

Building Predictive Models for Protein Tyrosine Phosphatase 1B Inhibitors Based on Discriminating Structural Features by Reassembling Medicinal Chemistry Building Blocks

Chihae Yang,^{*,†} Kevin Cross,[†] Glenn J. Myatt,[†] Paul E. Blower,[†] and James F. Rathman[‡]

Leadscope, Inc., 1393 Dublin Road, Columbus, Ohio 43215, and Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio 43210

Received April 12, 2004

A new approach to predicting the biological activity of small molecule pharmaceuticals is demonstrated. Structural features of medicinal chemistry building blocks are used as 2-D molecular descriptors. These descriptors include predefined structural features and macrostructures obtained from a supervised process in which features in the core library are reassembled to provide larger features that strongly differentiate the desired biological response variable. Chemical features derived in this manner can serve as predictor variables for diverse modeling algorithms, and application using partial least squares techniques is demonstrated here. Models are presented for inhibition by benzofuran and benzothiophene biphenyl analogues of protein tyrosine phosphatase 1B (PTP1B), a target for insulin-resistant disease states. Results are compared to models for PTP1B inhibitors available in the literature based on CoMFA-related techniques and 3-D molecular descriptors.

Introduction

Designing small molecules for desired pharmacological activity requires identification of compound classes based on structural features and methods to estimate the activities from such hypotheses. The ability to accurately predict biological responses from structural features of a chemical compound is one of the most important aspects in the drug discovery and optimization processes. QSAR model development based on molecular descriptors has been extensively studied.^{1–3} Most QSAR descriptors are derived either from energetics based on quantum mechanical or semiempirical calculations or from topological indices. Three-dimensional molecular descriptors are fundamental and give physical meaning to models; however, quantitative interpretation of their contributions is generally not straightforward and is often computationally prohibitive for large structure sets. Two-dimensional descriptors have been used in similarity searching and structure-based clustering for grouping similar compounds.^{4,5} Yet, despite their computational advantages, 2-D structural features have not been widely reported for the quantitative modeling of biological activities.

This paper describes a new methodology in which structural features of medicinal chemistry building blocks are used as a basis for 2-D molecular descriptors. In our recent paper, a new method to find discriminating structural features was presented. This methodology enables the building of predictive models based on structural hypotheses.⁶ These molecular descriptors include predefined structural features and predictive macrostructures. The macrostructures are reassembled from the original predefined features, guided by the

response variables to differentiate structures based on biological responses. These predictive features are optimized such that compounds containing the macrostructures tend to impart higher or lower activity compared to the mean of the training set. The predictive accuracy of these macrostructures can be assessed when the model is evaluated.

Chemical features derived in this manner can serve as predictor variables for many diverse modeling algorithms including multivariate least squares regression, principal component regression (PCR), *k* nearest neighbors (kNN), partial least squares (PLS) techniques, and neural network approaches. Although most of the fitting algorithms provide methods to reduce the high dimensionality of chemical feature space, preselection of descriptors is still key to building useful predictive models. Genetic algorithms, simulated annealing, recursive partitioning, and principal component analysis (PCA) are popular methods for descriptor selection. In this paper, we show that macrostructure assemblies (MSAs) are an intuitive and chemically relevant approach for reducing the dimensionality of the feature space. In principal component analysis, correlations and redundancy between features are removed by maximizing the variance captured by latent variables. The resulting principal components are linear combinations of the original predictors (structural features), which represent individual medicinal chemistry building blocks. This mathematical treatment is conceptually analogous to the process of reassembling the macrostructures from the individual features. In comparison to principal components, MSAs offer two advantages: (1) the connectivity of individual features is intrinsically included in structures; (2) the reassembly process is supervised by the selected biological response. Since MSAs represent actual physical structures, they are easier to

* Author to whom correspondence should be addressed. Tel: 614-340-3466. Fax: 614-675-3732. E-mail: cyang@leadscope.com.

[†] Leadscope, Inc.

[‡] The Ohio State University.

understand than the abstract latent variables generated by principal component analysis.

Examples illustrated in this paper are inhibitor models of benzofuran and benzothiophene biphenyl/naphthalene analogues for protein tyrosine phosphatase 1B (PTP1B), a novel target for insulin-resistant disease states.⁷ Models for PTP1B inhibitors available in the literature apply CoMFA-related techniques,^{8,9} docking simulations,⁹ or other 3-D QSAR descriptors.¹⁰ Results from these prior studies are used for comparisons and validation of the methodology presented in this paper. Advantages of using structural features as descriptors to intuitively understand chemical inference over more traditional QSAR methods are articulated. The importance of this chemical inference is emphasized throughout the model building process. The goal of the model building exercise in discovery and optimization is ultimately to be able to design compounds having the desired activities or properties. The chemical inference needed during the structure design process requires the intuitive connection of models back to the structural building blocks used in the model.

The objectives of this paper are to (1) confirm the importance of macrostructures in a molecular descriptor set for predicting activities; (2) predict the pIC₅₀ values of the 26 compound set used in the CoMFA modeling study by Murthy;⁸ and (3) accurately predict the activity classes of a 19-compound set discussed by Malamas et al.^{11,12} but for which pIC₅₀ values were not reported.

Methods

The overall modeling strategy includes the following steps: (1) diagnose the data set; (2) assemble macrostructures with predictive accuracy; (3) select descriptors—preselection of structural features plus addition of physicochemical properties; (4) employ appropriate model building algorithms; (5) evaluate the model with chemical inference; and (6) rebuild the model by refining the feature set.

Diagnosis of PTP1B Dataset. Prior to model building, data sets were diagnosed for data distribution and structural similarities between the training and test sets. The same structure set (118 compounds) described in Cross et al.⁶ was used to illustrate this modeling methodology. The training (92 compounds) and test (26 compounds) sets were partitioned as described by Murthy et al.⁸ These data sets were used to build both quantitative pIC₅₀ (−log IC₅₀) and classification models for PTP1B inhibition activity. For the binary classification model, each compound was classified as active or inactive. Compounds having pIC₅₀ values lower than the average of the 118 compounds (mean = 0.70) were classified as inactive, resulting in a data set with equal numbers of actives and inactives. A set of 19 additional compounds described by Malamas et al., but whose IC₅₀ values were not reported in that paper, was used to further challenge models built by this approach. Both test sets (26-test and 19-unknown) were not used in the model training process. For each of these 19 compounds, Malamas reports a percent inhibition measured at one concentration; concentrations ranging from 0.1 to 2.5 μM are listed for compounds in this subset. For the analysis presented here, compounds whose reported percent inhibition was lower than 50% at a concentration higher than 0.199 μM were classified as inactive, since the average IC₅₀ value of the 118 set was 0.199 μM (pIC₅₀ = 0.702).

Structural similarities between compounds in the training and test sets are assessed by comparing compound fingerprints. The fingerprint for each compound is represented as a vector of binary (1/0) values, each element indicating whether or not the compound has a specific chemical feature. Structural features were selected from a library of 27,000 medicinal

chemistry building blocks and the reassembled macrostructures.¹³ For each compound in the test set, a mean correlation *within* the test set is calculated by averaging all pairwise Tanimoto coefficients. Similarly, pairwise Tanimoto coefficients *between* each test set compound against all compounds in the training set are calculated and then averaged. The *within* and *between* Tanimoto coefficients are then correlated. If the test and training sets are structurally similar, the correlations *within* and *between* them should be approximately the same. The ratio of the within/between average correlation provides a quantitative measure: the closer this ratio is to unity, the more structurally similar the test set is to the training set.

Macrostructure Assembly. The key step in the model building process described in this paper is the generation of MSAs, chemical scaffolds constructed by combining smaller 2-dimensional molecular descriptors. Although a brief description of this process is presented here, the detailed process for generating macrostructures for the PTP1B set has been reported in a previous paper.⁶ The use of 2-D chemical fragments as descriptor variables in QSAR models is well-established and supported by extensive prior work. In this approach each compound is viewed as a combination of chemical fragments. The MSA method takes this approach a step further by dynamically building new and larger fragments in a supervised manner; this process is dynamic in the sense that MSA construction is an integral part of the data analysis process, as opposed to the traditional approach in which fragments are selected from a preexisting library of features. The MSA method offers two very important advantages: (1) MSAs are generally larger than the 2-D descriptors available in fragment libraries and therefore properly describe connectivity within molecules to a greater extent. (2) Models that include MSAs usually require significantly fewer total descriptors than models based solely on smaller fragments. In this light, the MSA method can be viewed as a chemically intelligent dimension reduction technique both simplifying and generalizing the model.

Preselection of Descriptors. Even for a relatively small number of compounds, the total number of predictor variables (basic chemical features, MSAs, physical properties) can be very large. An important first step is to select a subset of descriptors that will be most useful in the model building process. Preselection began by removing features present in only one compound or in every compound since, even if these features happen to be important, the compound set is not sufficiently diverse to measure their influence with any statistical certainty. These initial numbers of features can be further reduced before application of model building algorithms by simple significance tests such as *t*²-tests (when response variable is continuous) and χ^2 -tests (when response is categorical). This type of feature preselection, commonly employed in most supervised methods, was used here for pIC₅₀ (*t*²-test) and activity classification (χ^2 -test) models.

Relevant physicochemical properties or QSAR descriptors can be added after preselection of features. In this study, the following molecular properties, calculated within Leadscape software, were added: aLogP,¹⁴ polar surface area,¹⁵ parent molecular weight, parent atom count, number of hydrogen bond acceptors and donors, number of rotatable bonds, and Lipinski scores.¹⁶ The values for these eight properties for the full set of 137 compounds, and an sd-file of structures, are available at [http://www.leadscape.com/downloads/data/PTP1B\(137\)_JMedChem.zip](http://www.leadscape.com/downloads/data/PTP1B(137)_JMedChem.zip).

Model Building. The nonlinear iterative partial least squares (PLS) method was used to develop a pIC₅₀ model. For activity classification model, partial logistic regression (PLR) was employed.¹⁷ PLS involves the stepwise extraction of factors (linear combinations of predictor variables) that are highly correlated with a dependent response variable. Each factor is orthogonal to all others and has an associated vector of weights that describe the contribution of each independent variable. The weight vectors are similar to eigenvectors in PCA. Multiplication of the original predictor variables by the weight

vector yields the component scores for a particular factor. In conventional PLS, after extracting a desired number of factors, predicted values of the continuous response variable are calculated in a stepwise manner.¹⁸ In PLR, the first step is the same as above: the scores matrix is obtained by extracting a desired number of factors, treating the dependent variable as if it were continuous. For the prediction step, the stepwise process for constructing continuous response variables in PLS is no longer valid; instead, logistic regression is then performed using the scores matrix as the independent variables to model the nominal response data.

The parameters used in the PLS and PLR methods include the number of structural features preselected for use as predictor variables and the number of factors extracted during the model building process. Goodness of fit is compared using R^2 and RMSE (root-mean-square error). Leave-one-out cross-validation of the training set was used to determine the optimal number of preselected features and number of factors. Using these parameters, the models were then applied to predict activities of the test sets. Since the test sets were not used at any point during the training process, this approach provides a true test of the predictive ability of a model.

Results and Discussion

PTP1B inhibitors are negative regulators in insulin and leptin signaling cascades and hence play a role as potential agents for type II diabetes and obesity.^{19–22} Compound classes reported in the literature include naphthoic acid¹⁹ and sulfonamide²⁰ analogues of phosphonodifluoromethyl phenylalanines, and oxalyl-aryl-amino benzoic acid²¹ and salicylate-based ligands.²² Recent studies have suggested that apparent activities of some compounds may instead be explained by nonspecific mechanisms such as sequestration of protein by hydrophobic aggregates.²³ We selected the benzothiophenes and furan analogues presented by Malamas since the number and variety of the local neighbors are good from a modeling standpoint, and also because this allows comparison of this new modeling methodology with results from 3-D QSAR models that have been published based on this same data set. Structural classification of the data set and assessment of structural similarity of the training and test sets are first addressed prior to modeling the activities of the PTP1B inhibitors. The six steps described in the Methods section is followed for discussion.

Step 1. Diagnosis of Data Sets. Analyzing the dataset for structural diversity, similarity, and distribution is the first step in the model building process. The distribution of pIC_{50} values for the 92-training and 26-test sets is illustrated in Figure 1. The 26-test set contains a higher proportion of active compounds, as indicated by the shift in mean: the means and standard deviations of the 92-training and 26-test sets are 0.65 ± 0.55 and 0.89 ± 0.59 , respectively. Applying the t -test to compare these two sets, the mean pIC_{50} value of the 26-test set was significantly higher than that of the training set at the 95% confidence level (p -value of 0.026 in a one-tail test). The mean of a truly random test set should be approximately equal to that of the training set. In this paper, we are using the same test set identified by Murthy et al.⁸ so that an exact comparison of models can be made. Ensuring that the training set is representative of the test set is obviously an important criterion when selecting a single test set. This issue was addressed by running rigorous cross-validation during the model building process.

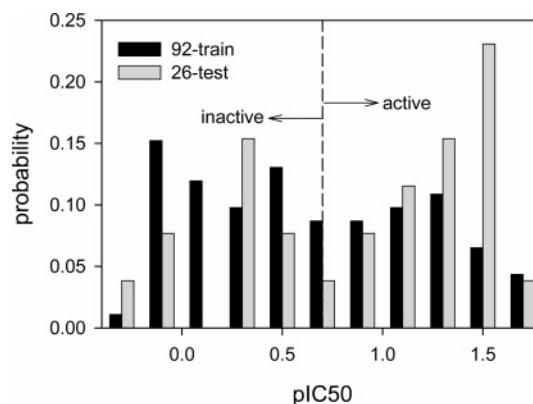
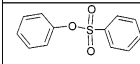
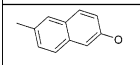
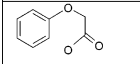
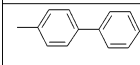


Figure 1. Histogram of pIC_{50} distribution. Compounds with pIC_{50} values lower than the mean of the training set were considered inactive. A significantly higher percentages of actives are observed in the 26-test set.

Table 1. Examples of Chemical Class Groupings of the Data Set

Major Classes	% frequency		
	92-training set	26-test set	19-unknown set
benzimidazole	0	0	5.3
benzofuran	53	61	63
benzothiophene	32	35	5.3
oxazole	7.6	3.8	21
pyridine	2.2	0	5.3
tetrazole	3.3	3.8	0
	10.9	19.2	0
	10.9	7.7	10.5
	63	62	47
	15	7.7	37

Measuring the structural similarity between training and test sets is also a critical preliminary step for any modeling approach. It is imperative that the chemical space of the test set lies within that of the training set. This paper presents three independent methods to diagnose the appropriate structural similarities between the test and training sets: compound class grouping, Sammon map, and structure similarity correlation.

According to chemical class grouping shown in Table 1, the 26-test set contains a distribution of compound classes similar to that of the 92-training set, with few exceptions. The 26-test set contains a higher frequency of benzofurans and sulfonyl groups than the 92-training set, while the number of compounds with pyridine and oxazole structures is less. The 19-unknown set turns out to be much less similar to the training set than the 26-test set. The 19 set contains the benzimidazole class, which was not part of the 92-training or 26-test sets. A greater proportion of compounds in the 19-unknown set contain phenol, oxazole, and benzofurans. This set also has no compounds with sulfonyl groups and much lower benzothiophene content than either the 92-training or 26-test sets.

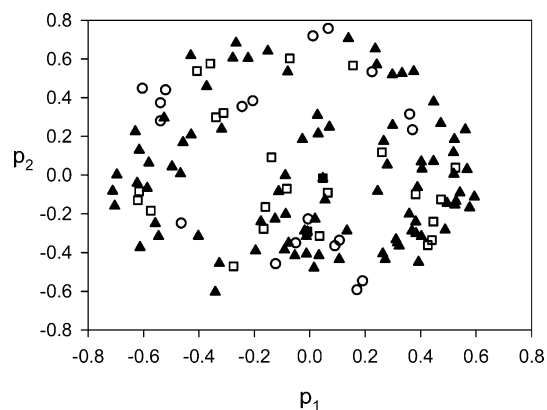


Figure 2. A 2-D Sammon map projection. The chemical feature space of the three datasets (▲ 92-training, ◻ 26-test set, ○ 19-unknown set) projected in two dimensions. Axis scales are Tanimoto similarity.

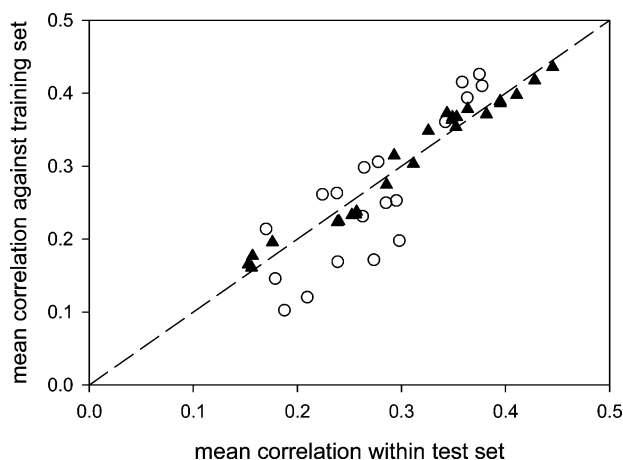


Figure 3. Structural correlations of the 19-test (○) and 26-test (▲) sets with the 92-training set.

The Sammon map²⁴ in Figure 2 illustrates a 2-D projection of the structural feature space of the 92-training, 26-test, and 19-unknown compound sets. Sammon maps provide only a semiquantitative measure of similarity and must be interpreted with caution since they are highly distorted due to projection of high dimensional data to only two or three dimensions for visualization. The 92-training and 26-test sets appear to be quite similar; there are only a small number of test set compounds that have no close neighbors in the training set. On the other hand, the 19-unknown set appears to have somewhat more dissimilar features from the other 118 compounds.

A quantitative similarity metric to estimate structural correlation between the test and training sets was calculated for each compound in the 26-test and 19-unknown sets against all compounds in the 92-training set. Ideally, the *within* (test–test) and *between* (test–training) average structural feature correlations should be approximately equal. Therefore, the extent to which points lie along the diagonal in Figure 3 indicates how well a test set is represented structurally in the training set. The compounds in 92-training and 26-test sets are correlated at 97%, whereas the correlation is only 68% between the 19-unknown and 92-training sets.

In summary, the analyses presented in Table 1 and Figures 2 and 3 provide valid methods to assess the similarity of the chemical space between training and

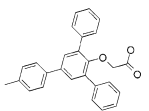
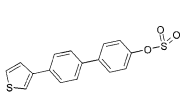
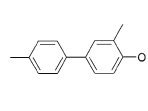
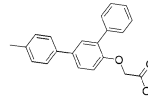
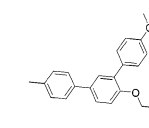
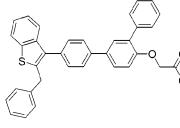
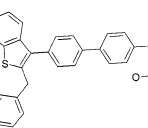
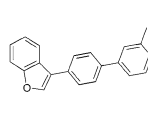
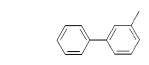
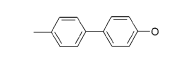
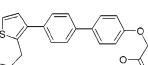
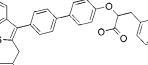
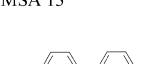
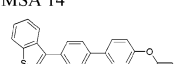
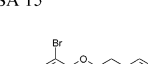
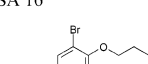
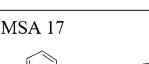
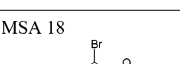
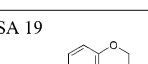
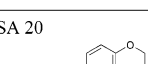
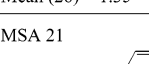
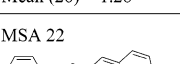
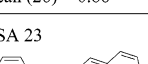
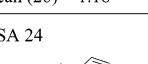
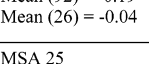
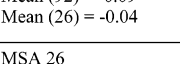
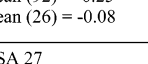
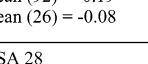
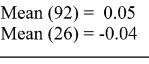
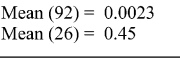
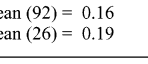
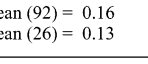
test sets. All three methods, chemical class grouping, Sammon map, and similarity correlations of the data sets, independently confirmed that the 26-test set is more similar to the 92-training set than the 19-unknown set. For these datasets we note that the similarities between test sets and the training set are not optimal, a fact that will ultimately limit the predictive accuracy of any model. As explained earlier, the reason for choosing these particular test sets was to allow comparison of the model presented here with results from prior studies. This is a common problem in this area and points out the need for a more systematic approach to evaluating informatics models, an approach in which models are judged based on their performance averaged over a large number of randomly selected test sets rather than in-depth analysis of a single test set. The issue of domain of applicability will be addressed in a subsequent paper.

Step 2. Assembly of Macrostructures. The data set used for modeling inhibition of PTP1B produced a total of 509 medicinal chemistry building blocks to describe 92 test compounds. On the basis of the original structural features, 71 macrostructures that discriminate the activity were assembled, providing a combined total of 580 features. Examples of MSAs used in modeling are presented in Chart 1.

Step 3. Preselection of Features. For the pIC₅₀ model, a subset of features was then selected based on a *t*-statistic calculated for each feature, comparing the mean pIC₅₀ for compounds having a particular feature with the mean for those without. Of the top 150 influential features, 41 were MSAs. Chart 1 lists macrostructures sorted in order of the *t*-statistic as well as a summary of compound groupings in training and test sets. The percentage of compounds classified by the predictive MSAs in the test and training sets is quite similar, which is another good indication of the structural proximity of the two sets. After preselection of structural features, calculated molecular properties described in the Methods section can be added. The final predictors, therefore, consist of both binary and continuous variables.

Step 4. Model Building. The nonlinear partial least squares (PLS) method was then used to model the reduced dataset. PLS further reduces the dimensionality of the features by extracting factors that are linear combinations of the input variables. Two parameters, the number of preselected features and the number of PLS factors to be extracted, were optimized by leave-one-out cross-validation to avoid overtraining and to maximize the prediction capability. The mean R^2 (averaged over 92 runs) for training sets and Q^2 for the leave-one-out results are reported in Table 2, which also lists the relevant modeling parameters: types of predictors (all structural features vs MSAs), number of preselected features, and the number of PLS factors. Only settings that yielded the best cross-validation are reported in Table 2. Figure 4 presents a detailed analysis of the effects of the number of preselected structural features and number of PLS factors. This figure clearly illustrates why preselecting a subset of differentiating features is necessary. As expected, essentially perfect fits ($R^2 = 1$) of the training set can be obtained using all features and a high number of PLS factors, but the

Chart 1. Examples of MSAs Used in the Model^a

<p>MSA 1</p>  <p>$t = 14.7$ PLS wt = 0.076 Mean (92) = 1.57 Mean (26) = 1.32</p>	<p>MSA 2</p>  <p>$t = 13.9$ PLS wt = 0.12 Mean (92) = 1.52 Mean (26) = 1.57</p>	<p>MSA 3</p>  <p>$t = 9.7$ PLS wt = 0.063 Mean (92) = 1.30 Mean (26) = 1.34</p>	<p>MSA 4</p>  <p>$t = 9.3$ PLS wt = 0.050 Mean (92) = 1.34 Mean (26) = 1.29</p>
<p>MSA 5</p>  <p>$t = 8.6$ PLS wt = 0.047 Mean (92) = 1.33 Mean (26) = 1.32</p>	<p>MSA 6</p>  <p>$t = 7.9$ PLS wt = 0.052 Mean (92) = 1.32 Mean (26) = 1.28</p>	<p>MSA 7</p>  <p>$t = 6.8$ PLS wt = 0.048 Mean (92) = 1.14 Mean (26) = 1.18</p>	<p>MSA 8</p>  <p>$t = 6.0$ PLS wt = 0.063 Mean (92) = 1.27 Mean (26) = 1.44</p>
<p>MSA 9</p>  <p>$t = 5.8$ PLS wt = 0.072 Mean (92) = 1.21 Mean (26) = 1.34</p>	<p>MSA 10</p>  <p>$t = 5.4$ PLS wt = 0.10 Mean (92) = 0.75 Mean (26) = 0.97</p>	<p>MSA 11</p>  <p>$t = 5.0$ PLS wt = 0.058 Mean (92) = 1.05 Mean (26) = 1.18</p>	<p>MSA 12</p>  <p>$t = 4.9$ PLS wt = 0.10 Mean (92) = 1.14 Mean (26) = 1.13</p>
<p>MSA 13</p>  <p>$t = 4.9$ PLS wt = 0.090 Mean (92) = 0.70 Mean (26) = 0.97</p>	<p>MSA 14</p>  <p>$t = 4.5$ PLS wt = 0.059 Mean (92) = 1.08 Mean (26) = 1.13</p>	<p>MSA 15</p>  <p>$t = 4.3$ PLS wt = 0.14 Mean (92) = 1.08 Mean (26) = 1.13</p>	<p>MSA 16</p>  <p>$t = 4.1$ PLS wt = 0.13 Mean (92) = 1.13 Mean (26) = 1.21</p>
<p>MSA 17</p>  <p>$t = 3.9$ PLS wt = 0.19 Mean (92) = 1.26 Mean (26) = 1.35</p>	<p>MSA 18</p>  <p>$t = 3.9$ PLS wt = 0.12 Mean (92) = 1.09 Mean (26) = 1.28</p>	<p>MSA 19</p>  <p>$t = 3.7$ PLS wt = 0.056 Mean (92) = 0.82 Mean (26) = 0.86</p>	<p>MSA 20</p>  <p>$t = 3.6$ PLS wt = 0.079 Mean (92) = 0.95 Mean (26) = 1.18</p>
<p>MSA 21</p>  <p>$t = -3.6$ PLS wt = 0.079 Mean (92) = 0.19 Mean (26) = -0.04</p>	<p>MSA 22</p>  <p>$t = -4.1$ PLS wt = 0.086 Mean (92) = 0.09 Mean (26) = -0.04</p>	<p>MSA 23</p>  <p>$t = -4.4$ PLS wt = 0.065 Mean (92) = 0.23 Mean (26) = -0.08</p>	<p>MSA 24</p>  <p>$t = -4.9$ PLS wt = 0.090 Mean (92) = 0.19 Mean (26) = -0.08</p>
<p>MSA 25</p>  <p>$t = -5.1$ PLS wt = 0.080 Mean (92) = 0.05 Mean (26) = -0.04</p>	<p>MSA 26</p>  <p>$t = -5.9$ PLS wt = 0.048 Mean (92) = 0.0023 Mean (26) = 0.45</p>	<p>MSA 27</p>  <p>$t = -6.1$ PLS wt = 0.094 Mean (92) = 0.16 Mean (26) = 0.19</p>	<p>MSA 28</p>  <p>$t = -6.4$ PLS wt = 0.096 Mean (92) = 0.16 Mean (26) = 0.13</p>
<p>MSA 29</p>  <p>$t = -6.5$ PLS wt = 0.087 Mean (92) = 0.15 Mean (26) = 0.44</p>	<p>Feature (additional 1)</p>  <p>Mean (92) = 1.07 Mean (26) = 1.53</p>	<p>Feature (additional 2)</p>  <p>Mean (92) = 1.22 Mean (26) = 1.53</p>	<p>Feature (additional 3)</p>  <p>Mean (92) = 1.28 Mean (26) = 1.59</p>

^a The IDs of the macrostructures are numbered in decreasing order of t values. Average PLS weights are calculated from the weights of the PLS factors for each MSA. (Only the top 10% of the structural features were considered.) The mean of the absolute average PLS weights across all selected 150 structural features was 0.074. The mean (92) and mean (26) represent the average pIC₅₀ values for the 92-training and 26-test datasets, respectively.

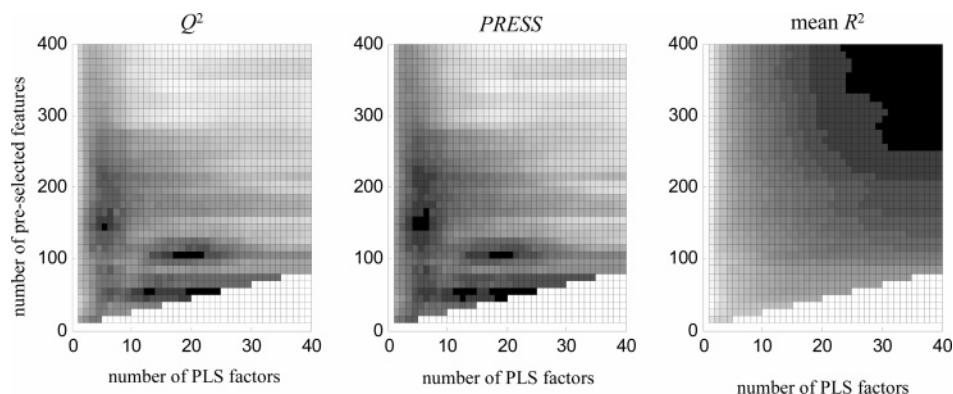


Figure 4. Determination of the number of PLS factors and preselected features to be used in model. Surface plots of the Q^2 correlation coefficient and PRESS (prediction error sum of squares) summarize results for leave-one-out cross-validation on the 92-training set. Plot on right shows the mean training set R^2 values for each cross-validation run. Shading is proportional to correlation coefficient for Q^2 and R^2 (darker indicates higher correlation) and inversely proportional to PRESS for a given set of parameter values.

Table 2. Parameter Optimization for Cross-Validation

predictor types	G^a	F^b	training set		leave-one-out CV	
			R^2	RMSE	Q^2	RMSE
all: base features + MSA + 8 properties	580 (all used)	3	0.77	0.27	0.57	0.37
base features + MSA + 8 properties	200	5	0.83	0.23	0.68	0.31
base features + 8 properties	150	4	0.81	0.24	0.70	0.30
base features + 150	150	12	0.89	0.18	0.62	0.34
base features + 150	150	20	0.93	0.15	0.57	0.39
base features + 100	100	6	0.78	0.26	0.66	0.32
base features + 100	100	19	0.91	0.17	0.71	0.31
base features + 50	50	12	0.83	0.23	0.71	0.30
base features + 50	50	5	0.76	0.28	0.68	0.31
base features only	150	4	0.71	0.30	0.56	0.37
base features + 8 properties	150	4	0.76	0.27	0.62	0.34
MSA only	71 (all MSA)	5	0.80	0.25	0.61	0.35
MSA + 8 properties	71 (all MSA)	5	0.84	0.22	0.68	0.32
8 properties alone	8	1	0.48	0.40	0.47	0.40

^a Number of preselected structural features. ^b Number of PLS factors.

cross-validation Q^2 and PRESS (prediction error sum of squares) statistic are very low, indicating that such models are overtrained and therefore not providing meaningful information. Overall, higher cross-validation Q^2 and PRESS values are observed for a relatively low number of PLS factors (<10) and with 100–200 preselected features. When all 580 structural features were used without preselection, models overtrain such that the cross-validation Q^2 values decrease rapidly with increasing number of PLS factors greater than 4.

This combination of preselection and the use of minimum number of PLS factors is important not only for model robustness but also for meaningful chemical inference. Finding structural features significantly impacting the models by decoding PLS factors becomes more intuitive with a small number of factors. The use of macrostructures alone as structural features also gave reliable models for this local neighborhood of benzothiophene and benzofuran biphenyl/naphthalene analogues. All 71 MSAs were included when building models with structural features made of only MSAs. Inclusion of eight molecular properties, including aLogP, enhanced the training and cross-validation by 0.03 log unit in the root-mean-square error.

Using the parameters optimized by cross-validations, the influence of each of the molecular descriptors on the

Table 3. Predictive Power of the Molecular Descriptors

descriptors	G^a	F^b	training set		26-test set		
			R^2	RMSE	Q^2	RMSE	
all descriptors: basic features + MSAs + 8 properties	150	4	0.80	0.24	0.70	0.72	0.32
basic features only	150	4	0.71	0.30	0.56	0.59	0.38
basic features + MSAs	150	4	0.78	0.26	0.67	0.68	0.33
basic features + 8 properties	150	4	0.76	0.27	0.62	0.69	0.33
MSAs only	71	5	0.79	0.25	0.61	0.64	0.36
MSAs + 8 properties	71	5	0.83	0.22	0.68	0.65	0.37
8 properties only	8	1	0.48	0.40	0.47	0.66	0.37
CoMFA model A*			0.72			0.51	
CoMFA model B (with aLogP)*			0.84			0.75	

^a Number of preselected features. ^b Number of PLS factors used.

model's ability to predict the 26-test set was investigated. Various models, using base structural features, MSAs, calculated properties, and combinations of these descriptor sets, are compared in Table 3. Overall, the results clearly demonstrate the usefulness of the structural features for modeling the potency of biphenyl and naphthalene analogues of benzothiophenes and benzofurans for PTP1B inhibitors. These results compare very favorably with results from Murthy's CoMFA models.⁸

MSAs alone work remarkably well as predictors for both cross-validation of the training set and modeling the true test set. Although including preselected base features as well as the eight physical properties improves prediction, MSA-only models are very helpful for establishing chemical inference at the evaluation stage. The MSAs allow medicinal chemists to evaluate prediction models by chemical inference without resorting to the more abstract mathematical treatment of structural features as in the case of principal component analysis or genetic algorithms for reducing high dimensionality and redundancy.

The inclusion of aLogP substantially enhanced the predictive accuracy of an earlier CoMFA study.⁸ Several other 3-D QSAR studies reported using a similar data set containing analogues of benzothiophene and benzofuran biphenyls.^{9–12} Descriptors such as hydrogen bonding and CPSA (charged polar surface area) improved model accuracy, whereas HOMO descriptors were marginal at best.^{8,10} In this study, the properties aLogP, parent molecular weight, and hydrogen bond acceptors

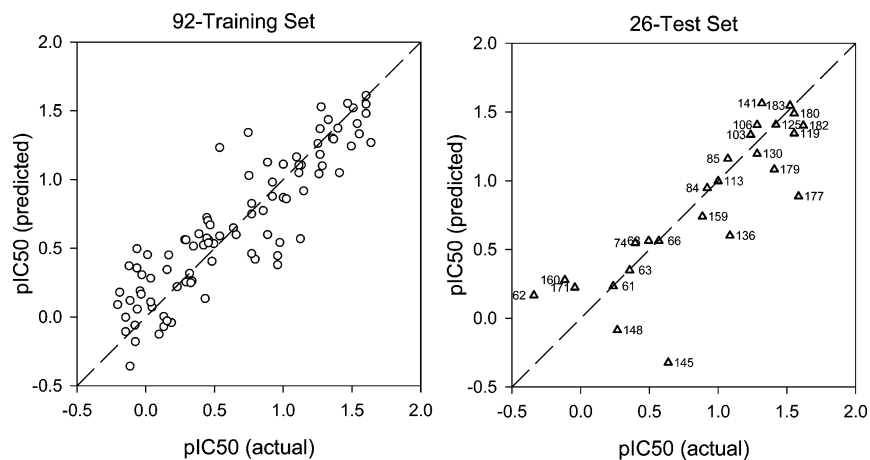


Figure 5. Comparison of actual and predicted pIC_{50} values for 92-training and 26-test sets. Predicted values obtained from PLS model using all features (basic structures, MSAs, physical properties) with 150 preselected features and 4 PLS factors. Numbers in plot on right denote compound IDs.

were significant in the model; however, the effects were not as dramatic as those observed in the CoMFA study or other 3-D QSAR studies. MSAs may already provide the information that LogP contributes to the model. Although the properties-only model performed as well as the MSA model for the 26-test set, this result is somewhat misleading because comparison of results using different randomly selected test sets shows that the properties-only model in general is not nearly as good. In both the 92-training and 26-test sets, strong correlations between the $aLogP$ and activities were observed. The 26-test set used here was selected specifically in order to compare with a previous CoMFA study.⁸ As discussed earlier, this set contains compounds whose average IC_{50} value is 1.8 times greater than that of the 92-training set. This shift toward higher activity with stronger correlation with $aLogP$ clearly leads to a greater dependence on $aLogP$ in the model for this particular test set.

Step 5. Evaluation of Model by Chemical Inference. One of the most obvious advantages of the modeling method presented in this paper is the ability to connect models directly to structural design. A detailed structural evaluation of the model is now attempted using best model for the 26-test set, which was built using all predictors (basic features, MSAs, physical properties) with 150 preselected features and 4 PLS factors. The actual pIC_{50} values are plotted against predicted values for both 92-training and 26-test sets in Figure 5. We discuss examples of compounds that are predicted accurately and explain why the model works. Equally important, we also present cases where the model fails. Only through an iterative process of chemical inspection and testing model hypotheses can the prediction capability be fairly assessed. For this discussion, Chart 1 lists the MSAs employed in this model, including the discriminating (t -statistic) and predictive power (PLS weights). Table 4 lists selected structural features that were weighted highly in the PLS model for the 26-test set.

To demonstrate chemical inference, examples of inactive and active compounds that are accurately and incorrectly predicted are considered. These compounds are listed in Chart 2. The pIC_{50} values for compounds 63 and 113 are predicted accurately in the 26-test set.

Compound 63 is an inactive compound whose pIC_{50} value was predicted correctly. Approximately one-fourth of the structures in the training set were found to be similar to compound 63, with pIC_{50} ranging from 1.367 to -0.121 . The compound contains a strongly negatively correlating feature, MSA 28, as well as several positively correlating features such as MSA 10 and 19. MSA 28 is a highly discriminating feature, representing the analogues of 1,2-disubstituted benzofurans, whereas the corresponding benzothiophene analogues show higher activities. All 14 compounds containing MSA 28 in the training set have pIC_{50} values lower than the average of the training set (0.65). On the other hand, the positively correlating feature MSA 10 appears in 19 benzofuran analogues that are above the average, while 21 of the analogues were lower. The presence of MSA 10 does not guarantee activity; however, the absence of this feature correlates well with lower activity. In the 92-training set, all 11 compounds without this feature are inactive. Thus, the overall prediction of pIC_{50} lower than the average of the training set is reasonable.

To assess whether these discriminating features provide predictive accuracy, it is useful to calculate the PLS weights averaged for all 4 factors employed in this model. PLS factor weights provide a direct measure of the influence of a descriptor on the model. Since each factor is a linear combination of all the structural features plus the properties, the higher this weight the greater impact this feature has on the model. If a feature exhibits high t -values and above average PLS weights, it can be considered to impart predictive accuracy. For compound 63, the discriminating MSAs 28 and 10 give high predictive accuracy, as listed in Table 4.

The next example is the active compound 113 ($pIC_{50} = 1.0$) that was predicted correctly. Approximately one-fifth of the structures were found to be similar in the training set (Chart 2). They represent the compound classes of 2-benzylbenzothiophene biphenyl oxo-acetic acid. Compound 113 is described by many positively correlating features (MSA 3, 4, 9, 7, 10, 11, 20, and 19) and contains no negatively correlating structural features. All structures containing MSA 3 and 4 belong to the active class (pIC_{50} values greater than average) in the training set. Prediction of this compound depends

Table 4. Predictive pIC₅₀ and Classification Models for the 26-Test Set^a

compd ID	exptl ^b pIC ₅₀	predicted models		significant features in prediction
		pIC ₅₀	class (prob) ^c	
61	0.237	0.232	inactive (0.041)	methane, 1-aryl-,1-phenyl-; MSA 10; MSA 13, benzene, 1-aryl-,4-hydroxy-; thiophene, 2-benzyl-
62	-0.34	0.166	inactive (0.014)	MSA 10; MSA 28; MSA 13; benzene, 1-aryl-,4-(2-oxoethoxy)-; MSA 19
63	0.357	0.348	inactive (0.037)	MSA 10, MSA 28, benzene, 1-aryl-,4-phenyl-; MSA 13, MSA 19
66	0.569	0.562	inactive (0.017)	methane, 1-aryl-,1-phenyl-; MSA 10; MSA 13; benzene, 1-aryl-,4-phenyl-; benzene, 1-aryl-,4-(2-oxoethoxy)-; MSA 19
68	0.495	0.562	inactive (0.017)	MSA 10; MSA 13; benzene, 1-aryl-,4-phenyl-; benzene, 1-aryl-,4-(2-oxoethoxy)-; MSA 19
74	0.398	0.547	inactive (0.018)	MSA 10; MSA 13; benzene, 1-aryl-,4-phenyl-; benzene, 1-aryl-,4-(2-oxoethoxy)-; MSA 19
84	0.921	0.947	active (0.98)	MSA 12, MSA 10, benzene, 1-aryl-,4-phenyl-; MSA 13; MSA 20
85	1.071	1.159	active (0.98)	MSA 12, MSA 10, benzene, 1-aryl-,4-phenyl-; MSA 13; MSA 20
103	1.237	1.335	active (1.0)	MSA 15; MSA 16; MSA 18; MSA 12; MSA 10; MSA 13; MSA 20
106	1.284	1.406	active (1.0)	methane, 1-aryl-,1-phenyl-, MSA 12, MSA 10, MSA 13, MSA 20, MSA 3
113	1	0.997	active (0.93)	MSA 10; MSA 13; MSA 20; MSA 9; MSA 3
119	1.553	1.343	active (0.99)	MSA 18; MSA 10; MSA 13; MSA 20; MSA 9; MSA 3
125	1.42	1.407	active (0.90)	MSA 16; MSA 18; MSA 10; benzene, 1-aryl-,4-phenyl-; MSA 13
130	1.284	1.196	active (0.89)	MSA 18; methane, 1-aryl-,1-phenyl-; MSA 10; benzene, 1-aryl-,4-phenyl-; MSA 13
136	1.086	0.6	inactive (0.47)	methane, 1-aryl-,1-phenyl-; MSA 10; benzene, 1-aryl-,4-phenyl-; MSA 13; benzene, 1,2,4-acyc
141	1.319	1.562	active (1.0)	MSA 10, MSA 13, MSA 1; MSA 12; MSA 3
145	0.638	-0.324	inactive (0.01)	MSA 10; MSA 28; MSA 27; benzofuran, 3-(alkyl, acyc)-; MSA 13; benzene, 1-aryl-,4-hydroxy-
148	0.268	-0.087	inactive (0.01)	MSA 10, MSA 28, MSA 27, benzofuran, 3-(alkyl, acyc)-; MSA 13; benzene, 1-aryl-,4-(2-oxoethoxy)-
159	0.886	0.74	active (0.67)	MSA 360, MSA 301, MSA 227, MSA 10; benzene, 1-aryl-,4-trifluoromethyl-; oxazole, 2-aryl-
160	-0.114	0.277	inactive (0.002)	MSA 28; MSA 23; MSA 27; MSA 24; naphthalene, 2-hydroxy-
171	-0.041	0.222	inactive (0.01)	MSA 28; MSA 23; MSA 27; MSA 24; MSA 22; MSA 25; MSA 21
177	1.585	0.886	active (0.89)	methane, 1-aryl-,1-phenyl-; 1-benzene-carboxylic acid, 2-hydroxy-; MSA 10; benzene, 1-aryl-,4-phenyl-; MSA 13
179	1.409	1.082	active (0.99)	methane, 1-aryl-,1-phenyl-; 1-benzene-carboxylic acid, 2-hydroxy-; benzene, 1-hydroxy-,3-sulfonyl-; MSA 10; MSA 13
180	1.553	1.489	active (1.0)	methane, 1-aryl-,1-phenyl-; 1-benzene-carboxylic acid, 2-hydroxy-; benzene, 1-hydroxy-,3-sulfonyl-; MSA 10; MSA 13
182	1.62	1.401	active (0.91)	MSA 2; methane, 1-aryl-,1-phenyl-; 1-benzene-carboxylic acid, 2-hydroxy-; benzene, 1-hydroxy-,3-sulfonyl-; MSA 10
183	1.523	1.546	active (0.94)	MSA 2; methane, 1-aryl-,1-phenyl-; 1-benzene-carboxylic acid, 2-hydroxy-; benzene, 1-hydroxy-,3-sulfonyl-; MSA 10

^a The features are listed in the order of significance (order of PLS weights) for each compound. ^b All experimental values are taken from the literature.^{11,12} ^c Probability calculated from logistic partial least squares model for compound classification.

heavily on MSAs 10, 13, 20, 9, and 3 as illustrated in Table 4. They are highly discriminating, as indicated by their *t*-values and PLS factor weights, whose combination allows good predictive accuracy.

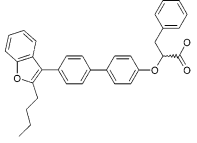
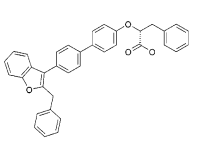
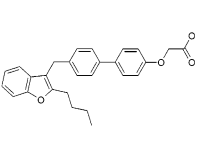
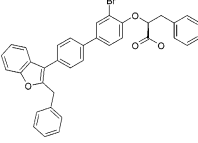
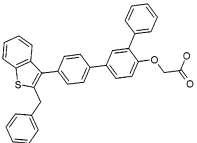
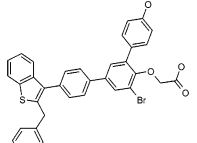
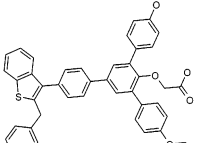
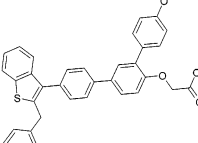
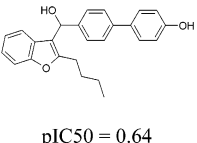
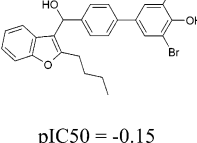
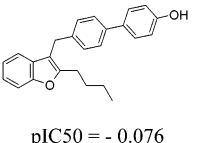
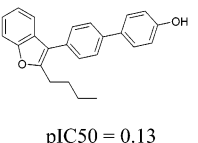
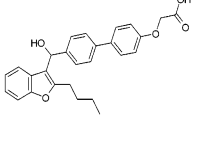
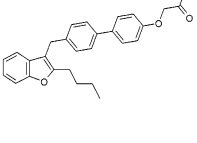
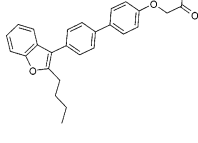
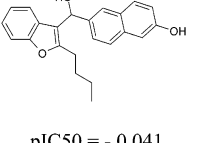
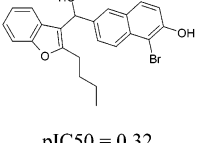
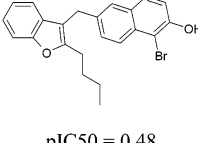
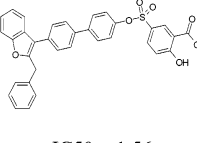
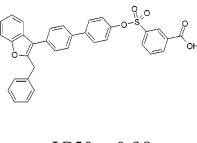
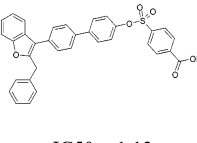
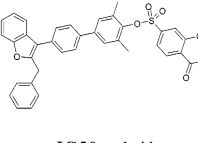
Chemical inference is especially important in evaluating a model in cases where predictions are not accurate. Applying this analysis to cases in which the model fails to predict correctly leads to understanding of the real robustness of a model or to the possible identification of experimental anomalies.

The pIC₅₀ values for compounds 145 and 177 are predicted inaccurately in the 26-test set. Compound 145 is a weakly active compound whose predicted pIC₅₀ value (-0.324) was much lower than its actual value (0.638). Chart 2 displays similar structures in the training set: compound 143 with a methylene spacer between the butylbenzofuran and the biphenylphenol groups; and compound 146, a dibromo analogue. The pIC₅₀ value of 145 (0.638) is much higher than those of both analogues (-0.076 and -0.146). The under-predicted activity may be due to the contribution from the biphenylphenol group since 14 of the 15 compounds containing this feature in the training set are inactive. Two factors, spacers between the benzofuran and biphenyl groups and bromide substitution, are further explored. First, the introduction of carbon spacers, methylene or hydroxymethylene, was reported by Malamas¹¹ to have a small effect on activity. As listed in Chart 2, addition of a methylene spacer to compound 52 to yield compound 143 decreases the activity, while

the same addition has the opposite effect in compounds 62 and 147. The addition of the hydroxymethylene spacer from compound 147 to 148 results in an increase in activity. On the basis of this observation, test compound 145 would be expected to be more active than 143, which indeed is the case; however, addition of the hydroxymethylene to 164 leads to lower activity in compound 163. The effect of bromide substitution is also not clear. In general, bromo analogues may yield higher activity, as observed here for naphthalene analogues 162 and 163. Malamas¹¹ pointed out that it is surprising that compound 146, the ortho-substituted dibromo analogue, exhibited lower activity than 145. Hence, the activity of the compound 145 cannot be explained by the dibromo substitution.

The effects of methylene and hydroxymethylene spacers and bromo substitution are therefore at best ambiguous in attempts to explain the observed activity of test compound 145 based on similar compounds in the training set. All analogues in the training set belong to the inactive class; 145 is the only compound in this group that seems to be somewhat active. In our model, this compound is described by MSAs 28, 27, and 29. These structural features are strongly correlated with lower activity in the training set; all but one of the compounds containing MSA 29 in the training set is inactive. Examining the structures of the analogues as well as the differentiating macrostructures, none of the compounds similar to 145 are active. Thus, the reported high pIC₅₀ for compound 145 could not be explained by

Chart 2. Examples of Structures Used for Chemical Inference

Test compounds	Similar compounds in training set		
Test-63  pIC ₅₀ = 0.36	67  pIC ₅₀ = 0.46	147  pIC ₅₀ = -0.061	124  pIC ₅₀ = 1.25
Test-113  pIC ₅₀ = 1.00	118  pIC ₅₀ = 1.54	121  pIC ₅₀ = 1.60	114  pIC ₅₀ = 1.10
Test-145  pIC ₅₀ = 0.64	146  pIC ₅₀ = -0.15	143  pIC ₅₀ = -0.076	52  pIC ₅₀ = 0.13
	148  pIC ₅₀ = 0.27	147  pIC ₅₀ = -0.061	62 (test set)  pIC ₅₀ = -0.34
	162  pIC ₅₀ = -0.041	163  pIC ₅₀ = 0.32	164  pIC ₅₀ = 0.48
Test-177  pIC ₅₀ = 1.56	175  pIC ₅₀ = 0.98	174  pIC ₅₀ = 1.12	176  pIC ₅₀ = 1.41

our model. It is also possible that the reported value might have been an experimental anomaly.

The second case of a prediction outlier is compound 177, an active compound whose pIC₅₀ value was predicted to be lower (1.19) than the actual value (1.585). Compound 177 (Chart 2) contains biphenyl derivatives with sulfonyl-salicylic acid with sulfonyl group meta to the carboxylic acid. There were several highly similar structures in the training set, from which the prediction was based. The structural features that were positively correlated with activities were oxybiphenyls (MSA 10, 13, 3) and biphenyls connected to sulfonylbenzene or *p*-sulfonylbenzoic acid. No negatively correlating MSAs were found in these similar compounds. As confirmed in Table 5, MSAs 10 and 13 were among the most significant features in the model. However, MSA 10 is

Table 5. Parameter Optimization for Classification Model

		training			cross-validation		
		% concor- dance	% sensi- tivity	% speci- ficity	% concor- dance	% sensi- tivity	% speci- ficity
<i>G</i>	<i>F</i>						
all 150	2	84.8	81.4	87.8	82.6	81.4	83.7
	8	94.6	95.9	93.0	75.0	65.1	83.7
all 50	2	83.7	81.4	85.7	82.6	79.1	85.7
	8	91.9	88.4	93.9	79.3	74.4	83.7
MSA (50) + properties	2	89.1	91.8	86	87.0	81.4	91.8
	9	95.7	95.3	95.9	83.7	81.4	85.7

not helpful in the local neighborhood of this compound since every compound has this feature. No MSAs that were negatively correlated with activity were found in the training set. On the other hand, compound 177 contains sulfonylsalicylic acid, a potent series of PTP1B

inhibition. In Chart 1, no MSAs containing the sulfonylbenzoic acid feature are included in the model because of relatively low *t*-values observed in the feature preselection step of the model building process. Therefore, predicting pIC_{50} lower than the actual can be explained.

Step 6. Refining Model by Chemical Inference.

As discussed above, some key features may be missed by the algorithmically reassembling MSAs. New features can be designed to test the structure activity relationship hypothesis and hence refine the model to improve the prediction accuracy. For example, sulfonylbenzoic acid features can be added to the existing structural feature predictors. Compound 177 contains the OH group ortho to the acid, although the sulfonyl group is meta to the acid. Examining similar compounds in the training set, salicylic acid (OH group ortho to the acid) is associated with higher activity than benzoic acid. In addition, a sulfonyl group para to the carboxylic acid enhances the activity. This suggests addition of new macrostructures to capture this relationship consisting of the sulfobenzene with ortho OH and carboxylic acid groups. On the basis of this analysis, three additional features (shown in Chart 1) were then included in the model. The model was built again by extracting 4 PLS factors from the 153 structural features. Without changing the overall goodness of fit, the predicted value for compound 177 increased from 0.886 to 0.947, somewhat closer to the experimental value of 1.54. This is a good example of how chemical inference can be used to guide the model.

Classification Model. For screening compound activities, quite often binary or categorical ranking of potency or efficacy is sufficient. Compounds with activities lower than 0.70 (average of the whole 118 set) were considered inactive. The same 92-training and 26-test sets were used to develop a classification model based on partial logistic regression (PLR). In PLR models, the probability of each compound being active or inactive is calculated. The overall accuracy and numbers of false positive and false negative predictions are determined.

A new set of MSAs was extracted from the dataset using the binary biological response; the resulting set was very similar to the set based on the pIC_{50} data. As before, optimal parameters were selected based on leave-one-out cross-validation. In all cases, the logistic model prediction power was best for a small number of extracted PLS factors ($F = 2$). Predictors consisting only of MSAs and augmented by the same 8 physical properties gave the best balance of the specificity and sensitivity. These classification models gave excellent results with low RMSE, high concordance (overall accuracy), specificity (true negatives), and sensitivity (true positives) as illustrated in Table 5.

The model using MSAs and 8 properties was applied to the 26-test set. Even compounds whose predicted pIC_{50} values deviated significantly from experimental results (compounds 62, 145, and 177) were classified correctly. When the pIC_{50} values were close to the average, classification was not as accurate even though the predicted pIC_{50} values were in good agreement with experimental values. For example, the experimental and predicted pIC_{50} values for structures 136 (26-test set) and 137 (92-training set) were close to the average of

Table 6. Classification Model of % Inhibition: 19 Compound Set

compd ID	% inhibition ^a	activity class		probability	IC ₅₀ predicted
		assigned ^b	predicted		
53	-51 (2.5 μ M)	0	0	0.004	0.78
54	-13 (2.5 μ M)	0	0	0.001	1.7
55	-54 (2.5 μ M)	0	0	0.003	1.5
59	-27 (2.5 μ M)	0	0	0.012	0.56
64	-47 (2.5 μ M)	0	0	0.005	0.54
65	-58 (2.5 μ M)	0	0	0.018	0.39
69	-38 (0.25 μ M)	0	0	0.149	0.22
70	-36 (2.5 μ M)	0	0	0.124	0.23
93	-41 (2.5 μ M)	0	0	0.047	0.36
94	-19 (1 μ M)	0	0	0.056	0.67
101	-59 (1 μ M)	0	0	0.031	0.56
102	-43 (1 μ M)	0	0	0.125	0.32
135	-59 (0.1 μ M)	1	1	0.929	0.080
144	-53 (2.5 μ M)	0	0	0.004	2.4
150	-56 (2.5 μ M)	0	0	0.005	1.9
154	-11 (2.5 μ M)	0	0	0.007	2.0
155	-47 (2.5 μ M)	0	0	0.013	1.2
161	-41 (2.5 μ M)	0	0	0.001	2.5
173	-41 (2.5 μ M)	0	0	0.117	0.72

^a % inhibition data were taken from Malamas et al. ^b Activity was assigned as described in the Methods section.

the training. In the logistic model, these two were incorrectly classified as inactive; however, the probabilities that determined these classifications were close to 0.5 (0.62 for structure 136, 0.54 for structure 137). Probabilities near 0.5 for logistic models suggest that the compound is in the neutral area, having activities close to the cutoff value used to define actives vs inactives.

For the 19-unknown test set, a new model was developed using all 118 compounds (92-training plus 26-test) as the training set employing MSAs and 8 properties as predictors. As summarized in Table 6, the prediction of the activity classes of this true unknown set was 100% accurate, with the model correctly predicting compound 135 to be the only active compound in the 19-unknown set.

Prediction of IC₅₀ for the Unknown Set. The Malamas paper¹¹ reported only % inhibitions at a concentration range of 2.5–0.1 μ M for the 19-unknown compound set. The pIC_{50} model developed here was used to predict IC₅₀ values and compare with the previously reported % inhibition and concentration data. As summarized in Table 6, the predicted IC₅₀ values seem quite reasonable. It is important to note that these predictions are only speculative and have not been validated with real experimental assays; however, it demonstrates how a reliable model can provide direction and insight into the design of molecules.

Conclusion

An approach for developing predictive models for chemical activity based on 2-D structural descriptors is presented. Macrostructure assemblies (MSAs), dynamically constructed from a set of compounds, provide an intuitive means of reducing the high dimensionality of feature space and also greatly improve the ability to perform meaningful chemical inference. Regression models using partial least squares for continuous response data and logistic partial least squares for binomial response data are demonstrated to accurately predict activity of test set compounds. When building

predictive models, preselection of features is important to avoid overfitting; however, preselection based on statistical methods may result in the elimination of important features, especially if the compounds are within local neighbors of small size. This problem can be addressed by an iterative model building approach that allows for the incorporation of additional information.

Models are demonstrated to perform as well as 3-D QSAR models, which are known to be less intuitive in connecting back to the structural features when designing molecules. 3-D models provide some insights on molecular connectivity, surface area, and hydrogen bonding and charges. However, the modeling methodology presented here, based on 2-D structural features, enables an intuitive, quantitative connection of structural features of medicinal chemistry building blocks to compound activity. Chemical inference transparent to structural features enables efficient evaluation of hypotheses needed in the design of structures.

Acknowledgment. The authors acknowledge the Technology Action Fund Grant No. 02-066 provided by Ohio Department of Development to support this collaborative work. Personal communications with Drs. M. Myers and M. Lajiness at Eli Lilly were helpful.

References

- (1) Cramer, R. D. I.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. The Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (2) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
- (3) Martin, Y. C. Pharmacophore Mapping. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington DC, 1998; pp 121–148.
- (4) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.
- (5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (6) Cross, K. P.; Myatt, G. J.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Blower, P. E., Jr. Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.* **2003**, *46*, 4770–4775.
- (7) Saltiel, A. R.; Kahn, C. R. Insulin signaling and the regulation of glucose and lipid metabolism. *Nature* **2001**, *414*, 799.
- (8) Murthy, V. S.; Kulkarni, V. M. 3D-QSAR CoMFA and CoMSIA on Protein Tyrosine Phosphatase 1B Inhibitors. *Bioorg. Med. Chem.* **2002**, *10*, 2267–2282.
- (9) Sippl, W. Development of biologically active compounds by combining 3D QSAR and structure-based design methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 825–830.
- (10) Patankar, S. J.; Jurs, P. C. Classification of Inhibitors of Protein Tyrosine Phosphatase 1B Using Molecular Structure Based Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 885–899.
- (11) Malamas, M. S.; Sredy, J.; Gunawan, I.; Mihan, B.; Sawicki, D. R.; Seestaller, L.; Sullivan, S.; Flam, B. R. New azolidinediones as inhibitors of protein tyrosine phosphatase 1B with anti-hyperglycemic properties. *J. Med. Chem.* **2000**, *43*, 995–1010.
- (12) Malamas, M. S.; Sredy, J.; Moxham, C.; Klatz, A.; Xu, W.; McDevitt, R.; Adebayo, F. O.; Sawicki, D. R.; Seestaller, L.; Sullivan, D.; Taylor, J. R. Novel benzofuran and benzothiophene biphenylys as inhibitors of protein tyrosine phosphatase 1B with antihyperglycemic properties. *J. Med. Chem.* **2000**, *43*, 1293–1310.
- (13) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. Leadscape: software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.
- (14) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
- (15) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
- (16) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and developmental settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (17) Nguyen, D. V.; Rocke, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **2002**, *18*, 39–50.
- (18) Geladi, P.; Kowalski, B. Partial least squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
- (19) Burke, T. R.; Kole, H. K.; Roller, P. P. Potent inhibition of insulin receptor dephosphorylation by a hexamer peptide containing the phosphotyrosyl mimetic F2Pmp. *Biochem. Biophys. Res. Commun.* **1994**, *204*, 129–134.
- (20) Liu, D. G.; Gao, Y.; Voigt, J. H.; Lee, K.; Nicklaus, M. C.; Wu, L.; Zhang, Z. Y.; Burke, T. R., Jr. Acylsulfonamide-Containing PTP1B Inhibitors Designed to Mimic an Enzyme-Bound Water of Hydration. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3005–3007.
- (21) Xin, Z.; Oost, T. K.; Abad-Zapatero, C.; Hajduk, P. J.; Pei, Z.; Szczepankiewicz, B. G.; Hutchins, C. W.; Ballaron, S. J.; Stashko, M. A.; Lubben, T.; Trevillyan, J. M.; Jirousek, M. R.; Liu, G. Potent, Selective Inhibitors of Protein Tyrosine Phosphatase 1B. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1887–1890.
- (22) Liu, G.; Xin, Z.; Liang, H.; Abad-Zapatero, C.; Hajduk, P. J.; Janowick, D. A.; Szczepankiewicz, B. G.; Pei, Z.; Hutchins, C. W.; Ballaron, S. J.; Stashko, M. A.; Lubben, T. H.; Berg, C. E.; Rondinone, C. M.; Trevillyan, J. M.; Jirousek, M. R. Selective Protein Tyrosine Phosphatase 1B Inhibitors: Targeting the Second Phosphotyrosine Binding Site with Non-Carboxylic Acid-Containing Ligands. *J. Med. Chem.* **2003**, *46*, 3437–3440.
- (23) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46*, 4477–4486.
- (24) Sammon, J. W., Jr. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **1969**, *18*, 401–409.

JM0497242