High Confidence Predictions of Drug−Drug Interactions: Predicting Affinities for Cytochrome P450 2C9 with Multiple Computational Methods

Matthew G. Hudelson,*,[†] Nikhil S. Ketkar,[‡] Lawrence B. Holder,*,[‡] Timothy J. Carlson,[§] Chi-Chi Peng,[∥] Benjamin J. Waldher,[‡,∥] and Jeffrey P. Jones*,[∥]

Department of Mathematics, Washington State University, Post Office Box 643113, Pullman, Washington 99164-3113, School of Electrical Engineering and Computer Science, Washington State University, Post Office Box 642752, Pullman, Washington 99164-2752, Department of Pharmacokinetics and Drug Metabolism, Amgen, 1120 Veterans Boulevard, South San Francisco, California 94080, and Department of Chemistry, Washington State University, Post Office Box 644630, Pullman, Washington 99164-4630

Four different models are used to predict whether a compound will bind to 2C9 with a K_i value of less than 10 μM. A training set of 276 compounds and a diverse validation set of 50 compounds were used to build and assess each model. The modeling methods are chosen to exploit the differences in how training sets are used to develop the predictive models. Two of the four methods develop partitioning trees based on global descriptions of structure using nine descriptors. A third method uses the same descriptors to develop local descriptions that relate activity to structures with similar descriptor characteristics. The fourth method uses a graph-theoretic approach to predict activity based on molecular structure. When all of these methods agree, the predictive accuracy is 94%. An external validation set of 11 compounds gives a predictive accuracy of 91% when all methods agree.

Introduction

Identifying drug−drug interaction potential early in drug discovery and development is important because drug−drug interactions can cause life threatening changes in drug levels. Early discovery of potential drug−drug interactions for a compound expedites the decision to eliminate that compound from consideration, thus lowering the cost of drug discovery. Virtual drug screening allows for the prediction of binding affinity prior to synthesis, and if prediction can be trusted, it can guide the drug discovery process. To date no single computational method has proven to be outstanding in this regard, and virtual screening still has not replaced in vitro screening methods or been routinely used in drug design. Drug interaction sites related to metabolism include UDP-glucuronosyltransferase, sulfotransferases, aldehyde oxidase, and the cytochrome P450 enzymes. Because a number of drug−drug interactions are observed for the cytochromes P450, affinity models have been developed for these enzymes. In particular, models have been developed for the three major drug metabolizing enzymes 2C9, 2D6, and 3A4.[1–3] These metabolic enzymes have broad overlapping substrate specificity, most often coupled with relatively low affinity. Almost all substrates are also competitive inhibitors of the enzyme that metabolizes them, however, not all inhibitors are substrates. The general rule for drug−drug interactions is that compounds that have K_i values, uncorrected for protein binding, greater than 10 μM are unlikely to exhibit important clinical drug-drug interactions. Since most compounds have K_i values higher than 10 μM, including

estimates of drug affinity in the early stages of drug design should reduce drug-drug interactions, while not significantly limiting the library of lead compounds for a given therapeutic target.

Many methods have been used to predict the affinity of a substrate for a given drug metabolizing enzyme. The most accurate of these models are three-dimensional (3D) models, which overlay a number of known inhibitors of a given enzyme. We and others have developed a number of these models.[4–7] An added advantage of 3D quantitative structure–activity relationship (QSAR[a]) models is that they provide an understanding of the molecular features important in binding. These methods have very high resolution when applied to related compounds, and the differences in affinity that can be predicted are much smaller than when docking to protein crystal structures is used. This high resolution is achieved by the cancelation of errors associated with the overlapping structures, leading to prediction of binding energy differences of less than 1 kcal/mol. While free energy perturbation methods can approach this type of resolution in well-defined crystal structures, the cost in computational resources is very high.[8] Other methods, such as that of Aqvist, have been applied to the cytochrome P450 enzymes[9] with good success and are much faster. Obviously, docking experiments and molecular dynamics approaches provide much more information about how a substrate binds and can provide valuable information on how to redesign a chemical to have lower affinity. Thus, each of these methods provides unique information, while the ligand-based models provide the highest throughput.

While 3D methods have specific advantages over 2D models, they suffer from the potential for multiple binding modes and in general are not easily adapted to very high-throughput. Methods based on crystal structures assume that the protein structure remains the same for each substrate. This is not true

---

* To whom correspondence should be addressed. Tel.: 509-335-5983(J.P.J., for general methods). Fax: 509-335-8867 (J.P.J.). E-mail: jpj@wsu.edu (J.P.J.); Tel.: 509-335-6138, Fax: 509-335-3818, E-mail: holder@wsu.edu (L.B.H., for SUBDUE); Tel.: 509-335-3125, Fax: 509-335-1188, E-mail: mhudelson@wsu.edu (M.G.H., for LWRP, NERP, Gravity).

† Department of Mathematics, Washington State University.

‡ School of Electrical Engineering and Computer Science, Washington State University.

§ Department of Pharmacokinetics and Drug Metabolism, Amgen.

∥ Department of Chemistry, Washington State University.

[a] Abbreviations: QSAR, quantitative structure–activity relationship; 2C9, cytochrome P450 2C; LWRP, line walking recursive partitioning; NERP, normal equation recursive partitioning; SUBDUE, Substructure Discovery Using Examples.

for 2C9, which appears to undergo major conformation changes that differ based on the substrate used.[10,11] In contrast, 2D models do not assume the substrate binds in a specific orientation, are high-throughput, and are not sensitive to protein conformational change. The major problems with 2D models are their inability to give information about protein structure and their dependence on the training set that can lead to overfitting, making a model seem statistically better than it actually is.

The utility of using multiple predictive methods has been demonstrated by De Groot and co-workers as a way to overcome the limitations of the individual methods.[3] They used a combination of two neural networks and one Bayesian model to increase the predictive capacity for 2D6 inhibition. When two models were combined, an impressive increase in correct classification resulted.

While all enzymes involved in the metabolism of drugs have the potential for drug−drug interactions, three P450 enzymes, 2C9, 2D6, and 3A4, account for the majority of drug metabolism and have been the major focus in understanding drug metabolism at the metabolic level.[12] Cytochrome P450 2C9 (2C9) is one of the major P450 enzymes responsible for the hepatic clearance of around 15% of clinically relevant drugs.[13] Given that 2C9 plays such an important role in drug metabolism and binds a number of compounds with high affinity, it is not surprising that drug−drug interactions occur for compounds metabolized by this enzyme.[12] We describe herein the use of four different models to predict whether a compound will bind to 2C9 with a K_i value of less than 10 μM. The modeling methods were chosen to maximize the differences in how training sets are used to develop the predictive models. Two of the four methods (line walking recursive partitioning (LWRP) and normal equation recursive partitioning (NERP)) develop partitioning trees based on global descriptions of structure using nine descriptors. A third method (we term the gravity method) uses the same descriptors to develop local descriptions that relate activity to structures with similar descriptor characteristics. The fourth method (SUBDUE) uses a graph-theoretic approach to predict activity based on molecular structure. When all of these methods agree, the predictive accuracy is 94%.

Methods

Inhibition constants for compounds not reported in the literature were determined as follows: Incubations were performed with 0.5 pmol 2C9/100 μL incubation for 10 min using diclofenac as a substrate probe. Diclofenac concentrations used were 1, 2, and 5 μM, which encompassed a range of approximately 0.5 times Km to 2 times Km. Inhibitors were incubated at five concentrations. The amount of 4′-OH diclofenac was assessed by LC/MS analysis to determine the rate of 4′-OH diclofenac formation at each combination of substrate and inhibitor cocentration, and K_i values were determined by nonlinear regression to Michaelis–Menten equation for competitive inhibition. 2C9 supersomes were obtained from BD Gentest.

Description of the Four Computational Methods for Developing Models. The LWRP, NERP, and gravity methods all rely on treating compounds as points in an m-dimensional "descriptor space", where m is the number of descriptors. For each compound c_i in the training set, we have a corresponding point $\mathbf{r}_i$ and its corresponding experimental value $g(\mathbf{r}_i)$, treated as a binary value (in this case, -1 if $pK_i < 5.0$ and $+1$ if $pK_i \geq 5.0$.) When a prediction is desired, the compound under investigation corresponds to another point $\mathbf{y}$ in the descriptor space. The compound's behavior is predicted based on where $\mathbf{y}$ resides relative to the points $\mathbf{r}_i$ corresponding to the compounds in the training set.

In contrast to the three descriptor-based methods, the graph-theoretic structure method treats compounds as graphs whose vertices and edges represent atoms, bonds, and their properties. The training set is analyzed as a collection of graphs; similarities among graphs in either subset are catalogued and then form the basis for prediction.

Line Walking Recursive Partitioning (LWRP). This method is described in detail in[1] and we give a synopsis here. Recursive partitioning as described in refs 1 and 14 consists of finding a hyperplane that partitions the space containing the training set into two pieces. Then the process continues by splitting each of the two pieces by a hyperplane and so on, until the training set has been partitioned into subsets where all of the points in a subset correspond to compounds whose experimental values $p(\mathbf{r}_i)$ are either all above or all below a predetermined threshold value. The goal is to make the final number of subsets small. The "line walking" portion of the algorithm described below was devised as a means of producing a hyperplane that splits a set of points into two more homogeneous sets with regard to the threshold value.

Recalling that there are m descriptors being used, the initializing step in choosing a splitting plane for a set S is to choose a set of vectors $R = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m\}$ from S at random.

Given an m-element subset $R = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m\}$ of S, the entire training set, a single iteration of the line-walking algorithm, consists of the following steps: Compute the vector $\mathbf{p}$ such that $\mathbf{p} \times \mathbf{r}_i = 1$ for all $\mathbf{r}_i$ in R; choose a value $\mathbf{r}_k$ at random from R' and delete it from R; form the matrix M whose rows are the remaining vectors in R; and find a solution distinct from $\mathbf{p}$ to the linear system $M\mathbf{q} = 1$. Ideally, this system will have a 1D solution space, but this is not required. In the rare situation the system is inconsistent, we solve the normal equation: $M^T M\mathbf{q} = M^T\mathbf{1}$.

Defining $\mathbf{L}(t) = t\,\mathbf{q} + (1 - t)\mathbf{p}$, determine for each $\mathbf{r}_s$ in S the value t_s such that $\mathbf{L}(t_s) \times \mathbf{r}_s = 1$. $\mathbf{L}$ is the "line" mentioned in the name of the algorithm. If no such value t_s exists, then the corresponding $\mathbf{r}_s$ is disregarded for steps 5 and 6 of the algorithm.

$f(\mathbf{L}(t_s))$ is maximized, where f is some objective function that measures how well $\mathbf{L}(t_s)$ splits the set R, and $\mathbf{r}_k$ is replaced with $\mathbf{r}_s$ in R. The new vector $\mathbf{p}$ is equal to $\mathbf{L}(t_s)$, so the next iteration begins at step 2.

There are several possibilities for conditions to halt the algorithm. The halting criterion chosen early in this research was if the maximized value of f remains unchanged for a predetermined number of consecutive iterations. Later, it was decided to permute the vectors in R and adopt each in succession as $\mathbf{r}_k$ in step 2 if the maximum value of f remained unchanged. The algorithm halts if all of the vectors in R are exhausted in this manner. This condition results in locating a local maximum for f in the sense that no $\mathbf{r}_s$ in R can be substituted, resulting in raising the value of $f(\mathbf{L}(t_s))$. The final vector $\mathbf{p}$ yields the decision plane, $P = \{\mathbf{x} \in \mathbf{R}^n : \mathbf{p} \times \mathbf{x} = 1\}$ The set S is then dissected into two pieces, $S^+ = \{\mathbf{r} \in S : \mathbf{p} \times \mathbf{r} \geq 1\}$ and $S^- = \{\mathbf{r} \in S : \mathbf{p} \times \mathbf{r} < 1\}$ We then repeat the algorithm on each piece, recursively, ultimately producing a "prediction tree".

Because the line-walking algorithm incorporates random decisions, a variety of decision trees can be produced from the same set S. This enables us to make consensus predictions based upon the outcomes of a simple majority of the trees.

Normal Equation Recursive Partitioning (NERP). This algorithm is like LWRP in that it also produces a recursive partitioning of chemical space by means of decision planes. Here, given a set of points, $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$, we find the hyperplane P that partitions S by the following steps: compute the vector $\mathbf{v} = (\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_n)$, where v_i is the mean of the values of the ith descriptor; form the $k \times n$ matrix N, whose i^{th} row is $\mathbf{s}_i - \mathbf{v}$; form the k-vector $\mathbf{g}$, whose i^{th} entry is 1 if the ith compound in the test set's experimental value is at or above the threshold and -1 otherwise; find a solution $\mathbf{p}$ to the normal equation $N^T N\mathbf{p} = N^T\mathbf{g}$.

Because this algorithm is deterministic, every NERP tree built from the set S is identical. To build a forest of different NERP trees upon which to base a consensus prediction, we use randomly generated subsets R of S to construct our trees.

Gravity. While LWRP and NERP partition the entire descriptor space into discrete regions of similar behavior, the gravity method is a "nearest-neighbor" type of algorithm that produces a real-valued
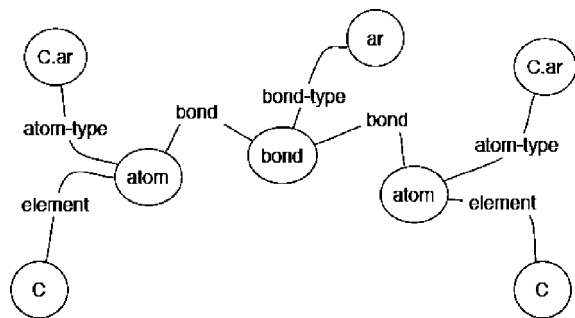
Figure 1. Graph representation example of a compound.

continuous prediction. The "gravity interpolation" algorithm begins with a collection of distinct points, called the training set, in some vector space, and a value $p(x)$ assigned to each point x. The predicted value $p(y)$ of any point y in the space is computed as a weighted average, $p(y) = \sum_i w_i p(x_i) / \sum_i w_i$, where the weights are given by the formula $w_i = 1/|y - x_i|^k$, $k > 0$; the $|y - x_i|$ term in the denominator denotes the Euclidean distance between x_i and y. When y is near one of the points x_i in the training set, then the weight assigned to this point overwhelms the other weights and so the prediction $p(y)$ is very close to $p(x_i)$. Adopting the convention that if $y = x_i$, then $p(y) = p(x_i)$ makes this process a continuous function over the vector space.

The parameter k can be modified to adjust the smoothness of the gradients among the points in the training set. As k is made very large, the prediction $p(y)$ approaches $p(x_{nearest})$, the value at the nearest point in the training set to y. When $k = 2$, then the weights are analogous to the "gravitational pull" of unit masses at the training points, hence, the name of the algorithm. Conversely, as k is made very small, the weights become close to 1 and so the prediction $p(y)$ approaches the mean of the $p(x_i)$ values.

Substructure Discovery Using Examples (SUBDUE). The SUBDUE method [15,16] differs from the previous methods in two fundamental ways. First, the data is represented as a graph, and the predictor for a class is represented as a set of subgraphs, such that if any of these subgraphs is present in a new example, then the example is predicted to belong to the class. Second, the top-level learning algorithm uses a set-covering approach, in contrast to a partitioning approach, that learns a predictor for a subset of the examples, removes this covered set from consideration, and then iterates until all of the examples are covered by at least one predictor.

In the current task of predicting the inhibitory behavior of a compound (i.e., whether its pK_i value is above or below 5.0), we represent each compound as a labeled graph, an example of which is depicted in Figure 1. Each atom is represented as a vertex-labeled "atom". An edge-labeled "element" connects the atom vertex to a vertex whose label is the atom's element. An edge labeled "atom-type" connects the atom vertex to a vertex whose label is the atom's type. Each bond is represented by a vertex labeled "bond", with two "bond"-labeled edges connecting it to the atoms involved in the bond. A third edge labeled "bond-type" connects the bond vertex to a vertex whose label is the bond's type. Alternatively, we could have used a representation in which each atom is a single vertex whose label is the atom type, and each bond is an edge labeled with the bond type and connecting the two involved atoms. However, this representation would force the predictors (subgraphs) to be based on specific atom types and bond types rather than having the ability to predict based on the mere presence of certain numbers of atoms and bonds, only some of which may need to be constrained to a specific element or type.

Given a set of positive example graphs Gp (i.e., compounds for which $pK_i \geq 5.0$) and a set of negative example graphs Gn (i.e., compounds for which $pK_i < 5.0$), the SUBDUE method proceeds as follows: (1) $H = \{\}$; (2) repeat, (a) find a subgraph g maximizing $V = (tp + tn)/(Gp + Gn)$, where tp is the number of graphs in Gp

containing g and tn is the number of graphs in Gn not containing g, (b) $H = H \cup \{g\}$, and (c) $Gp = Gp - \{$graphs in Gp containing $g\}$, until $Gp = \{\}$; (3) return H.

The search for the error-minimizing subgraph in step 2a is conducted using a heuristic search that finds a local minimum in the error $(1 - V)$. Candidate subgraphs are built starting from single vertices and expanded one edge at a time during the search based on edges existing in the data; therefore, only graph isomorphism tests (and not NP-complete subgraph isomorphism tests) are needed to determine the graphs in Gp and Gn containing g. The graph isomorphism test is constrained to run in time polynomial in the size of the graph. The returned hypothesis H is a disjunction of subgraphs, such that H predicts a graph is a member of the positive class if that graph contains at least one of the subgraphs in H. Because the subgraphs added to H in later iterations are likely to be overly specific (e.g., matching only one positive example), a stopping criterion based on tp or a limit on the size of H may be employed.

For SUBDUE we used a stopping criterion of $|H| \leq 5$, which was determined based on minimizing the error of a 10-fold cross-validation on the training set. We then ran SUDBUE on the entire training set to produce the predictor H, which was then used to classify the examples in the validation set. We used version 5.2 of SUBDUE, written in C, and available at www.subdue.org.

For the LWRP, NERP, and Gravity methods, we used nine descriptors, described in Table 1 [1] and a training set of 276 compounds whose pK_i values ranged from 1.9 to 7.6 and a validation set of 50 compounds, 23 of which had pK_i values below 5.0 and 27 of which had pK_i values above 5.0. For LWRP, we made our consensus predictions using 101 trees based on the entire 276 compound training set. For NERP, we also built 101 trees, but each was based on a randomly generated subset of 125 compounds from the training set. For the gravity method, we used the parameter $k = 4$. Computations for LWRP, NERP, and gravity were all done using the 2006.08 version of Molecular Operating Environment software.

Training and Validation Set Selection. The 326 compound database with Smiles string structures, training sets, validation sets, and descriptors for 2C9 inhibitors can be obtained at seeker.wsu.edu as excel spreadsheets. This database was split into a training set of 276 compounds and a diverse validation set of 50 compounds. The 50 compound validation set was obtained by ranking entries based on structural fingerprints with the distance being calculated by the Tanimoto Coefficient similarity matrix. The most structurally diverse subset was constructed using this metric. Other methods for selecting a diverse training set were explored with LWRP and very similar results were obtained (data not shown).

Results

We summarize the performance of the various methods in Tables 2 and 3.

In Table 2, the value λ is the Matthews coefficient, computed by the formula

$$\lambda = \frac{t_+ t_- - f_+ f_-}{\sqrt{(t_+ + f_+)(t_+ + f_-)(t_- + f_+)(t_- + f_-)}}$$

The Matthews coefficient λ has several useful properties. It can be shown that $-1.0 \leq \lambda \leq 1.0$, where a value of 1.0 indicates a perfect predictor and a value of -1.0 indicates a perfect antipredictor. Finally, if predictions are based solely on a random assignment, then the expected value of λ is 0.0.

Table 3 lists the breakdown of how well the four methods predicted the behavior of the 50 compounds in the validation set. All four methods correctly predicted the behavior of 31 of the 50 compounds and three of the four methods correctly predicted the behavior of 11 of the remaining 19 compounds. Thus, 42 out of 50 compounds were correctly identified by a majority of the four methods. Of the eight remaining com-

Table 1. Molecular Descriptors Used in LWRP, NERP, and Gravity Methods

descriptor	synopsis
vsa_hyd	The approximation to the sum of VDW surface areas of hydrophobic atoms.
vdw_vol	The van der Waals volume calculated using a connection table approximation.
apol	The sum of the atomic polarizabilities (including implicit hydrogens) with polarizabilities taken from ref *17*.
vdw_area	The area of van der Waals surface calculated using a connection table approximation.
weinerPol	Wiener polarity number: half-the sum of all of the distance matrix entries with a value of 3 as defined in ref *18*.
PEOE_VSA_NEG	The total negative van der Waals surface area; this is the sum of the v_i [b] such that q_i is negative.[a]
Zagreb	Zagreb index: the sum of d_i[b] over all heavy atoms i.[c]
SlogP	The log of the octanol/water partition coefficient (including implicit hydrogens).
bpol	The sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens) with polarizabilities taken from ref *17*.

[a] The variable q_i denotes the partial charge on atom i. [b] The v_i value denotes the accessible van der Waals surface area of atom i calculated from a connection table approximation. [c] The d_i value is defined as the number of heavy atoms to which atom i is bonded.

Table 2. Comparison of Performance among LWRP, NERP, Gravity, and SUBDUE Predictive Methods[a]

	LWRP	NERP	Gravity	SUBDUE
t_+	23	24	22	21
t_-	21	21	21	16
f_+	2	2	2	7
f_-	4	3	5	6
λ	0.762	0.800	0.726	0.475

[a] The variable t_+ denotes the number of "true positives", that is, those compounds that were correctly predicted to have $pK_i \geq 5.0$; t_- denotes the number of "true negatives"; f_+ denotes the number of "false positives", those compounds incorrectly predicted to have $pK_i \geq 5.0$; and f_- denotes the number of "false negatives". The variable λ denotes the Matthews coefficient.

Table 3. Aggregate Performance among LWRP, NERP, Gravity, and SUBDUE Predictive Methods

No. of the four methods producing correct predictions	No. of cmpds (out of 50)
4	31
3	11
2	4
1	2
0	2

pounds, four were evenly split among the four methods, two were incorrectly predicted by three of the methods, and two were incorrectly predicted by all four methods. In Tables 4 and 5 we show structures, experimental pK_i values, and results from the four methods on each of the eight underperforming compounds. In these tables, a "+" sign indicates a predicted pK_i value of at least 5.0 and a "−" sign indicates a predicted pK_i value below 5.0.

We note that three of these eight underperforming compounds (including the two worst performers) have experimental pK_i values within 0.2 units of the threshold value of 5.0.

Discussion

The computational prediction of whether a compound will bind tightly to a drug metabolizing enzyme can be an important tool in drug design. If high affinity compounds could be identified prior to synthesis, efforts could be focused on compounds less likely to have adverse effects via drug-drug interactions, decreasing the time required to develop a drug for a target. Unfortunately, most methods for predicting affinity are not accurate enough to convince the medicinal chemist that synthesis and testing are not appropriate. Herein we present four distinct modeling methods that when combined provide predictions that are as accurate as in vitro testing for predicting if a compound will have high affinity (<10 μM) for 2C9.

We used a number of common metrics to determine if a model is accurate at predicting if a validation set of 50

compounds are tight binding inhibitors of 2C9. "Concordance" denotes the percentage of compounds correctly predicted, "Specificity" denotes the percentage of weak binding compounds correctly predicted, "Sensitivity" denotes the percentage of tight binding compounds correctly predicted, and λ is the Matthews coefficient described in the methods section.
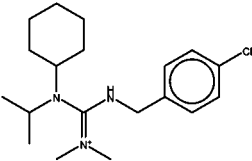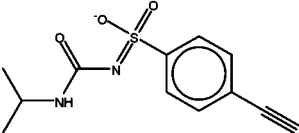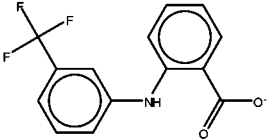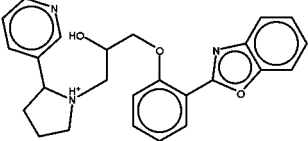
The LWRP method has been used in the past to predict if compounds are 2C9, 3A4, and 2D6 inhibitors at any concentration.[1] While this is useful information, the main reason to understand affinity is to predict if a compound will be a substrate or inhibitor. To determine if LWRP can predict if a compound will be a high affinity inhibitor of 2C9 we used a database of 326 compounds for which we have measured K_i values in the past. We selected 50 compounds from the 326 to act as a validation set and used the rest to train the LWRP model. As shown in Table 3, LWRP predicts 44 of the 50 compounds correctly to give a concordance of 88%. Out of the 27 compounds that bind with a K_i value of 10 μM or lower, LWRP predict 23 correctly, which corresponds to a sensitivity of 85%. Out of 23 compounds that bind with a K_i of over 10 μM, 21 are correctly predicted for a specificity of 91%.

NERP is a new method for dividing compounds into inhibitors and noninhibitors, related to LWRP, but with a significant savings in computing time; the details of the methods are presented in the computational methods section. As shown in Table 3, NERP predicts 45 of the 50 compounds correctly to give a concordance of 90%. Out of the 27 compounds that bind with a K_i value of 10 μM or better, NERP predicts 24 correctly, which corresponds to a sensitivity of 89%. Out of 23 compounds that bind with a K_i of over 10 μM, 21 are correctly predicted for a specificity of 91%.

Gravity is a new method for recursive partitioning that makes decisions based on a tree weighted by compounds that occupy a similar descriptor space. While LWRP and NERP attempt to make more global decisions, Gravity is essentially a nearest neighbor method. As shown in Table 3, Gravity predicts 43 of the 50 compounds correctly to give a concordance of 86%. Out of the 27 compounds that bind with a K_i value of 10 μM or better, Gravity predict 22 correctly, which corresponds to a sensitivity of 81%. Out of 23 compounds that bind with a K_i of over 10 μM, 21 are correctly predicted for a specificity of 91%.

SUBDUE is a totally different method, which has been used in the past for classification in several structural domains, including mutagenicity of compounds[19] and function of metabolic pathways.[20] As shown in Table 3, SUBDUE predicts 37 of the 50 compounds correctly to give a concordance of 74%. Out of the 27 compounds that bind with a K_i value of 10 μM or better, SUBDUE predicts 21 correctly, which corresponds to a sensitivity of 77%. Out of 23 compounds that bind with a K_i of over 10 μM, 16 are correctly predicted for a specificity of 69%.

Table 4. Compounds Correctly Predicted by Exactly Two out of the Four Predictive Methods

Compound	pK_i	NERP	LWRP	Gravity	SUBDUE
	1.7017	+	−	+	−
	4.3188	−	+	−	+
	5.0641	+	+	−	−
	5.2676	+	+	−	−

Interestingly, only two compounds out of the 50 compound training set are incorrectly predicted by all methods. This indicates that the methods produce models with different weaknesses and that combinations of the four models should have better predictive capacity than any individual model. All four models agree for 33 out of 50 compounds or 66% of the time. When the models agree, the prediction is correct 94% of time. The specificity and sensitivity are 92 and 95%, respectively. The two incorrectly predicted compounds are within 0.15 log units of the cutoff point for being a strong inhibitor (<10 μM), having K_i values of 14 and 8 μM. Given the normal uncertainty in K_i measurements, these values are not distinguishable from each other or the cutoff. Thus, for a large fraction (66%) of the compounds in the validation set, an extremely high degree of predictability can be obtained, and the results are essentially the same as determining the in vitro K_i value. Given this result if all four models agree that a compound will bind with a K_i value of less than 10 μM, this compound can confidently be removed from the set of compounds to be synthesized.
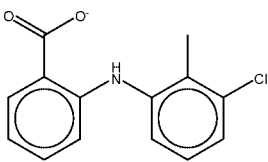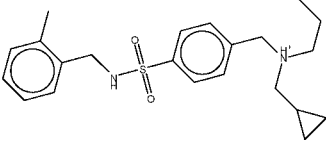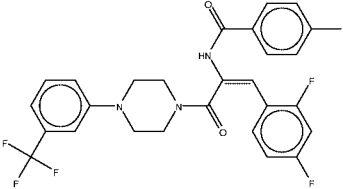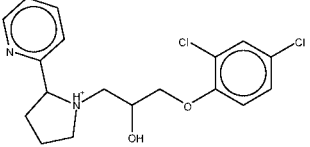
While using all four models gives very high confidence in the prediction, it could be argued that not getting a prediction for 33% of the compounds is too low a coverage for early predictive methods. We can increase coverage to 92% by scoring based on 3 out of four methods. Using three out of four of the models to get a prediction gives a correct prediction frequency of 91%. Out of the 24 compounds that bind with a K_i value of 10 μM or lower and that a prediction can be made, 21 are correctly predicted, which corresponds to a sensitivity of 87%.

Out of 22 compounds that bind with a K_i of over 10 μM, 21 are correctly predicted for a specificity of 95%.

After developing the models, we synthesized a new series of tight binding inhibitors related to benzbromarone analogs we synthesized earlier,[7] but with the benzofuran bicyclic ring system replaced with a chromanone bicyclic ring system (Figure 2). Details of the synthesis and characterization of binding will be presented in a separate manuscript. These compounds provide an external validation data set of 11 compounds that have no analogs in the training set. They also provide an example of compounds that bind so tightly to 2C9 that they would exhibit drug−drug interactions. All the available 327 molecules were used to develop new models using each method. When these models are used, all four of methods, LWRP, NERP, Gravity, and SUBDUE, predicted that 10 of the 11 compounds had K_i values below 10 μM. For the remaining compound, all methods but Gravity predicted that the K_i would be below 10 μM. We remark that SUBDUE identified all 11 compounds as binders because all of them contain the substructure (reported as subgraph "C.2-1-C.2-2-C.2-1-O.3") shown in Figure 3. None of the carbon atoms in this substructure are aromatic.

The results indicate that 11 out of 11 compounds would have high potential for drug−drug interactions by inhibiting 2C9, as predicted by using the consensus of three out of four models. If all four methods are required to agree for a prediction, 10 of 11 or a 91% correct prediction frequency is obtained, identical to the prediction frequency of the diverse validation set. This is an interesting result given that no chromenone analogs are in the data set. To see if this prediction accuracy was a result

Table 5. Compounds Correctly Predicted by Fewer than Two out of the Four Predictive Methods
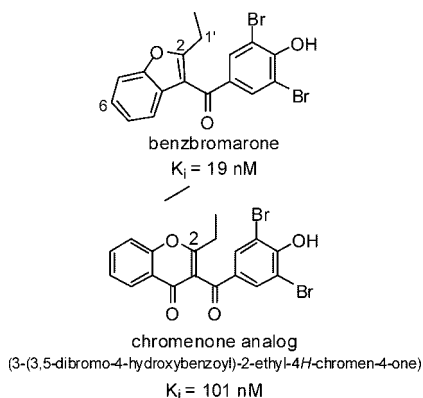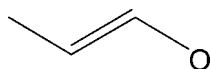
Compound	pK_i	NERP	LWRP	Gravity	SUBDUE
(chemical structure)	5.7305	---	---	+	---
(chemical structure)	5.8386	---	+	---	---
(chemical structure)	4.8520	+	+	+	+
(chemical structure)	5.1361	---	---	---	---

of the structure sharing common features with the tight binding benzbromarone compounds, we used Gravity to look at the three nearest neighbors in descriptor space for each of the 11 compounds. Four of the compounds did not have any benzbromarone analogs as the three nearest neighbors, while six others only had one benzbromarone as a nearest neighbor, the remaining compound had two benzbromarone analogs as nearest neighbors. Because Gravity was designed to be a nearest neighbor prediction method, these results indicate that the high predictive capacity for this external validation set is not just a result of close similarity of these compounds to the benzbromarone analogs in the training set. However, it should be noted that all these compounds are relatively tight binding inhibitors, and this external validation set does not test the models ability to predict diverse poor binders. While 27 compounds in the validation set of 50 are weak binders, many are related to structures in the training set.

Conclusions

Four different methods were used to predict if a diverse test set of 50 molecules and an external validation set of 11 tight binding inhibitors of 2C9 are likely to cause drug−drug interactions based on binding affinity. While each individual method had reasonable predictive accuracy, combining the models proved to be highly predictive, as was observed using other predictive methods by O'Brien and de Groot.[3] The best approach in terms of structural coverage and predictive accuracy appears to be to use the consensus of 3 out of 4 of the models. Thus, when any three models agree on the drug−drug interaction potential, the prediction is over 90% accurate. Most of the outliers are found to be relatively close to the 10 μM cutoff, indicating that the methods are capturing the relative affinities of the molecules. Given the high accuracy of the method, at least with the compounds tested, it appears that decisions on scaffolds for lead compound synthesis can be made with the

benzbromarone
K_i = 19 nM

chromenone analog
(3-(3,5-dibromo-4-hydroxybenzoyl)-2-ethyl-4*H*-chromen-4-one)
K_i = 101 nM

Figure 2. Benzbromarone and chromenone analog.

Figure 3. Substructure in common to all 11 chromenone analogs.

combined models. Of course, if the structures show a large variation from the training set, the predictive capacity would be reduced. In this case, the models may have to be retrained to include representative compounds from this new structural space.

References

(1) Hudelson, M. G.; Jones, J. P. Line-walking method for predicting the inhibition of p450 drug metabolism. *J. Med. Chem.* **2006**, *49*, 4367–4373.

(2) Ekins, S.; De Groot, M. J.; Jones, J. P. Pharmacophore and three-dimensional quantitative structure−activity relationship methods for modeling cytochrome P450 active sites. *Drug Metab. Dispos.* **2001**, *29*, 936–944.

(3) O'Brien, S. E.; de Groot, M. J. Greater than the sum of its parts: Combining models for useful ADMET prediction. *J. Med. Chem.* **2005**, *48*, 1287–1291.

(4) Jones, J. P.; He, M. X.; Trager, W. F.; Rettie, A. E. Three-dimensional quantitative structure−activity relationship for inhibitors of cytochrome P450 2C9. *Drug Metab. Dispos.* **1996**, *24*, 1–6.

(5) Haining, R. L.; Jones, J. P.; Henne, K. R.; Fisher, M. B.; Koop, D. R. Enzymatic determinants of the substrate specificity of CYP2C9: Role of B′−C loop residues in providing the pi-stacking anchor site for warfarin binding. *Biochemistry* **1999**, *38*, 3285–3292.

(6) Rao, S.; Aoyama, R.; Schrag, M.; Trager, W. F.; Rettie, A. A refined three-dimensional QSAR of cytochrome P450 2C9: Computational predictions of drug interactions. *J. Med. Chem.* **2000**, *43*, 2789–2796.

(7) Locuson, C. W.; Rock, D. A.; Jones, J. P. Quantitative binding models for CYP2C9 based on benzbromarone analogues. *Biochemistry* **2004**, *43*, 6948–6958.

(8) Jones, J. P.; Trager, W. F.; Carlson, T. J. The binding and regioselectivity of reaction of (*R*)- and (*S*)-nicotine with cytochrome P-450cam: Parallel experimental and theoretical studies. *J. Am. Chem. Soc.* **1993**, *115*, 381–387.

(9) Szklarz, G. D.; Paulsen, M. D. Molecular modeling of cytochrome P450 1A1: Enzyme-substrate interactions and substrate binding affinities. *J. Biomol. Struct. Dyn.* **2002**, *20*, 155–162.

(10) Wester, M. R.; Yano, J. K.; Schoch, G. A.; Yang, C.; Griffin, K. J.; et al. The structure of human cytochrome P4502C9 complexed with flurbiprofen at 2.0-angstrom resolution. *J. Biol. Chem.* **2004**, *279*, 35630–35637.

(11) Dickmann, L. J.; Locuson, C. W.; Jones, J. P.; Rettie, A. E. Differential roles of Arg97, Asp293, and Arg108 in enzyme stability and substrate specificity of CYP2C9. *Mol. Pharmacol.* **2004**, *65*, 842–850.

(12) Rettie, A. E.; Jones, J. P. Clinical and toxicological relevance of CYP2C9: Drug-drug interactions and pharmacogenetics. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 477–494.

(13) Evans, W. E.; Relling, M. V. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* **1999**, *286*, 487–491.

(14) Susnow, R. G.; Dixon, S. L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1308–1315.

(15) Cook, D. J.; Holder, L. B.; Su, S. B.; Maglothin, R.; Jonyer, I. Structural mining of molecular biology data. *IEEE Eng. Med. Biol.* **2001**, *20*, 67–74.

(16) Cook, D.; Holder, L. *Mining Graph Data*; John Wiley and Sons: New York, 2006.

(17) *CRC Handbook of Chemistry and Physics*; CRC Press: Boca Raton, FL, 1994.

(18) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 299–404.

(19) Gonzalez, J.; Holder, L.; Cook application of graph-based concept learning to the predictive toxicology domain. In *Proceedings of the Predictive Toxicology Challenge Workshop*, Freiburg, Germany, 2001.

(20) You, C.; Holder, L.; Cook, D. Application of graph-based data mining to metabolic pathways. *Workshop on Data Mining in Bioinfomratics, IEEE International Conference on Data Mining*, December 18−22, 2006, Hong Kong, 2006.

JM701130Z