HYBRID QUANTITATIVE STRUCTURE-ACTIVITY RELATIONISHIP MODELING OF HUMAN CYTOCHROME P450 ISOFORM 2C9 INHIBITION

by

DANIEL C. PINOLINI

A thesis submitted to the

Graduate School-Camden

Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Computational and Integrative Biology

Written under the direction of

Dr. Hao Zhu

And approved by

_____

Dr. Hao Zhu

_____

Dr. Kwangwon Lee

_____

Dr. Sunil Shende

_____

Camden, New Jersey, May 2015

ABSTRACT OF THE THESIS

Hybrid Quantitative Structure-Activity Relationship Modeling of Human Cytochrome P450 Isoform 2C9 Inhibition

by DANIEL C. PINOLINI

Thesis Director:

Dr. Hao Zhu

Purpose Human Cytochrome P4502C9 is a vital enzyme in human drug metabolism. Inhibition of P450 2C9 can cause critical Drug Drug Interactions (DDI). Great resources can be saved if the potential inhibition of new compounds (e.g. new drugs) can be evaluated before chemical synthesis. Computational models are promising tools to realize this goal. Previous Quantitative Structure Activity Relationship (QSAR) modeling works performed on this enzyme were not significant due to limitation of available data as training sets, and all suffer from shortcomings of traditional QSAR approaches, especially the issue of active cliffs. A successful large scale model that incorporates biological response data would be beneficial to future drug discovery.

Methods In this study, QSAR modeling approaches were employed to develop multiple computational models for P450 2C9 inhibition. A training set of 20,839 compounds and an external set of 20,655 compounds were compiled from PubChem assay data. After chemical descriptors were generated for each compound, random forest and support vector machine algorithms were used to develop QSAR models based on the training set. The results of individual models were averaged as consensus predictions. Individual and

consensus models were first validated using five-fold cross-validation. Then the validated models were used to predict the external sets.

Results The predictivity of external set compounds for developed models was acceptable for QSAR modeling (Consensus model statistics: Sensitivity = 67.3%, Specificity = 71.3%, Correct Classification Rate = 69.3%). Incorporation of biological response data as extra descriptor information into traditional QSAR approaches improved predictivity of the associated models (Sensitivity = 67.1%, Specificity = 75.8%, Correct Classification Rate = 71.5%). These improvements were shown to be statistically significant.

Conclusions In this study, QSAR models of CYP2C9 inhibition were successfully developed for a large set of compounds. Biological response data was successfully incorporated into traditional QSAR modeling procedure, leading to improvement in predictivity. This development could be used to more successfully predict the potential DDI of new compounds.

ACKNOWLEDGEMENTS

I would like to first and foremost thank my advisor and mentor, Dr. Hao Zhu. The time I have spent working with you has been invaluable for me both for the skills you helped me develop and the personal growth I experienced during my time with you. Your drive and determination have been a constant inspiration, and I am extremely grateful for the opportunity I have had to both learn and grow under your guidance.

I would also like to thank Dan Russo and Wenyi Wang, my fellow lab members, and now friends. All the work I have done would not have been anywhere near as enjoyable or proceeded as smoothly without your help and the constant positive environment in the lab.

Finally, I would like to thank Marlene Kim, my lab mate, my mentor, and my friend. I can comfortably say that none of what I have accomplished would have been possible without you. Thank you for your endless and readily available help, and for all of your advice to motivate me to do better and to be better.

DEDICATION

I would like to dedicate this work to Daniel and Jacqueline Pinolini, my mother and father, and to Thomas Pinolini, my brother, who have always been there to celebrate my success, and to pick me back up when I've fallen. Without all of your unconditional love and support nothing I have accomplished would have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

Human Cytochromes P450 (CYPs) are a group of hemoproteins belonging to the enzyme superfamily of monooxygenases, and are, in general, the terminal oxidase enzymes in electron transfer chains.[1] CYPs contain a heme group which is tethered to the hemoprotein by a cysteine thiolate ligand[2]. They are largely found embedded in the endoplasmic reticulum, and have been identified in all domains of life, including viruses.[1-3] These monooxygenase enzymes are best known for the catalysis of the oxygenation of an enormous number and variety of both endogenous and xenobiotic substrates including lipids, steroidal hormones, drugs, and toxic substances.[4] Most of this oxidative activity is due to hydroxylation reactions in which the reducing agent Nicotinamide adenine dinucleotide phosphate (NADPH) facilitates the insertion of one atom of oxygen into the aliphatic position of an organic substrate (RH), while the other oxygen atom is reduced to water, according to the stoichiometric equation[1]:

$$NADPH + O_2 + RH + H+ \rightarrow NADP^+ + ROH + H_2O \quad (1)$$

Among the human Cytochrome P450 family, one of the most important in terms of both functionality and research potential is the isoform 2C9 (CYP2C9).[5] CYP2C9 makes up approximately 15-20 percent of the total CYP protein in liver microsomes and metabolizes roughly one fifth of commercially prescribed drugs, usually those with a narrow therapeutic index.[6-8] Such examples include, but are not limited to warfarin (anticoagulant), phenytoin (anticonvulsant), tolbutamide (potassium channel blocker), losartan (angiotensin II channel receptor antagonist), and glipizide (high blood pressure),

other non-steroidal anti-inflammatory drugs, and endogenous compounds such as arachidonic and linoleic acid.[9,10] The majority of substrates for the CYP2C9 tend to be acidic, polar molecules containing an aromatic ring.[11-13] The prominence of the CYP2C9 isoform in the human metabolic system, as well as the wide range of drugs it metabolizes, makes the research into inhibition of the enzyme critical to avoiding potential drug-drug interactions.[14]

In recent years, quantitative structure-activity relationship (QSAR) approaches have become an integral part of the drug discovery process.[15] QSAR models are regression or classification models that relate a set of predictor variables, or descriptors, to the potency or the categorical value of a response variable. These descriptors consist of physico-chemical properties or theoretical molecular properties of chemicals. QSAR models aim to summarize a supposed relationship between chemical structures and biological activity within a data set of compounds, and to predict the activities of new chemicals. While initially useful, however, the QSAR hypothesis that chemically similar compounds tend to have similar activities has proven to be insufficient for accurate predictivity involving complex biological systems.[16] Systems involving such things as drug metabolizing enzymes expose this weakness in QSAR, as there is no mechanism built into the methodology to handle "activity cliffs," a phenomenon where compounds that are chemically dissimilar, yet elicit the same biological response.[17]

In this study, not only is there an attempt to model the chemical inhibition of a human cytochrome P450 enzyme on a much larger scale than any previously attempted studies, but also an attempt to rectify the innate problems of the QSAR modeling approach through the integration of bioassay data. It is expected that by using publicly

available bioassay data to create a hybrid model, improvements to traditional modeling techniques will lead to a more successful modeling approach.

Materials and Methods

CYP2C9 Inhibition Data Set

The CYP2C9 inhibition data set was curated from several PubChem bioassay sets, including AIDs 777, 1020, 1851, 344535, and 349322 . Ultimately, the CYP2C9 inhibition set used in this study was compiled from the PubChem AID 777, a high throughput screening (HTS) assay for the inhibition of CYP2C9 in an NADPH regenerating system which utilizes glucose-6phosphate dehydrogenase (G6PDH), as this assay possessed more strict criteria for activity compared to other inhibition assays.[18] For the modeling purpose of this study, compounds were considered inhibitors of the enzyme ("actives") if they exhibited more than 50 percent inhibition of the conversion of luciferin-H to luciferin at a concentration of 5 μM, and non-inhibitors ("inactive") if inhibition was less than 50 percent. Originally the data set consisted of 18,731 active compounds and 77,129 inactive compounds, for a total of 95,860 compounds. Following selection of this data set, the results were compared to a second assay, AID 1020, a counter screen to the original assay. This assay assessed compounds for the inhibition of G6PDH. Compounds that showed a positive result in both assays would be considered false positive results in AID 777, since it utilized a system dependent on G6PDH.[18] All compounds that showed a false positive result according to the counter screen were removed from the data set. The duplicate compounds, inorganic compounds, and

mixtures were also removed since the descriptor generators used cannot deal with these compounds. This data set is too large and out of the scope of the current QSAR approaches due to the limitation of the calculation power of computers. In order to make a QSAR modeling study feasible for the data set, the size of the set was reduced by randomly selecting 50 percent of the compounds.[14] Furthermore, the data set was heavily biased due to many more inactive compounds than active compounds, so an equivalent number of inactive compounds to active compounds were randomly selected, bringing the data set to a balanced ratio of active to inactive compounds.[14] The final modeling set consisted of 9,376 active compounds and 11,463 inactive compounds, for a total of 20,839 compounds.

The external set was organized using the remaining 50 percent of the compounds that were not included for the modeling set. Using random selection as described above, a balanced external set was created consisting of 9,292 active compounds and 11,361 inactive compounds, for a total of 20,653 compounds. Both data sets were larger than most previously used CYP2C9 inhibition datasets for modeling.

Chemical Descriptors

For the purpose of model creation in this study, two sets of chemical descriptors were generated. The first set of descriptors was generated using Molecular Operating Environment (MOE) Version 2011, and included topological indices, structural keys, E-state indices, physical properties (*e.g.* logP, molecular weight, molecular refractivity, etc.), and topological polar surface area. A total of 192 MOE descriptors were originally generated. The descriptors were normalized, and redundant descriptors were identified

and removed, leaving a total of 142 MOE descriptors for modeling implementation. In addition to the MOE descriptors, a second set of descriptors were generated using Dragon (DRAG) Version 6.0 (Talete SRL, Milano, Italy), which included E-state values, E-state counts, constitutional descriptors, walk and path counts, connectivity and information indices, 2D autocorrelations, Burden eigenvalues, molecular properties (*e.g.* hydrophilic factor), Kappa, hydrogen bond acceptor and donor counts, molecular distance edge, and molecular fragment counts. Following data normalization and removal of redundant descriptors, 1,108 Dragon descriptors fulfilled the criterion and were used for modeling.

Modeling Approaches

The chemical descriptors generated were used in the machine learning algorithms Random Forest (RF)[19] and Support Vector Machine (SVM).[20-23] Both the RF and SVM algorithms were implemented with R 3.0.2.[24]

RF is an ensemble learning method developed by Leo Breiman and Adele Cutler that generates multiple decision trees based on input variables from a data set for purposes of classification or regression modeling.[19] These decision trees are created by selecting with replacement *n* samples from a training set of N compounds. From M variables at each data point, *m* variables are selected at random, and the best split from these *m* variables is used. Each new data point is pushed down the tree, and the predicted output is averaged for a final predicted value.

SVMs are supervised learning models with associated learning algorithms developed by Vapnik et al that analyze data and recognize patterns for classification and regression analysis.[20-23] An SVM model works by representing elements of the training

set as points in n-dimensional space, mapped such that examples of separate categories are divided by as wide a gap as possible, known as a hyperplane. New examples are then mapped in the same space as the training set, and predicted to belong to a category, usually active or inactive, based on which side of the hyperplane they fall on. Classification with the largest marginalization possible is sought.

Modeling Workflow

Each individual model was developed using a combination of Dragon or MOE descriptors, and one of the modeling algorithms (RF or SVM). The use of both sets of descriptors and both modeling algorithms resulted in the development of four individual models: RF_MOE, RF_DRAG, SVM_MOE, and SVM_DRAG. Each individual model was validated using 5-fold cross-validation, a method in which the data set is partitioned into five complimentary subsets, and one subset is used as a test set, with the other four serving as a training set. This is repeated five times until each subset has been used as a test set. The prediction values from all four individual models for each compound were then averaged together, generating a consensus prediction model.[25,26] Models were then revised using a novel in-house tool discussed below for implementing biological similarity data into current QSAR modeling approaches.

Universal Statistical Evaluation of Model Performance

In order to create the models for this study, a variety of approaches and descriptor sets were used. Because of this, a universal statistical metric for comparison was needed to meaningfully assess the performance and predictive capability of each model. Thus, three statistical parameters were used for performance evaluation; sensitivity (percentage

of active compounds predicted correctly), specificity (percentage of inactive compounds predicted correctly), and correct classification rate (CCR, average of sensitivity and specificity). These statistical parameters are all defined as percentages according to the following equations:

$$Sensitivity = \left(\frac{True\ Positives}{True\ Positives+False\ Negatives}\right) x100 \quad (2)$$

$$Specificity = \left(\frac{True\ Negatives}{True\ Negatives+False\ Positives}\right) x100 \quad (3)$$

$$CCR = \left(\frac{Sensitivity+Specificity}{2}\right) x100 \quad (4)$$

Biological Similarity Analysis to Integrate Extra Biological Data as Descriptors

Following model creation and statistical analysis, improvement of modeling results was sought through the rectification of certain shortcomings of current QSAR techniques.[25,27-30] Presently used QSAR methods construct models based solely on chemical descriptors, which are only based on chemical structures of compounds, but give little to no consideration to the biological data that the chemicals elicit.[25,27-30] To remedy this shortcoming, a novel approach was developed to integrate biological assay and/or biological response data to current QSAR modeling approaches. The first part of the novel approach was to develop a metric of similarity search to use in the evaluation of biological data. This was accomplished in two parts. Before bioassay data could be applied to modeling, it was necessary to create a biological response profile of the compounds used for modeling by filtering redundant and noisy data. A profiling tool was developed to compile response data from bioassays available through PubChem.[23]

Compounds were assigned a binary score of 1, -1, or 0 based on assay response data, corresponding to active, inactive, or inconclusive response or missing value, respectively.[23] The number of responses of each type was recorded for each assay evaluated, and the top or most relevant assays were selected based on largest number of active responses and largest number of overall responses. Once this profile was constructed, it was used to calculate the weighted estimate of biological similarity (WEBS) score between every two compounds in the data set. The WEBS score was calculated based on inputted *in vitro* assay responses, as shown in the following equation:

$$\frac{\sum(p + (\omega)n)}{\sum(p + (\omega)n + d)}$$

where *p* is the number of assays in which both compounds showed the same response, *n* is the number of assays in which both compounds showed a different response, *d* is the number of assays in which comparison was impossible due to an inconclusive response, and ω is the negative response balancing weight, a variable that balances the ratio of active and inactive responses in the biological profile. In addition to similarity, the WEBS calculation also showed a confidence value, or a measure of how meaningful the similarity score is. When isolated, the top portion of the above equation can be compared to the total number of assays used to determine the confidence of the calculated similarity score. Following calculation of both similarity and confidence, the top five biological nearest neighbors (compounds with highest biological similarity score value) with an acceptable confidence value (20% of the number of assays manually selected) were identified for each compound in the modeling set. The experimental activity of each compound was compared to that of each of the aforementioned nearest neighbors, and

compounds with at least three of five nearest neighbors with different activities were removed from the modeling set, and performance statistics were recalculated (*e.g.* Compound CID 324 has an experimental activity of 1, or active, and has three nearest neighbors with experimental activities of -1, or inactive. Compound CID 324 is removed).

Results

Dataset Overview

The chemical space occupied by the compounds used in this study was analyzed by performing a Principle Component Analysis (PCA) using the MOE chemical descriptor values. Using the 142 MOE descriptors, principal components were calculated for the modeling set compounds. The three most important principal components, representing approximately 60% of the variance in the database, were selected to generate a three dimensional plot (**Figure 1**) of the 20839 compounds present in the modeling set. This plot is a visual representation of the chemical space occupied by the compounds comprising the modeling set.

Modeling Results

During modeling creation, four individual models and one consensus model were developed using 9,376 CYP2C9 inhibitors ("actives") and 11,463 non-inhibitors ("inactives"). Following individual model creation, a consensus model was generated by averaging the prediction values of each compound from each individual model into a single value, and then classifying the averaged value.

The five-fold internal cross validation results for all models are shown in **Figure 2a**. The sensitivity, specificity, and CCR values for this validation ranged from 65-68%, 68-73%, and 66-70% respectively. The best overall performance when ranked by CCR value was shown when using the Random Forest algorithm (RF_MOE and RF_DRAG models). The consensus model showed performance comparable or slightly superior to the aforementioned individual models in all three categories, with sensitivity, specificity percentages of 67%, 71%, and 69%, respectively.

Once the model was developed and underwent preliminary testing with internal cross-validation, it was used to predict an external set of 20,655 compounds (9,294 actives, 11,361 inactives). These compounds were randomly selected from the original assay, but were not part of the subset of compounds used to train the model, and thus were considered as unknown to the model. The model performance of this external validation is shown in **Figure 2b**. The ranges of the percentage values of the sensitivity, specificity, and CCR for the external validation were 63-68%, 68-73%, and 66-70% respectively. It is notable that despite the compounds being "unknown" to the model, model performance did not suffer as compared to the internal validation.

Using the biological similarity data to augment the QSAR modeling approach through the developed WEBS equation, the model performance statistics saw improvement. As seen in **Table 2**, the improvement shown in the models was demonstrated to be statistically significant by using an F-test. For each individual model, the $F_{Calculated}$ value was greater than the $F_{Critical}$ value, confirming statistical significance of results. **Figure 3** shows the model performance improvement as compared to performance using only QSAR modeling. On average, models showed an overall

improvement of approximately three percent. This is likely due to the fact that systematic removal of compounds based on aforementioned criteria from the modeling statistics results in less compounds with activities opposite of that of their nearest neighbors. This in turn results in a model with fewer compounds with uncertain predictivity.

Discussion

Evaluation of Initial Model Performance

Prior to the start of this study, most QSAR modeling work concerning human cytochrome P4502C9 was performed on a relatively small scale. As the 2C9 isoform of CYP450 is a vital enzyme to human drug metabolism[5,26,30], high percentage statistical performances was desired, however this required a sacrifice of the scope of the work, as high percentage performance depended on small modeling sets creating through meticulous selection. In addition to the standard hope for high percentage model performance, one of the primary goals of this study was to expand the scope of modeling work done on CYP2C9. Relative to datasets previously modeled, both the modeling set and the external validation set were enormous, and as a result possessed a much more extensive chemical space, chemical variance, and coverage. Given the size of the dataset used, as well the fact that greater compound diversity can complicate or reduce the effectiveness of QSAR modeling techniques, the performance of the individual models developed in this study is more than acceptable. The CCR range of 66-70% indicates that the results are as good as can be hoped for given the scope of the modeling attempted.

Inclusion of Biological Data

The 2C9 isoform of CYP450 acts on several diverse substrates, and is part of the complex biological pathway that is human drug metabolism. As such, the predictive success of models created using only traditional QSAR methods suffers.[25,28-30] Traditionally, QSAR techniques are applied using only chemical descriptors obtained through MOE or some other source.[15] This becomes problematic, however, when these same modeling approaches are applied to biological systems, as the use of only chemical data renders it impossible for such a model to make a distinction between compounds that are structurally very similar, but exhibit a very different biological response.[25,28-30] These "activity cliffs" limit the predictive power and application of the QSAR techniques.

The inclusion of bioassay response data into the modeling approach provides extra information about the target compounds that likely lead to the improvement shown in the modeling results. As shown in **Table 4**, there were many instances within the dataset where compounds that were chemically similar did not show similar biological response, and chemicals that showed similar biological response were not chemically similar. This is likely due to the nature of CYP2C9 as a key enzyme in drug metabolism.[5,6,8] CYP2C9 metabolizes various types of drugs,[9,10], which necessitates that an "active" biological response be elicited from compounds that are chemically dissimilar. An example of this is shown in **Table 4, Set 2**. Compounds CID 742428 and CID 2897066 are very chemically dissimilar, with a Tanimoto similarity coefficient of only 0.21, but exhibit an extremely similar biological response and are both active according to assay data.

The improved modeling performance shown after the incorporation of biological data reinforces that QSAR techniques alone are not sufficient for accurate

modeling of complex biological systems. This study demonstrates that while applying bioassay data in the form of a biological similarity score may not absolve QSAR of all fundamental shortcomings, it could lead to considerable progress in the resolution of the activity cliff problem that plagues computational modeling of biological system responses.[25,27-30] Successful integration of biological similarity and bioassay data into traditional QSAR techniques will lead to the development of models that show consistently higher predictivity, and thus will be far more useful in the drug discovery process.

Conclusion

In this study, a large and diverse data set of 9,376 CYP2C9 inhibitors and 11,463 non-inhibitors was compiled. This data set was used to develop several different QSAR models, which were then validated by using an internal 5-fold cross-validation approach. The resulting models were then used to predict an external dataset of 9,294 inhibitors and 11,361 non-inhibitors. In both the internal validation and external predictions, the consensus model showed similar predictivity to the best performing individual models, and the models as a whole showed excellent performance considering the scale of the data set as compared to previous studies.

Upon further examination of the dataset and the modeling results, it became apparent that attempted prediction of the response of a complex biological system on such a large and diverse dataset was resulting in a decreased predictivity due to activity cliff issues. In an attempt to resolve these issues, a novel approach was used to incorporate biological response and bioassay data into traditional QSAR modeling

approaches via a biosimilarity score. The resulting hybrid model showed improved performance as compared to the models created using only QSAR methods, and suggests the necessity of incorporating biological response data into future modeling approaches for optimal performance.

Figures

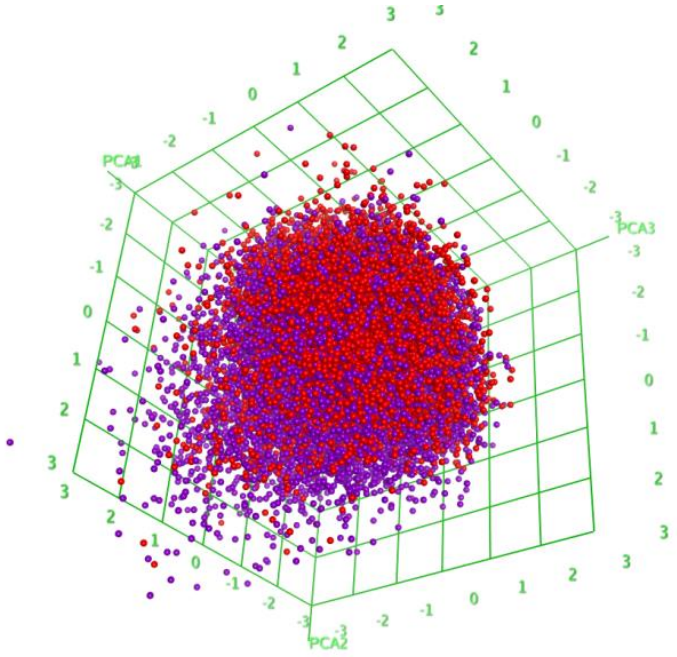
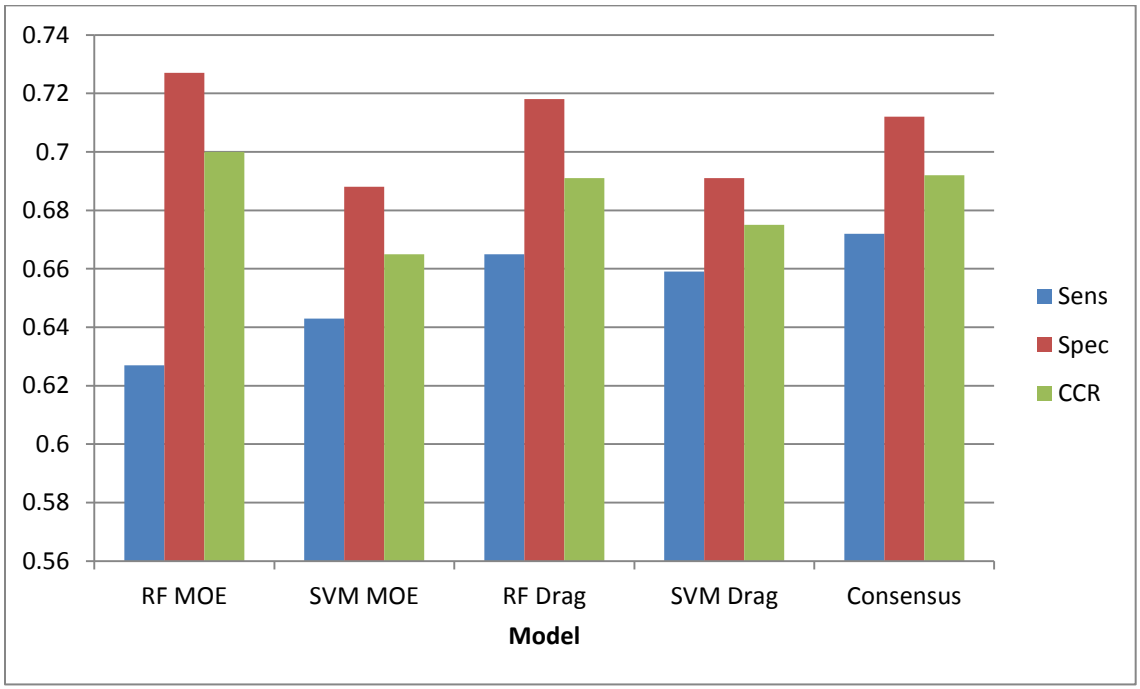
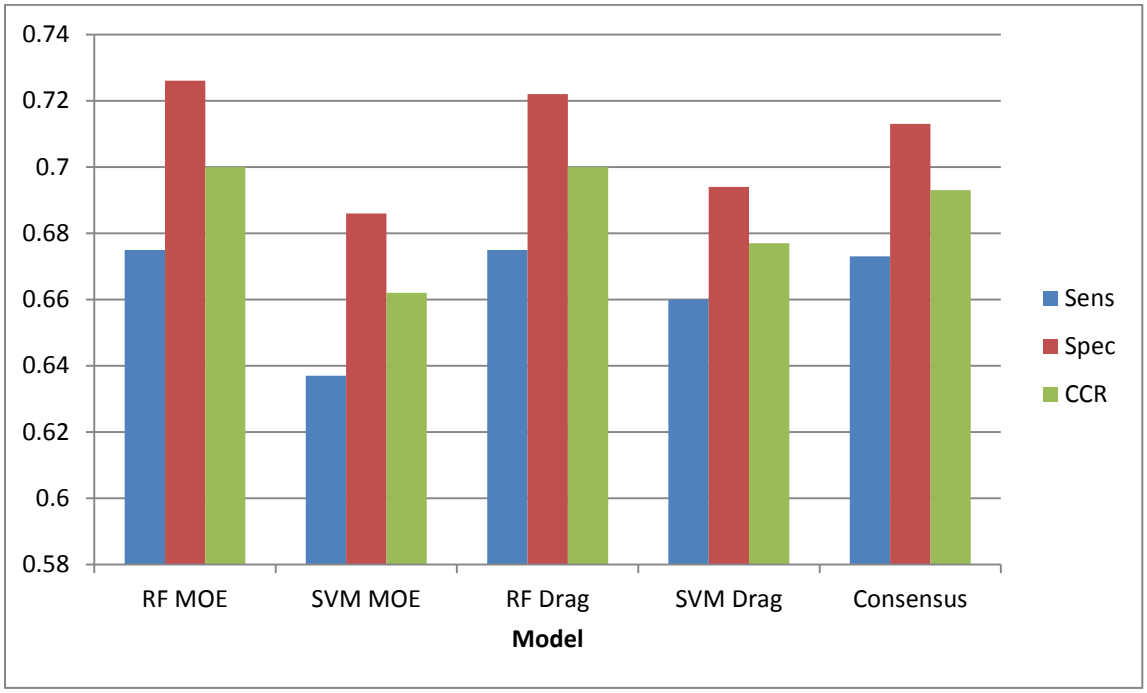


Figure 1. Chemical space occupied by modeling set of 9,376 actives (CYP2C9 Inhibitors, shown in red) and 11,463 inactives (CYP2C9 non-inhibitors, shown in purple) shown using top 3 principle components of MOE descriptors.

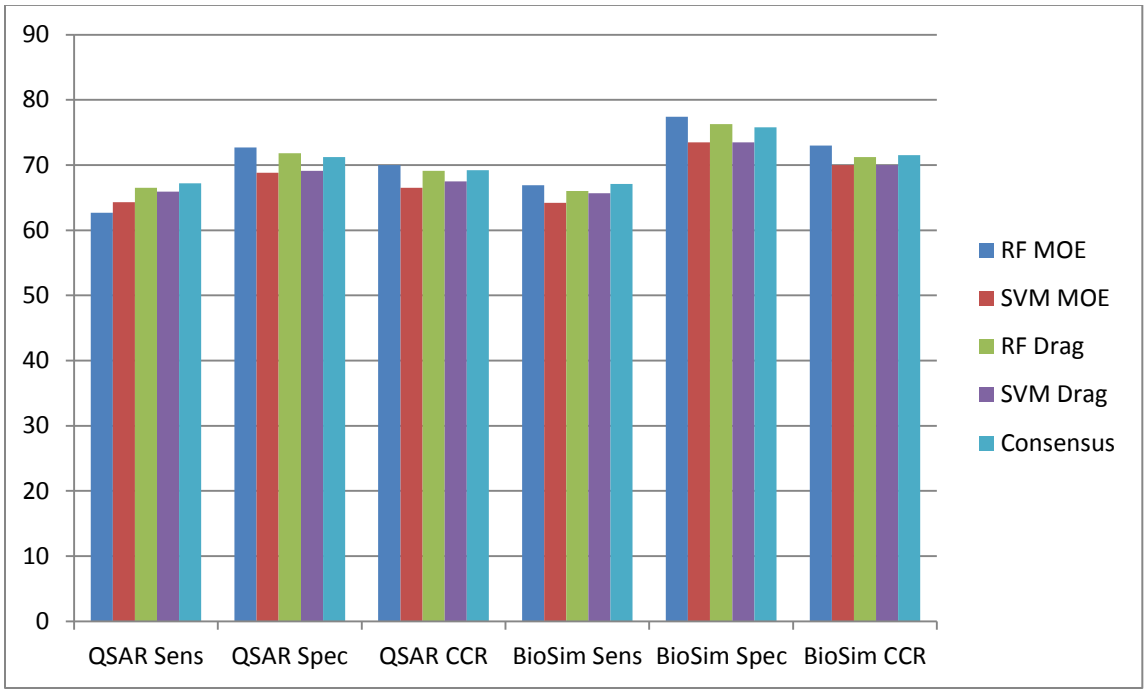


(a)

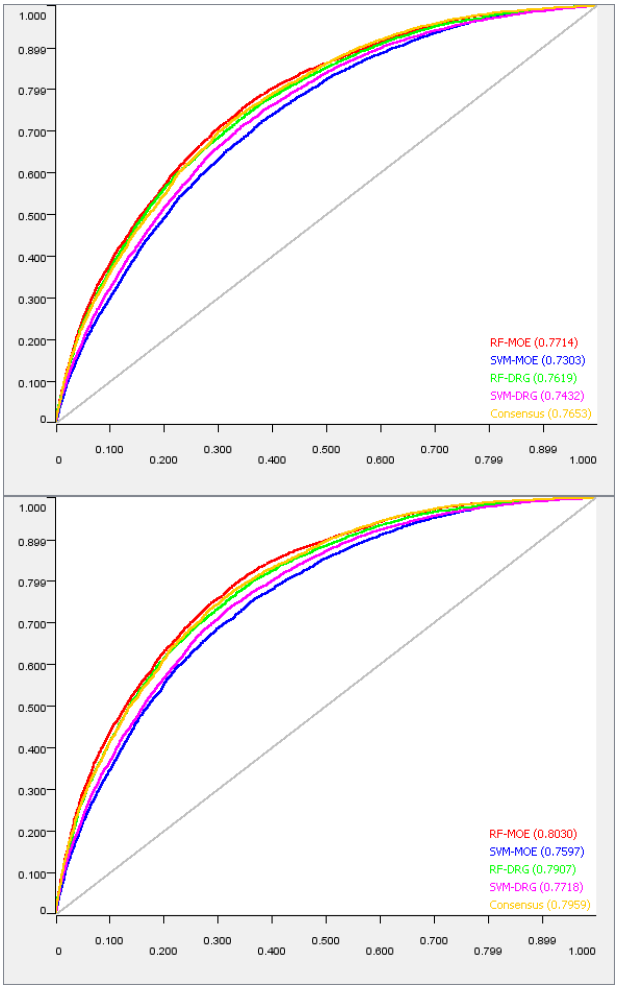


(b)

Figure 2. Statistics for performance evaluation for the four individual models and the consensus model for (a) internal five-fold cross validation; (b) prediction results for 20,655 compound external validation set



(a)



(b)

Figure 3. (a) Comparison of statistical performance of models using only QSAR and models using QSAR supplemented with Biological Similarity data. (b) ROC Curve comparison of model performance using only QSAR (left) with using QSAR supplemented with Biological Similarity data (right)

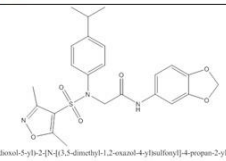
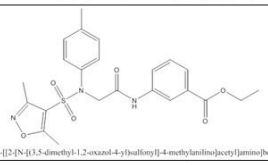
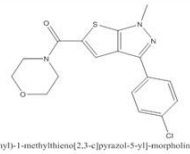
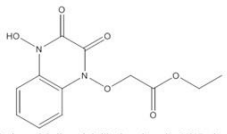
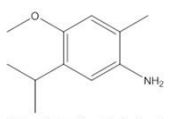
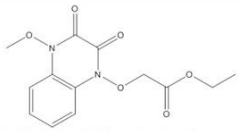
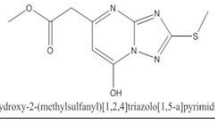
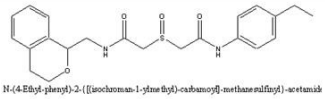
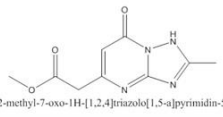
		CID	Illustration	Chemical Similarity	Biological Similarity	Experimental Activity
Set One	Compound	5307557	 N-(1,3-benzodioxol-5-yl)-2-[N-[(3,5-dimethyl-1,2-oxazol-4-yl)sulfonyl]-4-propan-2-ylanilino]acetamide	N/A	N/A	Inactive
	Biological Nearest Neighbor	3237614	 ethyl 3-[[2-[N-[(3,5-dimethyl-1,2-oxazol-4-yl)sulfonyl]-4-methylanilino]acetyl]amino]benzoate	0.43	0.97	Inactive
	Chemical Nearest Neighbor	5308778	 [3-(4-chlorophenyl)-1-methylthieno[2,3-c]pyrazol-5-yl]-morpholin-4-ylmethanone	0.9	0.82	Active
Set Two	Compound	742428	 ethyl [(4-hydroxy-2,3-dioxo-3,4-dihydroquinoxalin-1(2H)-yl)oxy]acetate	N/A	N/A	Active
	Biological Nearest Neighbor	2897066	 5-Isopropyl-4-methoxy-2-methyl-phenylamine	1	0.21	Active
	Chemical Nearest Neighbor	658428	 ethyl 2-(4-methoxy-2,3-dioxoquinoxalin-1-yl)oxyacetate	0.91	0.93	Inactive
Set Three	Compound	1474397	 methyl 2-[7-hydroxy-2-(methylsulfanyl)[1,2,4]triazolo[1,5-a]pyrimidin-5-yl]acetate	N/A	N/A	Active
	Biological Nearest Neighbor	646035	 N-(4-Ethyl-phenyl)-2-({[isochroman-1-ylmethyl]-carbamoyl}-methanesulfinyl)-acetamide	1	0.36	Active
	Chemical Nearest Neighbor	666935	 methyl 2-(2-methyl-7-oxo-1H-[1,2,4]triazolo[1,5-a]pyrimidin-5-yl)acetate	0.9	0.92	Inactive

Table One. Examples of sets of compounds from the data set. Included in each set is a predicted compound, the biological nearest neighbor of that compound, the chemical nearest neighbor of that compound, the biological and chemical similarity scores of each neighbor relative to the predicted compound, and the experimental assay activity of each compound. Biological Similarity was calculated using WEBS. Chemical similarity was calculated using Tanimoto Coefficient.

	RF MOE	SVM MOE	RF Dragon	SVM Dragon	Consensus

$F_{Calculated}$	5.442	2.337	6.059	2.112	3.892
$F_{Critical}$	1.026	1.026	1.026	1.026	1.026

Table 2: Results of an F-test performed to evaluate if improvement seen in hybrid modeling approach were statistically significant.

References

1. Danielson, P. (2002). The Cytochrome P450 Superfamily: Biochemistry, Evolution and Drug Metabolism in Humans. *Current Drug Metabolism*, 561–97.

2. Sligar, S. G., Cinti, D. L., Gibson, G. G., & Schenkman, J. B. (1979). Spin state control of the hepatic cytochrome P450 redox potential. *Biochemical and Biophysical Research Communications*, *90*(3), 925–932.

3. Lamb, D. C., Lei, L., Warrilow, A. G. S., Lepesheva, G. I., Mullins, J. G. L., Waterman, M. R., & Kelly, S. L. (2009). The first virally encoded cytochrome p450. *Journal of Virology*, *83*(16), 8266–9.

4. Guengerich, F. P. (2008). Cytochrome p450 and chemical toxicology. *Chemical Research in Toxicology*, *21*(1), 70–83.

5. Sigel, R., Sigel, A., & Sigel, H. (2007). *The Ubiquitous Roles of Cytochrome P450 Proteins: Metal Ions in Life Sciences*. New York: Wiley.

6. Cohan, K., Paine, S., & Waters, N. (2008). Advancements in Predictive In Silico Models for ADME. *Current Chemical Biology*, *2*, 215–228.

7. Miners, J. O., & Birkett, D. J. (1998). Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *Br J Clin Pharmacol*, *45*(6), 525–538.

8. Sun, H. (2012). Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data, *51*(10), 2474–2481.

9. Inoue, K., Inazawa, J., Suzuki, Y., Shimada, T., Yamazaki, H., Guengerich, F. P., & Abe, T. (1994). Fluorescence in situ hybridization analysis of chromosomal localization of three human cytochrome P450 2C genes (CYP2C8, 2C9, and 2C10) at 10q24.1. *The Japanese Journal of Human Genetics*, *39*(3), 337–43.

10. Rettie, A. E., & Jones, J. P. (2005). Clinical and toxicological relevance of CYP2C9: drug-drug interactions and pharmacogenetics. *Annual Review of Pharmacology and Toxicology*, *45*, 477–94.

11. Bourrié, M., Meunier, V., Berger, Y., & Fabre, G. (1999). Role of cytochrome P-4502C9 in irbesartan oxidation by human liver microsomes. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *27*(2), 288–96.

12. Kusama, M., Toshimoto, K., Maeda, K., Hirai, Y., Imai, S., Chiba, K., … Sugiyama, Y. (2010). In Silico Classification of Major Clearance Pathways of Drugs with Their Physiochemical Parameters. *Drug Metabolism and Disposition*, *38*(8).

13. Terfloth, L., Bienfait, B., & Gasteiger, J. (2007). Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *Journal of Chemical Information and Modeling*, *47*(4), 1688–1701.

14. Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., & Tropsha, A. (2012). Does rational selection of training and test sets improve the outcome of QSAR modeling? *Journal of Chemical Information and Modeling*, *52*(10), 2570–8.

15. Li, H., Sun, J., Fan, X., Sui, X., Wang, Y., & He, Z. (2008). Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J Comput Aided Mol Des*, *22*, 843–55.

16. Maggiora, G. (2006). On outliers and activity cliffs--why QSAR often disappoints. *J Chem Inf Model*, *46*(4), 1535.

17. Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M., & Van Drie, J. (2009). Navigating structure-activity landscapes. *Drug Discov Today*, *14*(13-14), 698–705.

18. PubChem. https://pubchem.ncbi.nlm.nih.gov/

19. Breiman, L. (2001). Random Forests. Mach Learn., *41*, 5–32.

20. Kovatcheva, A., Golbraikh, A., Oloff, S., Feng, J., Zheng, W., & Tropsha, A. (2005). QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. *SAR QSAR Environ Res.*, *16*, 93–102.

21. Kovatcheva, A., Golbraikh, A., Oloff, S., Xiao, T., Zheng, W., & Wolschann, P. (2004). Combinatorial QSAR of ambergris fragrance compounds. *J Chem Inf Comput Sci*, *44*, 582–95.

22. Votano, J., Parham, M., Hall, L., Hall, L., Kier, L., & Oloff, S. (2006). QSAR modeling of human serum protein binding with several modeling techniques utilizing structure information representation. *J Med Chem*, *49*, 7169–81.

23. Zhang, J., Hsieh, J.-H., & Zhu, H. (2014). Profiling animal toxicants by automatically mining public bioassay data: a big data approach for computational toxicology. *PLoS One*, *9*(6), e99863.

24. Dalgaard, P. (2008). Introductory Statistics with R.

25. Kim, M., Sedykh, A., Chakravarti, S., Saiakhov, R., & Zhu, H. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches. *Pharm Res*.

26. Solimeo, R., Zhang, J., Kim, M., Sedykh, A., & Zhu, H. (2012). Predicting chemical ocular toxicity using a combinatorial QSAR approach. Chem Res Toxicol. *American Chemical Society*, *25*(12), 2763–9.

27. Li, X., Chen, L., Cheng, F., Wu, Z., Bian, H., Xu, C., … Tang, Y. (2014). In silico prediction of chemical acute oral toxicity using multi-classification methods. *Journal of Chemical Information and Modeling*, *54*(4), 1061–1069.

28. Low, Y., Sedykh, A., Fourches, D., Golbraikh, A., Whelan, M., Rusyn, I., & Tropsha, A. (2013). Integrative chemical-biological read-across approach for chemical hazard classification. *Chemical Research in Toxicology*, *26*(8), 1199–208.

29. Zhu, H. (2013). From QSAR to QSIIR: searching for enhanced computational toxicology models. *Methods in Molecular Biology (Clifton, N.J.)*, *930*, 53–65.

30. Zhu, H., Ye, L., Richard, A., Golbraikh, A., Wright, F. a., Rusyn, I., & Tropsha, A. (2009). A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environmental Health Perspectives*, *117*(8), 1257–1264.