × 19 810 matrix onto the first 20 components of the DrugBank PCA, resulting in a 170 × 20 feature matrix.

Correlation of Expression Similarity and Chemical Similarity. We computed the pairwise Tanimoto similarity of all 170 drugs using ECFP4 fingerprints and the pairwise correlation of the expression values of the 3830 probes used for training. We fit a linear model using the lm function in the R base package with the similarity in gene expression as the outcome variable and the similarity of the drugs as the predictor variable.[25]

Machine Learning Approach. We generated 10 different training and evaluation data sets by randomly choosing 80% of the 170 drugs for training (n = 136) and 20% of the drugs for evaluation (n = 34) for each of the 10 iterations. We then performed machine learning on each of the 10 training data sets with the following process:

(1) Generate a drug × PCA loading feature matrix. We selected drugs for which there were at least 20 positives (expression > 1.5) and 20 negatives (expression < 1.1), in our expression data set.

(2) For each of the 3830 probes, generate 25 bootstrap samples and use L1 constrained logistic regression models by training on 80% of the drugs (a 109 × 20 matrix), chosen randomly for each iteration, and measuring the area under the receiver operating characteristic curve (AUROC) using the remaining 20% of the drugs (n = 27). These AUROC values serve as a performance metric for how well the models classify each of the probes as up 1.5 fold.

The L1 logistic regression models were generated using the cv.glmnet function from the glmnet package in R.[26,27] We set the parameters of the cv.glmnet function to perform cross-validation using 10 folds and choose a value of that minimized the error within one standard error of the minimum, as recommended to avoid overfitting.[26] We calculated ROC curves and AUROC values using the ROCR package in R.[27]

We performed an additional 10 iterations of this machine learning approach with the positive and negative labels shuffled on the training data, resulting in baseline performance metrics for 100 randomly labeled training sets. The metrics from these values indicate what performance we would expect to see at random.

Machine Learning Evaluation. We calculated and plotted the mean and standard deviation of the AUROC values across the 25 iterations for all of the probe classifiers within each of the 10 training sets. A comparison with the randomly shuffled results indicated that any classifier with a mean AUROC ≥ 0.534 had significant performance, and classifiers with a mean AUROC ≥ 0.7 were 1 standard error above the 0.534 cutoff. We therefore selected all classifiers which had a mean AUROC ≥ 0.7 in any of the training sets and evaluated the performance of the models by generating an AUROC using the evaluation drugs (n = 34) for each iteration. If a classifier had a mean AUROC ≥ 0.7 in more than one training set, we tested it in each of the models.

External Validation. We downloaded the GSE 4959 data set[28] from the Gene Expression Omnibus and analyzed the pregnenolone 16alpha-carbonitrile treated versus control samples using the GEO2R tool.[29] This data set measures the rat liver response to pregnenolone 16alpha-carbonitrile, a drug that is not present in the original data set. We chose all gene expression classifiers that had a mean AUROC ≥ 0.7 in four or more of the iterations (n = 87) for external validation and generated complete models using all 170 drugs to train the L1 constrained logistic regression. We then generated the ECFP4 features for pregnenolone 16alpha-carbonitrile and then projected those features into the top 20 PCA components using the PCA rotation matrix from the DrugBank data set. We matched 62/87 of the probes from GSE4959 to probes in our data set using Affymetrix identifiers. Finally, we scored each of these 62 probes and normalized their scores between 0 and 1 to generate a single ROC curve. We computed the Tanimoto similarity of pregnenolone 16alpha-carbonitrile to all 170 molecules in the data set to determine its maximum similarity. We also computed the Tanimoto similarity of the gene expression signature sets consisting of all genes that were up-regulated at least 1.5 fold in the presence of pregnenolone 16alpha-carbonitrile versus the expression signatures for all 170 molecules from the DrugMatrix data set.

Validation and Analysis of CYP1A2 Classifier. The best performing classifier in our data set was for a transcript encoding Cytochrome P450 1A2 (CYP1A2), an important

enzyme involved in the metabolism of xenobiotics. We searched the PubMed database with the terms "CYP1A2" and "induction" and identified three small molecules that have been shown to induce expression of CYP1A2, listed in Table 1,

Table 1. Three Drugs Known to Induce Expression of CYP1A2 and Their Predicted Score Using the CYP1A2 Classifier[a]

drug name	reference	max Tanimoto similarity	score
omeprazole	Rost et al. (1992)[33]	0.46	1.59
3,3-diindolylmethane	Lake et al. (1998)[34]	0.29	1.28
RO4938581	Bundgaard et al. (2013)[35]	0.14	0.31

[a]Also shown is the maximum Tanimoto similarity to any of the 170 drugs in the data set. The similarity score was calculated using ECFP4 fingerprints.

which were not in our data set. We generated the ECFP4 fingerprints for these drugs and projected the resulting feature vectors into our feature space using the top 20 principal components of the PCA rotation matrix from the DrugBank data set. We scored these molecules using the classifier to determine if they were correctly predicted to induce expression of CYP1A2.

To analyze the features used to predict up-regulation of CYP1A2, we took the complete model for CYP1A2 and computed a weight vector $\vec{w}$, using the following equation:

$$\vec{w} = \sum_{i \in \{\text{CYP1A2 PCs}\}} \beta_i \cdot \overrightarrow{\text{PCA}_i}$$

where {CYP1A2 PCs} is the set of principal components chosen by the L1 constrained logistic regression to have nonzero beta coefficients and each β_i is the beta coefficient for the corresponding principal component. The resulting vector $\vec{w}$ is a vector of 19 810 real values, $\{w_1, ..., w_{19\,810}\}$ where each value w_j represents the contribution that the presence of ECFP4_j makes to the score of any given molecule. We tested for enrichment of each of the ECFP4s in the true positives or true negatives using Fisher's Exact test with multiple hypothesis testing correction using the false discovery rate.[30] We checked for the presence of the ECFP4 fragments with the largest positive and negative weights using ChemAxon's "jcsearch" program and generated visualizations of the fragments with ChemAxon's "mview" program.

RESULTS

Gene Expression. The Affymetrix Rat 230 2.0 GeneChip data set we used contains multiple time points and multiple doses for some drugs (657 distinct drug-dose combinations in rat liver tissue with measurements on 31 042 probes). For the purpose of this work we were interested in building classifiers, which requires choosing cutoffs to determine which genes were true positives (up-regulated) and which were true negatives (not up-regulated). In order to avoid potential mislabeling of probes we focused on the more extreme signals by labeling probes with fold changes of 1.5 or greater as true positives (up-regulated) and probes with fold changes of 1.1 or less as true negatives (not up-regulated). Any gene measurements in between these two values were not considered in this analysis. In order to ensure the data set had a sufficient number of positive and negative examples we filtered for probes that were up-regulated at least 1.5 fold in the presence of at least 20 distinct drugs and 1.1 fold or less for at least 20 distinct drugs. After this filtering we were left with 3830 probes. The distribution of fold change values is shown in Figure 2.

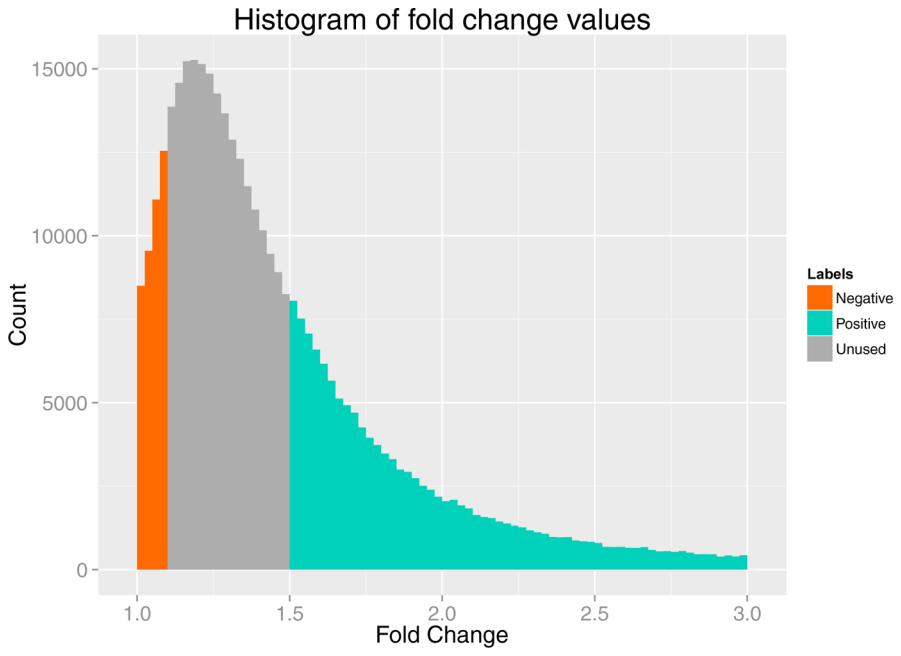


Figure 2. Distribution of fold change values for transcripts up-regulated by drugs. The area highlighted in cyan shows the expression values which are at least 1.5 fold and were used as positive examples (n = 126 040); the area highlighted in orange shows expression values less than or equal to 1.1 times and were used as negative examples (n = 41 185). The gray area shows probes with expression between 1.1 and 1.5 (n = 198 266), which were not used.

dx.doi.org/10.1021/ci3005868 | J. Chem. Inf. Model. 2013, 53, 2765−2773

Molecular Descriptors for Drugs. The principal components analysis of the DrugBank ECFP4 matrix (4069 drugs × 19 810 ECFP4 identifiers) resulted in a projection of the drugs into a reduced feature space that describes most of medically relevant chemical space. A scree plot of the first 100 components for this projection (Figure 3) shows that there is a considerable drop in variance explained after the first few components, followed by a gradual drop off. We projected the ECPF4 features into the top 20 principle components and used those 20 principal component loadings as the features to describe the drugs. Taken together these 20 components explain 30% of the variance in the extended connectivity fingerprints for 4069 drugs taken from DrugBank. This is a considerable reduction in feature space and leaves a significant proportion of the variance unexplained. However, it focuses our search on molecular fragments that occur in multiple drugs and greatly reduces the chance of over fitting our classifiers. Using this projection we were able to reduce our feature set from 2372 distinct ECFP4 identifiers to 20 loadings on the top 20 principal components, resulting in a 170 drug by 20 principal component loading matrix.

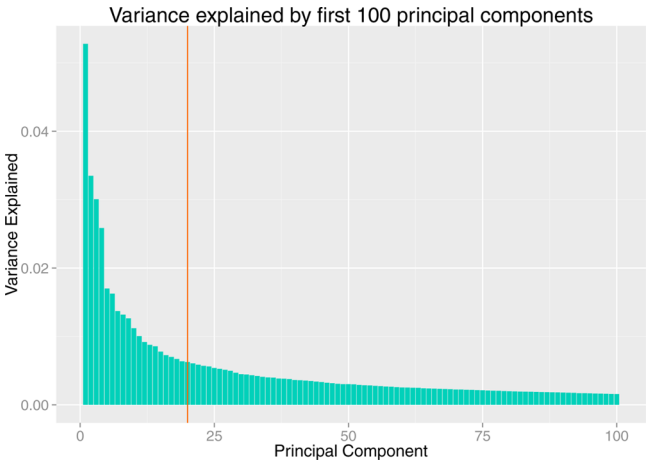


Figure 3. Scree plot of first 100 principal components of extended connectivity fingerprint identifiers for 4069 drugs from DrugBank. Taken together these components essentially describe all current pharmacologically relevant chemical fingerprint space. We chose the first 20 components, delimited by the vertical line, to represent the features of drugs, which accounts for 30% of the variance in the DrugBank ECFP4 (4069 drug × 19 810 ECFP4) matrix.

Correlation of Expression Similarity and Chemical Similarity. Interestingly, the Tanimoto similarity of the drugs was not correlated with similarity in gene expression. The correlation of between similarity in structure and similarity in expression was 0.10.

Performance of Classifiers. Figure 4A plots the mean AUROC of each gene in one training set and shows that most genes did not perform well and have a mean AUROC at or near 0.5, which indicates that their performance is no better than random. However, some classifiers performed well in many iterations and have performance considerably greater than 0.5.

The results of our permutation analysis are shown in Figure 4B. A mean AUROC > 0.534 would occur by chance only 5% of the time, giving an estimated p-value, $\hat{p} \leq 0.05$. These genes are highlighted in purple in Figure 4A and B. With this estimated p-value there are 1198 probe classifiers in Figure 4A with significant performance. To identify the strongest signals, we focused on the probes with an even greater mean AUROC ≥ 0.7, and in the case of the data shown in Figure 4A, there 171 of these genes. All 10 of the summary plots for the data sets are provided in Supporting Information Figure 1.

The performance of the probe classifiers also varied depending upon which of the 10 training and evaluation sets we used. Some probe classifiers had performance ≥0.7 in just one of the 10 iterations and some performed well in all 10 iterations. Figure 5 shows box plots summarizing the mean

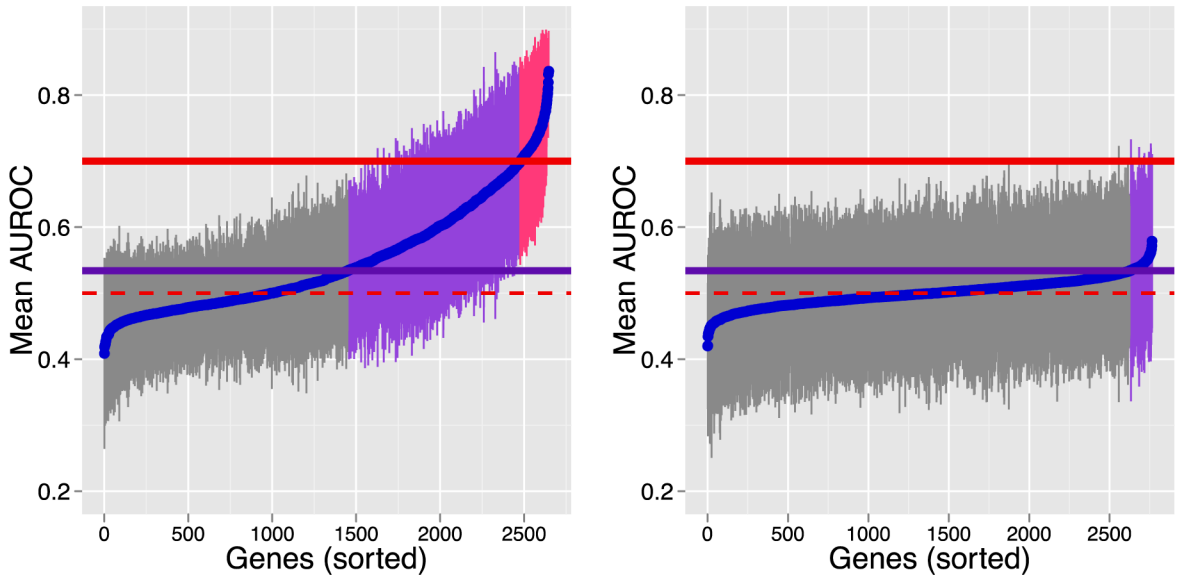


Figure 4. Plot of the mean area under receiver operating characteristic curve (AUROC) for classifiers for each of 3830 genes sorted in increasing order for real data (A) and randomized labels (B). The dotted red line indicates 0.5, where performance is equivalent to random chance. The solid red line indicates a mean AUROC of 0.7. Classifiers above the red line will rank a randomly chosen true positive greater than a randomly chosen true negative 70% of the time. (A) Larger number of high scoring genes. (B) Only 5% of genes with randomly permuted labels have a mean AUROC > 0.534, indicated by the purple line.

dx.doi.org/10.1021/ci3005868 | J. Chem. Inf. Model. 2013, 53, 2765−2773

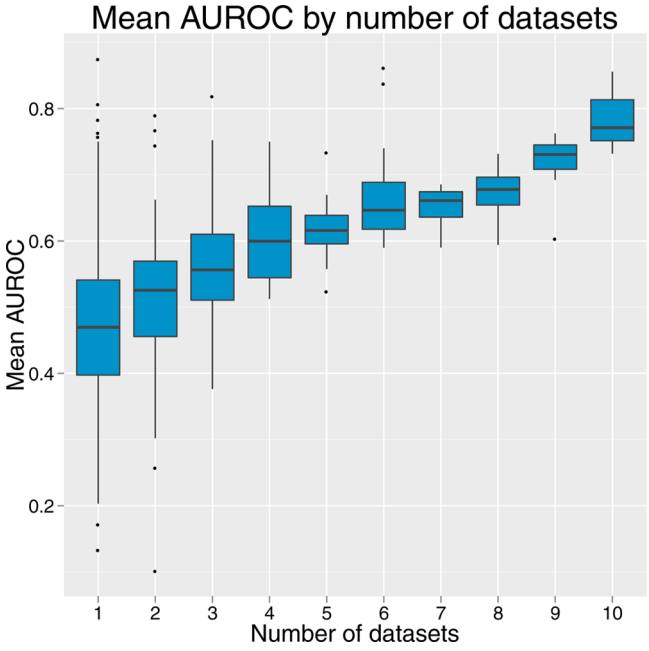


Figure 5. Box plots summarizing results from 10 randomly sampled training (n = 136) and test (n = 34) data sets from DrugMatrix. Scores of all probe classifiers with an AUROC ≥ 0.7 in at least one of the 10 data sets is included. As the number of data sets in which probe performs well increases the mean AUROC increases. We chose the probes that had a mean AUROC ≥ 0.7 in four or more data sets for external validation.

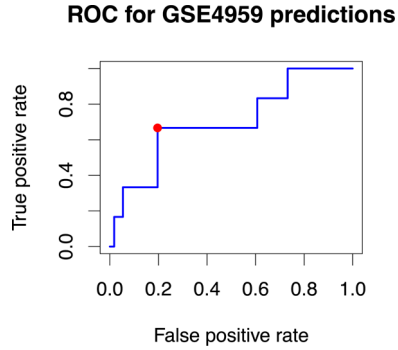


Figure 6. Receiver operating characteristic (ROC) curve for 62 probes from GSE4959, measuring the rat liver response to pregnenolone 16alpha-carbonitrile. We matched 62 of the 87 probe classifiers to probes in GSE4959. We normalized the scores generated by the classifiers between 0 and 1 and then used the true positive labels to generate the ROC curve. With a false positive rate of 0.2 the true positive rate is 0.67, indicated by the red circle. The area under this ROC curve is 0.7, which is in alignment with the results from our cross-validation.

AUROCs for probe classifiers that have mean AUROC ≥ 0.7 in one or more of the 10 training data sets. As expected, the mean performance of the probe classifiers increases as they occur in the top results of more data sets. The mean performance of probe classifiers which performed well in only one or two data sets is near 0.5, but if a probe classifier performed well in four or more data sets then the mean AUROC > 0.6 across all 10 data sets. To account for this variation we chose to use the 87 probe classifiers that were consistently up in four or more data sets for external validation.

External Validation. The maximum Tanimoto similarity of pregnenolone 16alpha-carbonitrile to any of the 170 drugs from DrugMatrix was 0.3286. The maximum Tanimoto similarity of the expression profile of genes up-regulated at least 1.5 fold was 0.389. The ROC curve for predicting the rat liver gene expression response to pregnenolone 16alpha-carbonitrile in the 62 chosen classifiers is shown in Figure 6 and shows an AUROC of 0.7. The list of 62 probes used to generate this ROC curve is shown in Supporting Information Table 1.

Validation and Analysis of CYP1A2 Classifier. The probe with the best predictive classifier performance was 1387243_at, which is a transcript encoding Cytochrome P450 1A2 (CYP1A2). The classifier for this probe had a mean AUROC of 0.86 with a standard error of 0.09 across all 10 of the training and evaluation sets. We scored three molecules, shown in Table 1, which are known to induce the expression on CYP1A2 using this classifier and were not present in our data set. None of molecules were significantly similar to any other molecules in our data set, the maximum Tanimoto similarity of the molecules to any of the 170 drugs ranged from 0.14 to 0.46. The three known CYP1A2 inducing molecules that we identified, omeprazole, 3,3-diindolylmethane, and RO4938581, scored within the range of scores for the true positive drugs, as shown in Figure 8. Using the weight vector $\vec{w}$, as described in the methods, we identified the ECFP4 identifiers that contributed the five largest positive and five largest negative weights to each of the CYP1A2 validation drugs. Supporting Information Figure 2 shows each of the drugs and highlights the positive and negative features associated with the ECFP4 identifiers.

DISCUSSION

We have shown that there is a subset of genes in the liver that can reliably be predicted to be up-regulated based on the chemical information found in ECFP4s. Because there is minimal correlation between the ECFP4 fingerprint similarities of the drugs and the gene expression that the drugs induce (ρ = 0.10) the connections between ECFP4 and gene expression must be made through the presence of coordinated subsets of ECFP4 identifiers rather than overall chemical similarity. We created our classifiers using a drastically reduced set of features generated by projecting our data into the first 20 principal components from an ECFP4 matrix consisting of all the ECFP4 identifiers generated from the drugs in DrugBank. These 20 principal components only capture 30% of the variance in the DrugBank ECFP4 matrix and many ECFP4 fragments in the DrugMatrix data set are not represented in the feature vectors. By using this reduced feature space the classifiers were forced to find combinations of principal components that best predict up-regulation of the genes. For the majority of the genes the classifiers were unable to make accurate predictions. However, 87 genes were reliably predictable across multiple bootstrap samples including CYP1A2, which we were able to validate in three drugs known to induce CYP1A2. Using a deeper analysis of the CYP1A2 classifier we identified which ECFP4 features of these drugs made the largest contributions to their positive and negative scores.

Strengths and Weaknesses. There are many concerns when evaluating a machine learning experiment with large feature sets, such as overfitting, validation, and interpretation. In this study we addressed these concerns at each stage of our experiment: (1) generation of our drug feature space; (2) machine learning and validation; and (3) analysis of features.

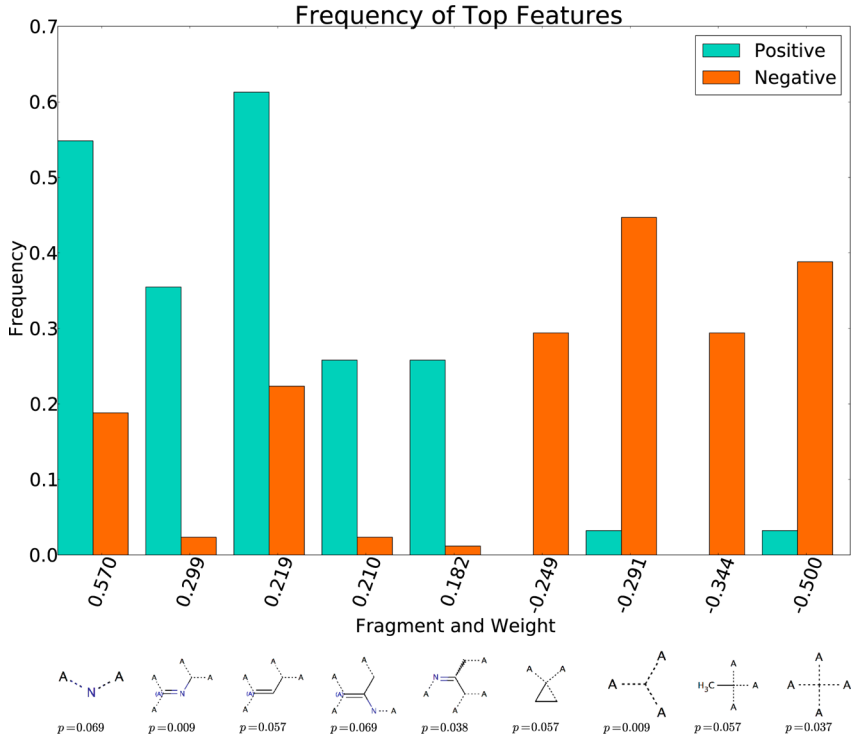


Figure 7. Histogram showing the frequency of ECFP4 identifiers that are enriched in either the true positive drugs, which up-regulate CYP1A2 (in cyan) and true negative drugs, which do not up-regulate CYP1A2 (in orange). A fragment is considered significantly enriched if the adjusted *p*-value for enrichment in the set using Fisher's exact test was less than 0.1. The weight for each ECFP4 is shown above each fragment, indicating how much the presence of each identifier contributes to the score for a given drug.

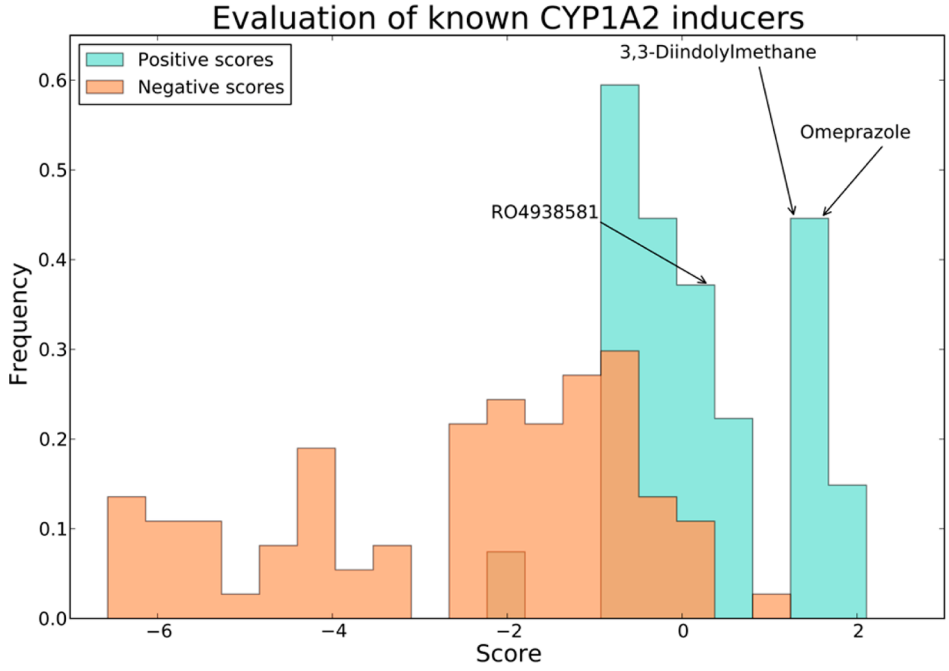


Figure 8. Histogram showing scores of true negative drugs that do not up-regulate CYP1A2 at least 1.5 fold (in orange) and scores for true positive drugs (in cyan) that do up-regulate CYP1A2. The scores of three known CYP1A2 inducers (from Table 1) are annotated on the histogram. The three drugs score well within the range of the true positives.

Generation of Feature Space. Rather than generate a feature space based on the only drugs in the training data set we used 4069 drugs from DrugBank, which resulted in 19 810 ECFP4s. Given that DrugBank contains FDA approved drugs, experimental drugs, and nutraceuticals, this set of ECFP4s describes nearly all known features relevant to known drug based medical therapy. Even so, the ECFP4s in the training data included 501 ECFP4s that were not in the DrugBank set because not all drugs and drug variants in the training data are present in DrugBank. We then transformed the 170 feature

dx.doi.org/10.1021/ci3005868 | J. Chem. Inf. Model. 2013, 53, 2765−2773

vectors for the training drugs into the top 20 PCA components of the DrugBank ECFP4 feature matrix. These top 20 features describe 30% of the variance in the DrugBank features. Given this vast reduction in feature space and by using the additional feature selection embedded in L1-regularized regression, we eliminated any concerns of over fitting. With larger high-quality expression data sets for more drugs it would be possible to use more PCA components that would explain more of the variance.

Machine Learning. In order to address statistical significance in our machine learning approach we used repeated bootstrap sampling in order to find predictive models that were useful in repeated data sets. Randomized labels resulted in a mean AUROC greater than 0.534 in only the top 5% of cases (Figure 4B), indicating that a mean AUROC > 0.534 will occur by chance only 5% of the time. By only considering classifiers with a mean AUROC > 0.7 in several bootstrap samples, we ensure that we are not selecting classifiers which are effective only due to chance or which are only effective on the training data. The validation of these classifiers in the external pregnenolone 16alpha-carbonitrile data set (Figure 6) further affirms the reliability of the models.

Analysis of Features. A benefit of this approach is that it allows for interpretation of the significant features used for prediction. Figure 7 shows the most significant ECFP4s for predicting if CYP1A2 is up-regulated, the frequency in which they appear in the true positives and true negatives, and the weight that each ECFP4 identifier contributes to the score. Five ECFP4 identifiers were enriched in the true positive drugs and four ECFP4 identifiers were enriched in the true negative drugs, using Fisher's exact test ($p < 0.1$). It is important to recognize that even if an individual feature is enriched in the true positives or true negatives, that feature alone does not determine the score. As seen in Figure 7, there may be similarities and even overlap between the features as well as correlations between the ECFP4s that tend to appear together in drugs. The total score for a given drug comes from the sum of all the weights for the features present in the drug, including the overlapping and correlated features. Despite the limitations in looking at just a few chemical features, it is informative to look at the features that make the largest contributions to the score for a drug. The three validation drugs, known to induce expression of CYP1A2, are shown in Supporting Information Figure 2 in three panels that highlight the features that have the largest positive and largest negative weights. It is interesting to note how the features tend to overlap over common regions of the molecular graphs so that a given portion of a molecule may contribute both positive and negative weights. However, the net contribution of those overlapping features results in either a positive or negative change to the score due to the differences in the size of the values.

Relationship to Previous Work. Others have shown that chemical structure can be related to gene expression. For example, Blower et al. calculated correlations between the structure of cancer drugs and gene expression in NCI-60 cell lines.[18] In that work the authors were able to successfully construct substructure queries for chemical classes that induced similar gene expression sets. However, that work was focused on cancer cell lines and entire expression signatures, instead of individual gene classifiers based on structural features. ECFPs have been used to predict activity classes; Heikamp et al. showed that feature selection on ECFPs with gain-ratio analysis could be used to create reduced fingerprint feature sets for searching activity classes, suggesting that such feature sets could be used for scaffold hopping.[6,31] Similarly, the work presented here uses machine learning and feature selection in order to find sets of features that are predictive of an increase in gene expression. Bender et al. used chemical structure to predict ligand binding in a selection of targets associated with adverse events to successfully develop predictive models for adverse events.[32] We anticipate that additional training data would further increase the number of genes that could be reliably predicted to have increased expression. Given a large enough set of genes, there would be intriguing drug development applications such as identifying drugs or combinations of drugs with increased likelihood of adverse or toxic effects.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figures S1−S3 and Table 1. The file ci3005868_si_005.zip (expression_matrix.tsv) is a tab separated file containing a 170 × 3380 ternary matrix with values −1, 0, or 1. There is one column for each of the 3380 probes described in the study and one row for each of the 170 drugs studied. Each *i,j* entry has a value of 1 if the drug *i* increased the expression of probe *j* by 1.5 fold or greater, a value of −1 if the expression was 1.1 fold or less, and 0 otherwise. The file ci3005868_si_006.zip (feature_matrix.tsv) is a tab separated file containing a real valued 170 × 20 matrix of PCA loadings. Each row contains the drug name and the 20 PCA loadings which were used as features for that drug, as described in the manuscript. The rows for both files are in the same order. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: russ.altman@stanford.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

NIH grants NIH LM07033, R24GM61374 (to G.H.F.) and the NIH/NIGMS Pharmacogenetics Research Network and Database and the PharmGKB resource NIH R24GM61374 (to R.B.A.).

■ REFERENCES

(1) Shen, D.; Kunze, K.; Thummel, K. Enzyme-catalyzed processes of first-pass hepatic and intestinal drug extraction. *Adv. Drug Delivery Rev.* **1997**, *27*, 99−127.

(2) Liguori, M. J.; Waring, J. F. Investigations toward enhanced understanding of hepatic idiosyncratic drug reactions. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 835−846.

(3) Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Timmerman, H. *Handbook of Molecular Descriptors*; John Wiley & Sons: Weinheim, Germany, 2000.

(4) Willett, P. Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* **2011**, *672*, 133−158.

(5) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(6) Heikamp, K.; Bajorath, J. How do 2D fingerprints detect structurally diverse active compounds? Revealing compound subset-specific fingerprint features through systematic selection. *J. Chem. Inf. Model.* **2011**, *51*, 2254−2265.

(7) Heikamp, K.; Bajorath, J. Large-Scale Similarity Search Profiling of ChEMBL Compound Data Sets. *J. Chem. Inf. Model.* **2011**, *51*, 1831−1839.

(8) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *J. Med. Chem.* **2010**, *53*, 5707−5715.
(9) Barrett, T.; Troup, D. B.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Muertter, R. N.; Holko, M.; Ayanbule, O.; Yefanov, A.; Soboleva, A. NCBI GEO: archive for functional genomics data sets−10 years on. *Nucleic Acids Res.* **2011**, *39*, D1005−10.
(10) Hughes, T. R.; Marton, M. J.; Jones, A. R.; Roberts, C. J.; Stoughton, R.; Armour, C. D.; Bennett, H. A.; Coffey, E.; Dai, H.; He, Y. D.; Kidd, M. J.; King, A. M.; Meyer, M. R.; Slade, D.; Lum, P. Y.; Stepaniants, S. B.; Shoemaker, D. D.; Gachotte, D.; Chakraburtty, K.; Simon, J.; Bard, M.; Friend, S. H. Functional discovery via a compendium of expression profiles. *Cell* **2000**, *102*, 109−126.
(11) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313*, 1929−1935.
(12) Sirota, M.; Dudley, J. T.; Kim, J.; Chiang, A. P.; Morgan, A. A.; Sweet-Cordero, A.; Sage, J.; Butte, A. J. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **2011**, *3*, 96ra77.
(13) Iorio, F.; Tagliaferri, R.; di Bernardo, D. Identifying network of drug mode of action by gene expression profiling. *J. Comput. Biol.* **2009**, *16*, 241−251.
(14) Ganter, B.; Tugendreich, S.; Pearson, C. I.; Ayanoglu, E.; Baumhueter, S.; Bostian, K. A.; Brady, L.; Browne, L. J.; Calvin, J. T.; Day, G.-J.; Breckenridge, N.; Dunlea, S.; Eynon, B. P.; Furness, L. M.; Ferng, J.; Fielden, M. R.; Fujimoto, S. Y.; Gong, L.; Hu, C.; Idury, R.; Judo, M. S. B.; Kolaja, K. L.; Lee, M. D.; McSorley, C.; Minor, J. M.; Nair, R. V.; Natsoulis, G.; Nguyen, P.; Nicholson, S. M.; Pham, H.; Roter, A. H.; Sun, D.; Tan, S.; Thode, S.; Tolley, A. M.; Vladimirova, A.; Yang, J.; Zhou, Z.; Jarnagin, K. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.* **2005**, *119*, 219−244.
(15) Ganter, B.; Snyder, R. D.; Halbert, D. N.; Lee, M. D. Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. *Pharmacogenomics* **2006**, *7*, 1025−1044.
(16) Fielden, M. R.; Brennan, R.; Gollub, J. A Gene Expression Biomarker Provides Early Prediction and Mechanistic Assessment of Hepatic Tumor Induction by Nongenotoxic Chemicals. *Toxicol. Sci.* **2007**, *99*, 90−100.
(17) Natsoulis, G.; Pearson, C. I.; Gollub, J.; P Eynon, B.; Ferng, J.; Nair, R.; Idury, R.; Lee, M. D.; Fielden, M. R.; Brennan, R. J.; Roter, A. H.; Jarnagin, K. The liver pharmacological and xenobiotic gene response repertoire. *Mol. Syst. Biol.* **2008**, *4*, 175.
(18) Blower, P. E.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Yu, L.; Richman, S.; Weinstein, J. N. Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J.* **2002**, *2*, 259−271.
(19) Huang, R.; Wallqvist, A.; Thanki, N.; Covell, D. G. Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute's anticancer screen and drug mechanism of action. *Pharmacogenomics J.* **2005**, *5*, 381−399.
(20) Chen, X.; Xu, J.; Huang, B.; Li, J.; Wu, X.; Ma, L.; Jia, X.; Bian, X.; Tan, F.; Liu, L.; Chen, S.; Li, X. A sub-pathway-based approach for identifying drug response principal network. *Bioinformatics* **2011**, *27*, 649−654.
(21) Silberberg, Y.; Gottlieb, A.; Kupiec, M.; Ruppin, E.; Sharan, R. Large-scale elucidation of drug response pathways in humans. *J. Comput. Biol.* **2012**, *19*, 163−174.
(22) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−41.
(23) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−7.
(24) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49*, 108−119.
(25) Team, R. C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012.
(26) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1−22.
(27) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21*, 3940−3941.
(28) Guzelian, J.; Barwick, J. L.; Hunter, L.; Phang, T. L.; Quattrochi, L. C.; Guzelian, P. S. Identification of genes controlled by the pregnane X receptor by microarray analysis of mRNAs from pregnenolone 16alpha-carbonitrile-treated rats. *Toxicol. Sci.* **2006**, *94*, 379−387.
(29) Barrett, T.; Troup, D. B.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Muertter, R. N.; Holko, M.; Ayanbule, O.; Yefanov, A.; Soboleva, A. NCBI GEO: archive for functional genomics data sets−10 years on. *Nucleic Acids Res.* **2010**, *39*, D1005−D1010.
(30) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **1995**, *57*, 289−300.
(31) Krueger, B. A.; Dietrich, A.; Baringhaus, K.-H.; Schneider, G. Scaffold-hopping potential of fragment-based de novo design: the chances and limits of variation. *Comb. Chem. High Throughput Screen* **2009**, *12*, 383−396.
(32) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2007**, *2*, 861−873.
(33) Rost, K. L.; Brösicke, H.; Brockmöller, J.; Scheffler, M.; Helge, H.; Roots, I. Increase of cytochrome P450IA2 activity by omeprazole: evidence by the 13C-[N-3-methyl]-caffeine breath test in poor and extensive metabolizers of S-mephenytoin. *Clin. Pharmacol. Ther.* **1992**, *52*, 170−180.
(34) Lake, B. G.; Tredger, J. M.; Renwick, A. B.; Barton, P. T.; Price, R. J. 3,3′-Diindolylmethane induces CYP1A2 in cultured precision-cut human liver slices. *Xenobiotica* **1998**, *28*, 803−811.
(35) Bundgaard, C.; Badolo, L.; Redrobe, J. P. RO4938581, a GABAAα5 modulator, displays strong CYP1A2 autoinduction properties in rats. *Biochem. Pharmacol.* **2013**, *85*, 1363−1369.