
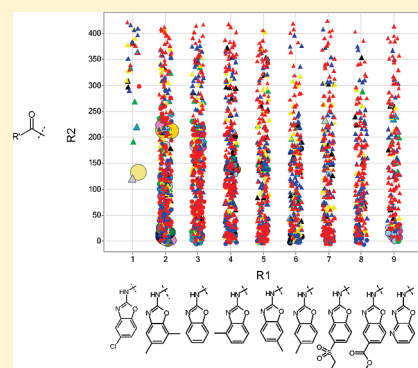


# Statistical Analysis and Compound Selection of Combinatorial Libraries for Soluble Epoxide Hydrolase

Li Xing<sup>\*,†,‡</sup> Robert Goulet,<sup>†</sup> and Kjell Johnson<sup>§</sup><sup>†</sup>Structural and Computational Chemistry, Pfizer Global Research and Development, 700 Chesterfield Parkway West, Chesterfield, Missouri 63017, United States<sup>‡</sup>Pfizer Global Research and Development, 200 CambridgePark Drive, Cambridge, Massachusetts 02140, United States<sup>§</sup>PharmaTherapeutics Statistics, Pfizer Global Research and Development, Groton, Connecticut 48105, United States Supporting Information

**ABSTRACT:** Inhibitors of soluble epoxide hydrolase (sEH) have been extensively pursued as antihypertensive therapies as well as potential treatment for other cardiovascular dysfunctions and prevention of renal damage. In this study we report quantitative structure–activity relationship (QSAR) models for 1223 structurally diverse sEH inhibitors produced by combinatorial library design and synthesis. Daylight fingerprints, MOE 2D and DragonX descriptors were generated for QSAR modeling approaches. Using these descriptors, a number of statistical models were trained and validated. Of these methods, gradient boosting machines (GBM), partial least-squares (PLS), and Cubist methods demonstrated the best performance on training and test set validation in terms of their leave-group-out cross-validated (LGO–CV)  $Q^2$  and correlation coefficient  $R^2$  ( $Q^2_{\text{GBM-training}} = 0.79$ ,  $R^2_{\text{GBM-test}} = 0.81$ ;  $Q^2_{\text{PLS-training}} = 0.75$ ,  $R^2_{\text{PLS-test}} = 0.75$ ;  $Q^2_{\text{Cubist-training}} = 0.91$ ,  $R^2_{\text{Cubist-test}} = 0.78$ ). A final model was constructed using the consensus approach of the three individual models and showed robust statistics and prediction of the external validation set. The Gaussian process modified sequential elimination of level combinations (G-SELC) method was then used to expand the chemical space beyond what has been explored by combinatorial synthesis. This approach identified 50 new compounds that are structurally diverse and potentially desirable for sEH inhibition based on prior knowledge. The activities of the suggested compounds were then predicted by the consensus QSAR model, and the results supported that the compounds were more likely to exist in the active parts of the chemical space. This study illustrates that the balanced approach by G-SELC could provide a general method for combinatorial library design, to effectively identify promising compounds to be created in the laboratory.



## INTRODUCTION

Epoxide hydrolases catalyze the hydration of epoxides to their corresponding diols. In particular, the cytosolic or soluble form of epoxide hydrolase (sEH) is the major enzyme that degrades epoxyeicosatrienoic acids (EETs) to dihydroxyeicosatrienoic acids (DHETs) in human blood vessels.<sup>1</sup> EETs are reported in the literature to be potent vasodilators and to have anti-inflammatory<sup>1,2</sup> and antiproliferative effects.<sup>1</sup> Therefore, inhibitors of sEH could enhance the circulating concentrations of EETs and be useful in the treatment of inflammation and cardiovascular dysfunction as well as end organ damage. Discovery of novel sEH inhibitors with improved pharmacological and physiological properties could serve potential development as antihypertensives in the clinic.

Combinatorial chemistry has been proven to be a powerful drug discovery tool for efficient synthesis of diverse compound libraries.<sup>3</sup> In the previous study, structure-based virtual screening was applied to design combinatorial libraries for the discovery of novel and potent sEH inhibitors.<sup>4</sup> X-ray crystallography revealed unique interactions for a benzoxazole template in addition to the conserved hydrogen bonds within the sEH catalytic machinery.

Taking advantage of the strengthened interactions, large numbers of reagents from commercial sources were identified to form the benzoxazole-based products. Guided by structure-based virtual screening a few hundred products were selected for synthesis. Biological assay of the library yielded exceptional hit rates (as high as 90%) at 10  $\mu\text{M}$  screening concentration. Combined with the follow up libraries, the combinatorial approach in total generated more than 1200 compounds, of which more than 300 are submicromolar sEH inhibitors by  $\text{IC}_{50}$  determination. In summary the successful library design not only yielded structure–activity relationships in an efficiently parallel way but also identified a number of extremely potent, single-digit nanomolar sEH inhibitors.

A generally challenging question in the pharmaceutical industry regarding combinatorial library design is: Given the existing data, how do we design the next round of compounds to maximize the learning from library approach? In this study we present QSAR analysis of the sEH compounds that have been

Received: February 28, 2011

Published: May 26, 2011

produced, with the goal of prioritizing remaining compounds in the library for subsequent synthesis. High-quality statistical models were created by extensive validation analysis. Accounting for prior knowledge on sEH inhibition, we employed the Gaussian process modified sequential elimination of level combinations (G-SELC) method<sup>5</sup> to design new compounds that have not been synthesized and applied QSAR models to predict their sEH activities.

The original SELC algorithm employs principles of experimental design and algorithmic optimization (genetic algorithms) to identify combinations of factors that are likely to improve upon the desired response.<sup>6–8</sup> Mandal et. al provided examples of the use of the SELC algorithm including a proof-of-concept example that showed the SELC identified monomers were related to activity improvement.<sup>7</sup> Using only 15% of the original resources, the method provided enrichment to ligand space, while also finding active compounds. G-SELC is a recent modification to the original SELC algorithm by improving upon the modeling relationship among independent and dependent variables through Gaussian spatial processes.<sup>5</sup> We herein apply this method to identifying promising compounds from a large collection of feasible compounds based on the library chemistry.

## MATERIALS AND METHODS

**Data Sets.** The sEH library employed 9 R1's (benzoxazole amines) and 416 R2's (carboxylic acids and/or acid chlorides) as foundational monomers. Of the 3744 exhaustive combinations of R1 and R2's, 1221 compounds (33% of the combinatorial space) were synthesized, and their sEH percent inhibitions were determined in an enzymatic inhibition assay at 10  $\mu$ M concentration.

**Division of Training, Test, and External Data Sets.** To build QSAR regression models, the 1221 compounds made in the sEH library were split into a training, test, and external validation data sets. A 'balanced random' split was used such that the data would be distributed evenly in proportion to the response. The data were first split so that 20% would be used for the external validation set, and the remaining data were then further split into training (80%) and test (20%) sets. This resulted in 783 compounds for the training set, 194 for the test set, and 244 for the external validation set. Distributions of sEH enzyme inhibition were plotted in Figure S1, Supporting Information. As can be seen, all data sets have similar distributions to the overall data set, with a large number of compounds clustering at 80% sEH inhibition level.

**Descriptor Generation.** An internally developed statistical modeling tool called the in silico model generator, or isMG, was used for all modeling analysis. This program allows one to easily input structures from multiple formats, apply a consistent structure conversion for input to the various descriptor calculations, and merge the end points and descriptor data necessary for model building and prediction. Among the numerous 2D and 3D descriptors that can be calculated, for the analysis in this study we used the MOE 2D descriptors ( $d = 258$ ),<sup>9</sup> DragonX descriptors ( $d = 1664$ ),<sup>10</sup> Daylight Fingerprints,<sup>11</sup> and ClogP<sup>12</sup> descriptors. MOE affords a wide range of 2D properties, such as connectivity and shape indices, atom and bond counts, physical and hydrogen-bond properties, subdivided surface areas, partial charge, and E-state descriptors. DragonX generates additional descriptors, including functional group counts, molecular properties, geometrical, constitutional, and topological descriptors as well as connectivity, eigenvalue-based indices. For Daylight Fingerprints, the 1000 most variable bits of the 2048 fingerprints from the whole training set were included.

The zero variance (and near zero variance) descriptors were eliminated, and then the correlated descriptors, at the level of 0.85 correlation coefficient, were removed. For correlated descriptors, the one that had the highest correlation with the end point, the sEH percent inhibition specifically, was kept. This reduced the number of descriptors for MOE to 102, DragonX to 461, and Daylight Fingerprints to 551 bits. ClogP was found to correlate with other descriptors hence was removed from the list that was used for regression analysis. The remaining descriptors were then centered and range-scaled within each dimension.

**Regression Methods.** The modeling methods used in this analysis included RuleQuest's Cubist,<sup>13</sup> linear least-squares (LM), multivariate adaptive regression splines (MARS),<sup>14</sup> neural network (NNET),<sup>15</sup> partial least-squares (PLS),<sup>16</sup> random forest (RF),<sup>17</sup> regression trees (RPART),<sup>18</sup> gradient boosting machine (GBM),<sup>19</sup> support vector machine polynomial kernel (SVMp), and radial kernel (SVMr).<sup>20</sup> The additional statistical methods were incorporated into isMG using functions available in the varet package in R.<sup>21</sup> When building a model, all methods used a leave-group-out cross-validation (LGO-CV) procedure where the training set data is randomly split into five groups. Each group is then eliminated from the training set, and the compounds sEH activities are predicted by the model built from the remaining compounds. The process is repeated for 20 times to report an averaged LGO-CV  $Q^2$ . The final models are the ones that showed the best statistics, e.g.,  $Q^2$  and root mean squared error (RMSE) for regression models, for a certain set of tuning parameters. Different models can be received, for example, by altering the number of trees in RF methods.

**Model Building and Validation.** Although many of the regression methods resulted in reasonable models based on their  $R^2$  and RSME values (Table S1, Supporting Information), visualizing the plots for the training and test sets by observed versus predicted values revealed some interesting trends. As shown in Figure S2, Supporting Information, the training set plots for RF, SVM radial, and polynomial methods displayed a bias of underpredicting compounds of low sEH inhibition as well as overpredicting compounds of high enzyme inhibition. These same trends were seen in the plots of the test set data. On the other hand, the Cubist, GBM, and PLS predictions are devoid of such systematic bias and showed respectable  $R^2$  values. Therefore the Cubist, GBM, and PLS methods were selected for determining descriptors to use in building the final model.

The statistical significance of the models were estimated by the LGO-CV  $Q^2$  in the training set, a coefficient of determination  $R_0^2$ , and a linear fit predictive  $R^2$  for both internal and external test sets.  $R_0^2$  is the correlation coefficient of the predicted versus observed values while forcing the linear regression line to have a zero intercept. Model acceptability cutoffs were  $Q^2 > 0.60$  for training set and correlation coefficient  $R^2 > 0.70$  for the internal test set. All models that satisfied both criteria were applied to external validation sets.

**External Validation and Y-Randomization Test.** We took vigorous procedures to validate a QSAR model by assessing its prediction accuracy for an external set that was entirely excluded from model building. In all cases, the prediction accuracy for the external validation set had to satisfy the following conditions:<sup>22</sup>

$$R^2 > 0.60$$

$$(R^2 - R_0^2)/R^2 < 0.10 \text{ and } 0.85 < k < 1.15$$

Scheme 1

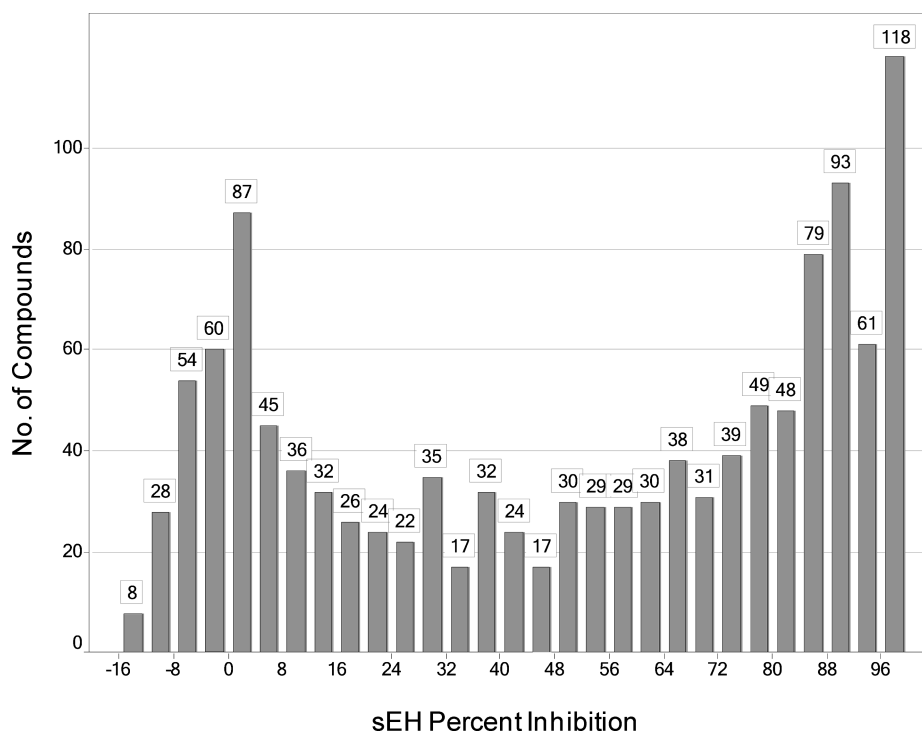
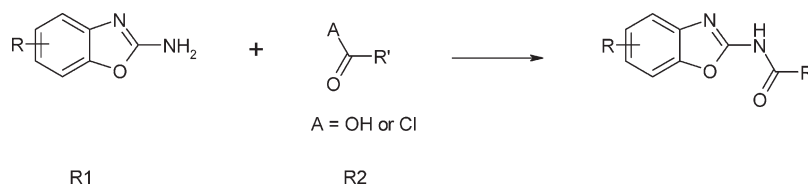


Figure 1. Distribution of percent inhibitions from historical sEH libraries.

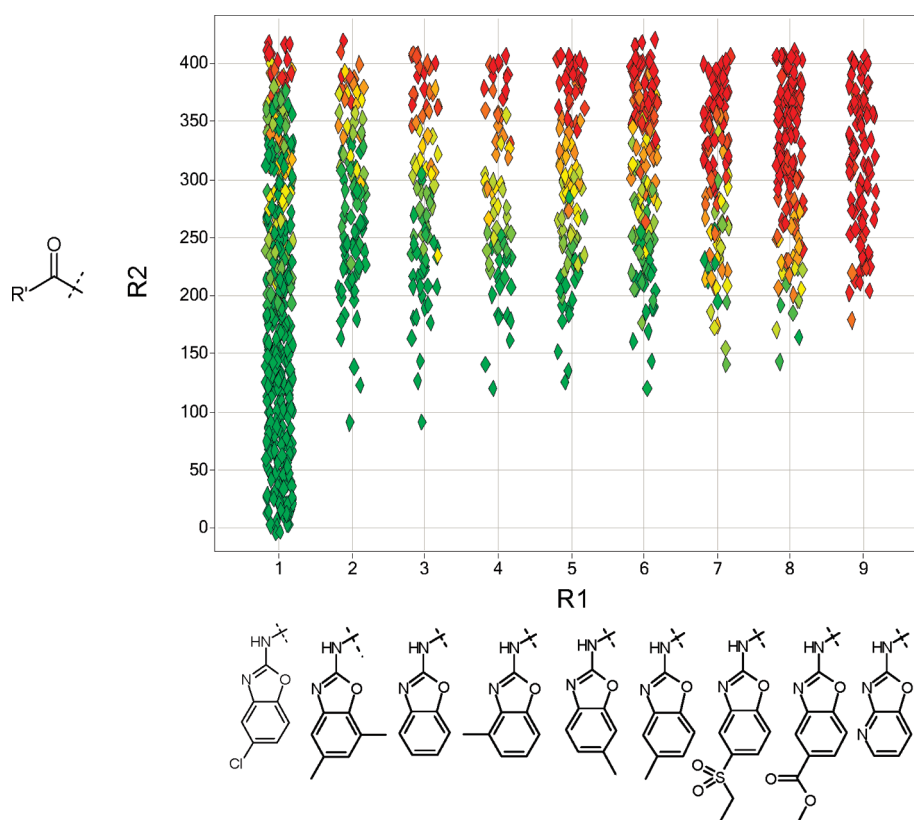
Where  $k$  is the slope of the regression lines (predicted versus observed activities) through the origin. The consensus QSAR prediction scheme was applied to all validation set compounds found within individual applicability domains of models used in consensus prediction.

In addition to external validation, Y-randomization test was carried out to establish model robustness. The test consists of rebuilding models using shuffled activities of the training set and the evaluation of such models' predictive accuracy in comparison with the original model. It is expected that models obtained for the training set with randomized activities should afford significantly lower values of statistical parameters such as  $Q^2$ ,  $R^2$ , etc., for training and, especially, test sets. Therefore, if most QSAR models generated in the Y-randomization test exhibit relatively high values of the statistical parameters for either training or test sets, then it implies that a reliable QSAR model cannot be obtained for the given data set. The test was applied to all QSAR approaches in this study.

**SELC and G-SELC Methods.** The SELC method was developed by Mandal et al.<sup>8</sup> and was a modification of the sequential elimination of levels method.<sup>23</sup> In short, SELC employs principals of experimental design and algorithmic optimization (genetic algorithms) to identify combinations of factors that are likely to

improve upon the desired response. When the SELC is used prospectively, it is initialized with an efficient space covering experimental design, e.g., an orthogonal array. After the experiment is performed based on the orthogonal array, the data are analyzed with a statistical model, which is typically a general linear model, and the statistical significance of each factor is determined. If a factor is significant, then its levels are weighed such that levels that are positively related to the response get more weight, whereas levels that are negatively related to the response get less weight. Levels of factors that are not statistically significant receive equal weight. These weights are then used in a weighted mutation scheme to guide the genetic algorithm toward combinations of levels that are potentially optimal. If the SELC method is used retrospectively, then a model is built on the historical data. Levels of factors that are statistically significant are reweighted, and levels of factors that are not significant retain equal weights.

Next, the design points from the orthogonal array and/or the historical data that yielded the worst performance are placed into a forbidden array. Combinations of factor levels that appear in this array are forbidden to appear in any future design point. For example, suppose the experiment has three factors and the following design point is placed in the forbidden array: (1, 3, 2). The



**Figure 2.** Historical R1, R2 monomer combinations. Each monomer axis is ordered from highest average percent inhibition (left for R1, bottom for R2) to lowest average percent inhibition (right for R1, top for R2). Color gradient indicates percent sEH inhibition for each R1 and R2 combination (green=high, yellow=median, and red=low).

**Table 1.** Three Individual Statistical Models using 88 Descriptors

statistical method	model statistics <sup>a</sup>					
	LGO-CV ( <i>n</i> = 783)		training set ( <i>n</i> = 783)		test set ( <i>n</i> = 194)	
	<i>Q</i> <sup>2</sup>	RMSE	<i>R</i> <sup>2</sup>	RMSE	<i>R</i> <sup>2</sup>	RMSE
Cubist	0.91	11.46	0.94	9.64	0.78	18.11
GBM	0.79	17.30	0.95	8.45	0.81	16.63
PLS	0.75	18.90	0.80	16.86	0.75	19.24

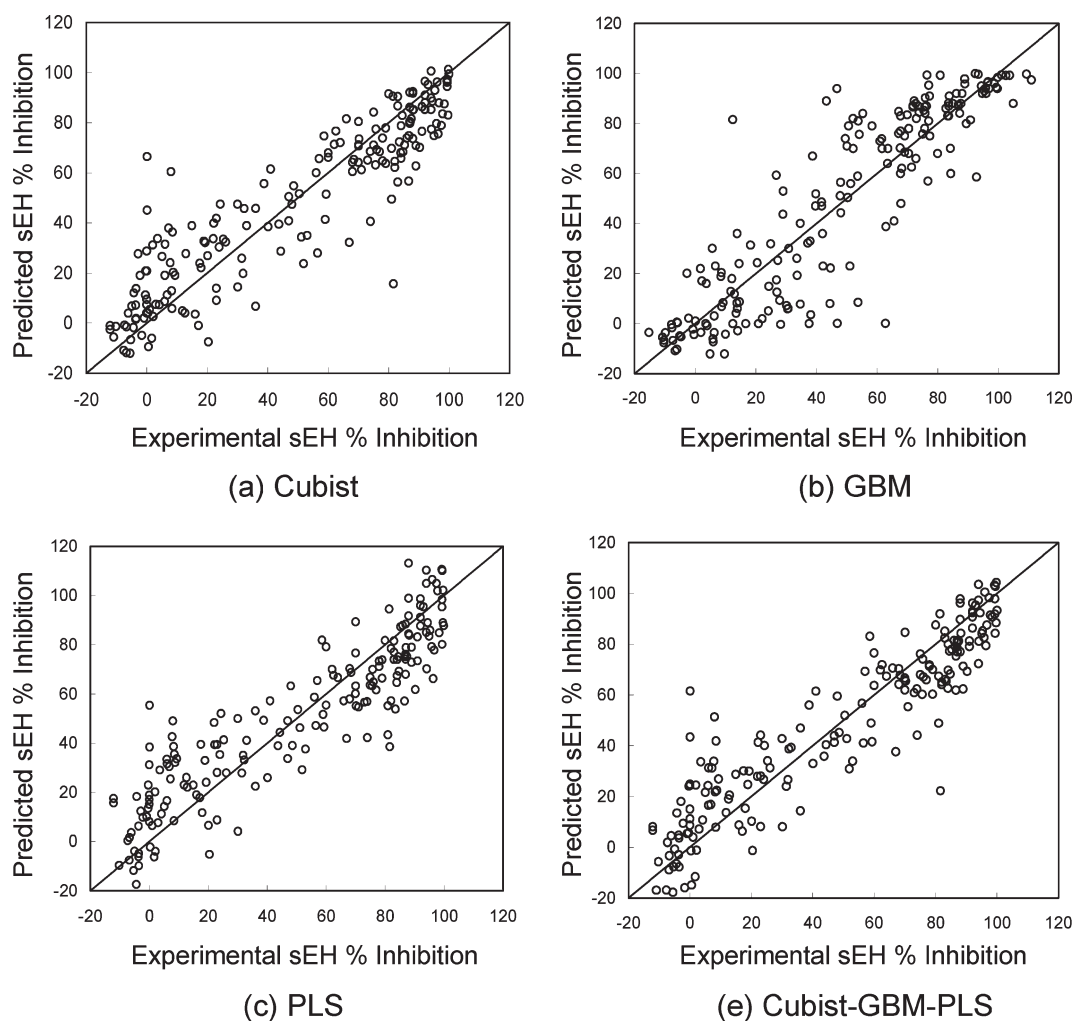
statistical method	Y-randomization test <sup>b</sup>			
	LGO-CV ( <i>n</i> = 783)		test set ( <i>n</i> = 194)	
	<i>Q</i> <sup>2</sup>	RMSE	<i>R</i> <sup>2</sup>	RMSE
Cubist	0.210	32.856	0.094	42.074
GBM	0.005	38.000	0.006	38.156
PLS	0.006	38.100	0.014	38.851

<sup>a</sup> *R*<sup>2</sup> and RMSE. The model statistics are from the LGO-CV regression analysis. The training set and test set statistics are for running the training or test set through the model and comparing the predicted to observed values. <sup>b</sup> Model and test set statistics (observed versus predicted comparison) when building same models as in table *R*<sup>2</sup> using randomized end point data (Y-randomization).

following design points would then not be allowed to be created in future experiments: (1, 3, \*), (1, \*, 2), and (\*, 3, 2), where \* represents any level of a factor. If a new design point is in the forbidden array, then it is removed, and another design point is created. Hence, the forbidden array and the weighted mutation

scheme guide the SELC toward optimal points. This process is continued as budget allows, until a sufficient number of optimal design points are identified or until optimality criteria are met. As Mandal et al. showed, the SELC more efficiently identifies optimal design points than traditional combinatorial design





**Figure 3.** Observed versus predicted sEH percent inhibition for 194 test set compounds run through the individual models Cubist (a), GBM (b) PLS (c), and the consensus model Cubist-GBM-PLS (d).

approaches.<sup>5</sup> The algorithm also has the ability to escape local optimums and search other parts of the combinatorial space.

To enhance the SELC algorithm and improve the efficiency of its search capabilities, Mandal et al. replaced the general linear modeling approach in the SELC algorithm with a Gaussian process (GP) model and termed the enhanced method G-SELC.<sup>5</sup> In addition, G-SELC uses the concept of expected improvement to dampen the greedy search nature of the Gaussian spatial processes.<sup>24</sup> The new G-SELC process can be run in batches as follows:

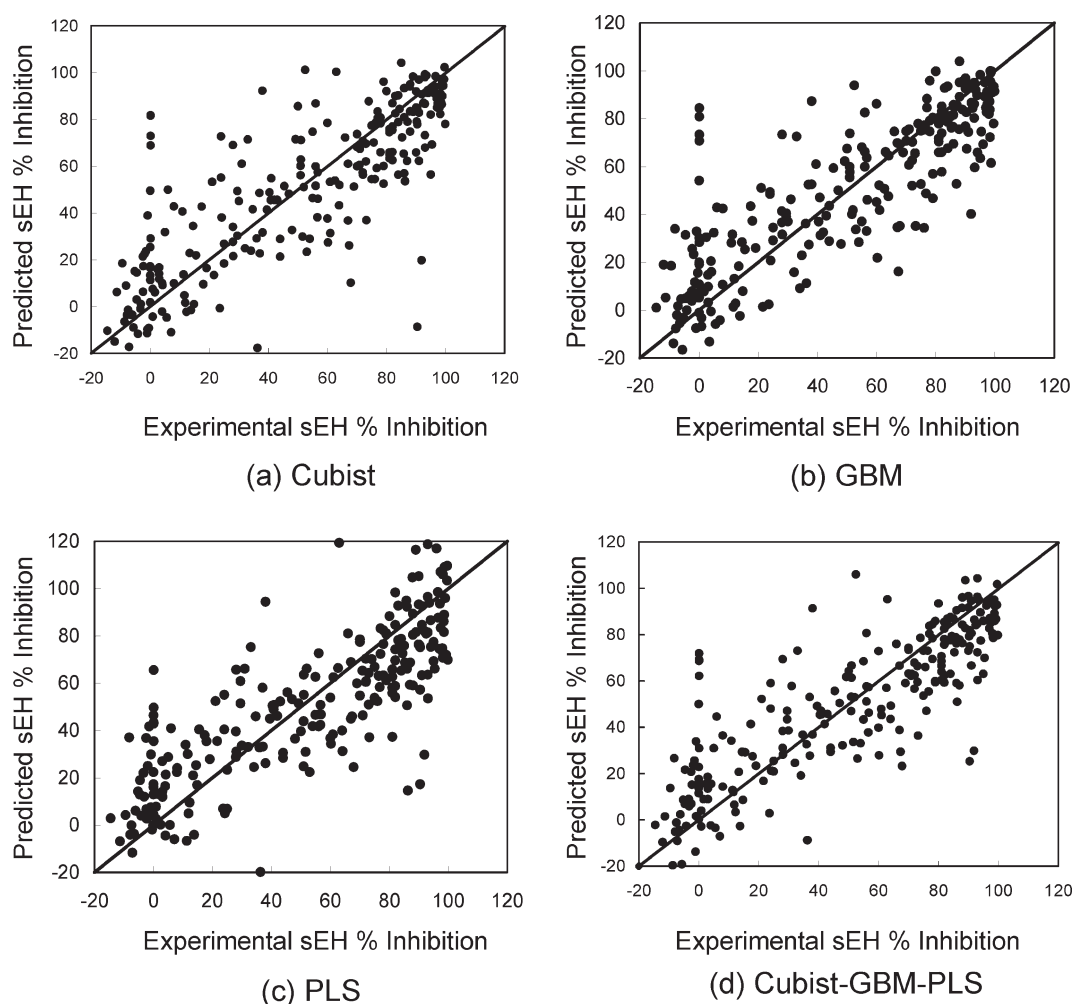
- (1) Select an initial set of design points, i.e., monomer combinations, ideally capturing a balanced, wide range of information about each attachment point.
- (2) Evaluate the new molecules and fit a GP model to the data.
- (3) Choose a threshold response above which compounds are considered optimal and identify those compounds that meet this criterion.
- (4) Determine the unique number of clusters of the optimal compounds, and  $\alpha$ , the proportion of space covered by the optimal clusters relative to the entire designed or observed combinatorial space.
- (5) Generate  $b$  new design points as follows:

**Table 2.** External Validation ( $n = 244$ ) Results for Each of the Individual Models As Well as the Consensus Model

modeling method	statistics		
	$R^2$	$R_0^2$	RMSE
Cubist	0.66	0.61	21.73
GBM	0.70	0.62	20.12
PLS	0.65	0.59	22.07
Cubist-GBM-PLS	0.73	0.65	19.60

- (a) Based on the predictions from the GP model, select the top  $\alpha b$  design points using the expected improvement score.
- (b) Select the remaining  $(1 - \alpha)b$  design points using the SELC algorithm.
- (6) Create the new design points and update the data.
- (7) Repeat steps 2–5 as budget allows or until a sufficient number of optimal design points have been obtained.

As with the SELC, the G-SELC algorithm can be used either prospectively or retrospectively. When using this method retrospectively, the algorithm is seeded with historical data. In our application, we have a rich set of historical data, which we used to initialize the G-SELC algorithm.



**Figure 4.** Observed versus predicted sEH percent inhibition for 244 external validation compounds run through the individual models Cubist (a), GBM (b), PLS (c), and the consensus model Cubist–GBM–PLS (d).

## RESULTS AND DISCUSSION

**sEH Compound Library.** The sEH library employed 9 R1's and 416 R2's as foundational monomers in an amide coupling reaction to form the benzoxazole products (Scheme 1). Of the 3744 potential combinations of R1 and R2, 1221 (33% of the combinatorial space) were synthesized and tested for sEH inhibition in percentage. Figure 1 represents the distribution of percent inhibition for the combinatorial library. Clearly, the historical library covers a broad range of percent inhibition space, with modalities at 0 and 100. Figure 2 displays the coverage of the R1, R2 combinatorial space, where the monomers are ordered by highest average percent inhibition (lower left, blue) to lowest (upper right, red). For example, monomer R1 = 1 has the highest average percent inhibition (across all R2s), while R1 = 9 has the lowest average percent inhibition. This figure helps understand the relationship between the sEH inhibition and the monomer combinations as well as the synthetic coverage of the 2D combinatorial space. Apparently, chemists had made more follow-up combinations with the first monomer R1 = 1 than the ninth R1 = 9.<sup>4</sup>

Figure 2 manifests a number of important characteristics of this library. First, the gradient in color from lower left to upper right is not linear. This indicates that the relationship between different monomers is not additive in their effect on sEH inhibition.

Second, the historical combinatorial library was biased toward using the first R1 monomer: 92% of the possible R2 monomers are paired with this R1 monomer, while the remaining R1's were on only coupled with 17–30% of the R2's. Furthermore, the figure illustrates that more than 100 of the R2 monomers showing high sEH inhibition in combination with the R1 = 1 have not been combined with R1 monomers 2–9. Given this historical information, the next rounds of combinatorial synthesis should be guided toward exploring novel SAR space with a higher propensity of producing sEH potency using R1 reagents from 2 to 9.

**Regression Analysis.** Although a number of the regression methods yielded reasonable models based on their  $R^2$  and RSME values (Table S1, Supporting Information), visualizing the plots for the training and test sets by observed versus predicted values revealed systematic biases by certain methods including MARS, RF, SVMr, and SVMp methods (Figure S2, Supporting Information). Combining this information with the model statistics and the accuracy they exhibited in predicting the training and test sets, the GBM, PLS, and Cubist methods were chosen to build the final models. To further reduce the number of descriptors used in the final models, only the variables found to have a 'variable importance' greater than or equal to 65% were preserved. Merging the variables from the GBM, PLS, and Cubist

Table 3. Statistics for Final Models Using Full Data Set

modeling method	training models <sup>a</sup>			
	model statistics		training set statistics	
	LGO-CV $R^2$	LGO-CV RMSE	$R^2$	RMSE
Cubist	0.91	11.27	0.97	6.19
GBM	0.80	17.00	0.98	4.92
PLS	0.73	19.70	0.78	17.57

modeling method	Y-randomization test <sup>b</sup>			
	model statistics		training set statistics	
	LGO-CV $R^2$	LGO-CV RMSE	$R^2$	RMSE
Cubist	0.290	31.592	0.001	39.70
GBM	0.006	38.000	0.028	37.32
PLS	0.007	38.200	0.051	38.55

<sup>a</sup> Statistics for building models using all the data ( $n = 1221$ ). The  $R^2$ 's are for the model and the correlation of the predicted versus observed when running training data through model. <sup>b</sup> The same statistics are shown for randomizing the end point before building the models.

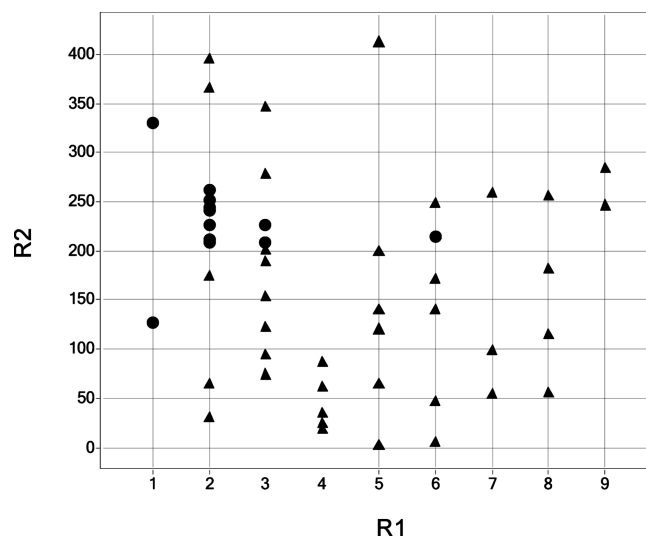
methods that passed the criterion resulted in a set of 88 descriptors to be used in the models construction, which are listed in the Supporting Information.

Table 1 displays the statistical results for the Cubist, GBM, and PLS models using the selected descriptors, including the training and test sets of compounds. Good statistical measures were observed, with training set LGO-CV  $Q^2$  and test set  $R^2$  of 0.91 and 0.78, 0.79 and 0.81, and 0.75 and 0.75 for Cubist, GBM, and PLS models, respectively. The observed versus predicted sEH inhibition for 194 test set compounds are plotted in Figure 3 for the three models. For the most part, the data points display an even distribution along the 45° line, suggesting an unbiased quality of these models.

To ensure that the models did not merely capture noise, the end points were randomized relative to the descriptors within the training set, and the models were rebuilt using the randomized data. As expected, all models using shuffled activity data produced close-to-zero LGO-CV  $Q^2$ 's and  $R^2$ 's (Table 1), corroborating the robustness of the original models. Furthermore the plots of the observed versus predicted for the test set compounds from Y-randomization showed absolutely no correlation (Figure S3, Supporting Information). These results confirmed that the statistical models uncovered nonspurious correlations between the molecular descriptors that were selected for the investigational compounds and their sEH inhibitory activities.

**Consensus Model.** It has been demonstrated that more accurate results can be obtained by consensus, that is, by averaging predictions from multiple QSAR models.<sup>25–27</sup> We took this approach to build a consensus model using combinations of the best individual models. Since Cubist, GBM, and PLS models showed high statistical significance and appeared to be robust by training and test sets predictions as well as the end point randomization validation, a consensus model was built using all three individual models.

The resulting training and test set  $R^2$ 's for the Cubist–GBM–PLS consensus model is 0.89 and 0.81, respectively. Setting the  $y$ -intercept for the trend-line to be 0 yielded  $R_0^2$  of



**Figure 5.** G-SELC suggested 50  $R_1$ ,  $R_2$  combinations. The EI combinations are represented by a circle (●), while the sequential elimination of level combinations (SELC) is represented by triangles (▲). EI focuses on improvement through the  $R_1$  monomers (i.e., 11 of 12 combinations are with  $R_1 = 1–3$ ), while the SELC focuses on improvement through the  $R_2$  monomers (i.e., 21 of 38 are with  $R_2 < 150$ ).

0.87 and 0.79 for the training and test sets respectively.  $R^2$  reflects the similarity in relative ranking of compounds based on actual versus calculated activities rather than the accuracy of the activity prediction and has been traditionally applied as an indicator of the predictive power of statistical models. On the other hand,  $R_0^2$  directly compares the actual versus predicted activities because it estimates the fitness of the data to the line with the intercept of 0 and the slope of 1, thus giving a better measurement of how well the model predicts compounds' activities.<sup>28,29</sup> The high  $R^2$  and  $R_0^2$  values conjunctively indicate good performance of model predictions. Figure 3d shows the observed versus predicted results for the test set compounds from the consensus model. In this case, the consensus model yielded moderately higher correlations between the observed and the predicted activities than the individual models for both the training and the test data sets.

**Model Validation Using External Data Set.** It has been previously reported that training set accuracy does not necessarily correlate with model performance for external data sets,<sup>30</sup> thus the external validation for evaluating the model robustness is critically needed. In this study the external validation set contains sEH inhibitors that were fully independent of the 977 (783 training and 194 test sets) compounds used for building the statistical models. Therefore, the practice can be considered a real test of the predictability of the QSAR models.

The individual Cubist, GBM, and PLS as well as the consensus Cubist–GBM–PLS models were used to predict the inhibitory activity of the external validation set (Table 2 and Figure 4). The results suggest that both the individual and the consensus models afforded reasonable performances, yielding  $R^2$  in the range of 0.65–0.73. Among them the Cubist–GBM–PLS consensus model showed the best performance, with an  $R^2$  of 0.73,  $R_0^2$  of 0.65, and RMSE of 19.7 for 244 compounds in the external validation set. A robust model shall pass the tests of all three indicators of  $R^2$ ,  $R_0^2$ , and RMSE, and this is showcased by the Cubist–GBM–PLS consensus model as well as the Cubist, GBM, and PLS individual models.

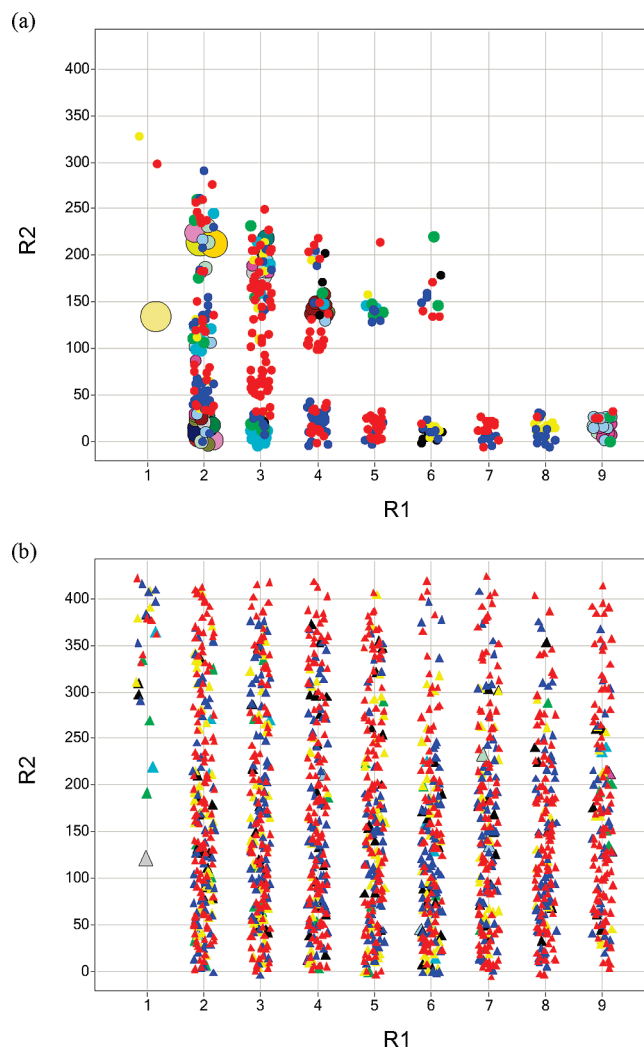
All models showed respectable performances that were consistent with the internal cross-validation and the test set prediction. The numbers of compounds in the external set that are predicted to be within  $\pm 10\%$  range of the experimental sEH activities are: 126 (51.6%) for Cubist, 130 (53.3%) for GBM, 97 (39.8%) for PLS, and 124 (50.8%) for Cubist–GBM–PLS. This range is considered to be within the experimental signal-to-noise ratio. It has been previously shown that the consensus models afford higher prediction accuracy for the external validation data with the highest space coverage as compared to individual constituent models.<sup>22,31</sup> The same observation was made in the present study as well.

An atom-pair similarity measure was used to gain inference on the confidence of prediction. At a given similarity level the number of training set data points similar to the predicted data point is determined. It is generally true that the more training set data points there are with a similarity cutoff to the predicted data point, the higher the confidence in the prediction.<sup>32</sup> At the Tanimoto similarity level of 0.7 the majority of the compounds was found to be similar to at least one compound in the training set, for compounds in both the test and external data sets. Specifically, only two compounds from the test set and eight from the external validation set do not have any similar pairs in the training set at 0.7 Tanimoto cutoff. At an even higher similarity measure of 0.8, the number of compounds dissimilar to the training set was still low, 11 in test set and 23 in external set specifically, in respect to the total number of compounds in the test ( $n = 194$ ) and external ( $n = 244$ ) data sets. These results suggested high confidence in prediction based on the overlapping chemical spaces between the training and test compounds.

**Final Models.** Our goal was to fully exploit the diversity of the chemical space defined by the experimental data available. The extensive validation analysis described previously demonstrated the robustness of the models, allowing selection of the most important descriptors in combination with the suitable statistical methods. In order to maximize the flexibility and predictability of the QSAR models, the final models were built by pooling all available data into the training set, including as diverse a set of compounds as possible. The same set of 88 descriptors were used, and GBM, PLS, and Cubist models were built using the same settings as before on all 1221 compounds. Table 3 shows the statistics for the three models as well as the statistics when applying a Y-randomization to the data. Likely owing to the rigorous model development procedures that were employed, the quality of the models preserved after expanding to the full data set as indicated by the high LGO-CV  $Q^2$ 's (0.73–0.91) and  $R^2$ 's (0.78–0.98). For a given compound the consensus of the Cubist–GBM–PLS model would afford the final prediction.

**G-SELC Compound Selection.** Building upon the historical compound data set, we used the G-SELC algorithm to suggest the next round of compounds to create. Two approaches were employed: First, we seeded the algorithm with the historical data, requested 50 new combinations of R1's and R2's, and evaluated these compounds using the final consensus QSAR model. Second, we ran the algorithm 100 times and asked for 50 new combinations each time. We then looked at the distribution of new combinations to understand how the algorithm explores the combinatorial space.

Before we provide the results of the algorithm, it is worthwhile to revisit Figure 2. In this figure, the R1 and R2 monomers are ordered from highest to lowest average response (left to right on the  $x$ -axis, bottom to top on the  $y$ -axis). Points are also colored by



**Figure 6.** Distribution of G-SELC R1, R2 combinations from 100 iterations: (a) Distribution of combinations selected by EI, (b) distribution of combinations selected by SELC. The sizes and colors of the markers represent selection frequencies. EI focuses on combinations that are most likely to produce desirable responses. SELC is more divergent in its selection of combinations, exploring a wide range of space while also capturing combinations that are likely to produce desirable responses.

activity (blue = highest activity, white = median activity, and red = lowest activity). In this ordering scheme, the optimal combinations in terms of sEH inhibition appear in the lower left of this graph, while the worst are in the upper right. Hence, we would expect any algorithm to be driven toward the lower left.

The original data were input to G-SELC, and 50 new combinations were requested in this space. Figure 5 presents the combinations suggested by the algorithm: expected improvement (EI) suggested 12 combinations, while SELC suggested 38. Based on the algorithm, EI is directed toward combinations where we would expect the most improvement, while SELC incorporates randomness through genetic algorithms to find optimal combinations and avoid local optimums. In these 50 combinations, EI focuses on improvement through the R1 monomers (i.e., 11 of 12 combinations are with R1 = 1–3), while SELC focuses on improvement through the R2 monomers (i.e., 21 of 38 are with R2 < 150). In this example, EI is directed toward already well-explored high-activity space, whereas SELC



**Table 4.** Final Prediction of the G-SELC Suggested Compounds Using Consensus Cubist–GBM–PLS mode, Nearest Neighbors in the sEH Data Set, and Their Experimental Percent Inhibition, Using a Similarity Cutoff of 0.7

ID	% sEH inhibition predicted	number of nearest neighbors	min/max similarity at cutoff	min/max %inhibition of training set similar compounds	average %inhibition of similar training set compounds
1	67.1	52	0.70/0.91	−10.2/93.8	35.8
2	94.7	5	0.79/0.92	81.4/94.0	90.3
3	39.3	30	0.70/0.91	−12.0/94.0	47.3
4	82.1	19	0.70/0.89	−12.1/98.9	79.1
5	49.4	2	0.71/0.89	90.4/97.8	94.1
6	43.2	11	0.70/0.91	−1.01/90.6	49.3
7	34.5	13	0.70/0.90	−11.2/90.6	46.1
8	−2.7	15	0.71/0.90	−6.51/85.6	17.9
9	52	8	0.72/0.91	80.0/99.7	91.7
10	87	13	0.70/0.90	89.0/99.7	95.7
11	77.1	29	0.70/0.91	−9.5/94.0	50.6
12	41.5	1	0.91/0.91	88.0/88.0	88.0
13	84.3	7	0.73/0.89	35.4/98.8	74.7
14	33.5	12	0.70/0.84	8.00/93.1	45.5
15	49.6	19	0.70/0.91	−14.5/93.1	36.7
16	72.2	10	0.70/0.81	25.0/99.2	76.0
17	46.7	9	0.71/0.85	−1.88/90.6	41.4
18	50.6	8	0.71/0.85	−1.0/90.6	49.3
19	44	7	0.71/0.86	−2.58/79.0	32.6
20	55.1	4	0.71/0.84	0.37/72.2	37.9
21	35.7	11	0.70/0.91	−2.58/81.0	36.4
22	32.8	4	0.70/0.84	0.10/81.0	52.3
23	37.4	9	0.71/0.91	−12.8/77.0	33.6
24	79.8	19	0.71/0.84	37.0/98.6	73.8
25	69.2	12	0.71/0.92	14.9/99.1	85.21
26	66.8	2	0.88/0.94	99.5/99.9	99.7
27	60.8	15	0.71/0.92	0.10/99.4	64.2
28	9.1	3	0.70/0.94	0.00/90.4	30.2
29	66.7	2	0.74/0.94	76.0/85.0	80.5
30	65.3	34	0.71/0.92	0.88/99.8	74.7
31	54.8	3	0.71/0.93	6.0/62.0	25.1
32	21.7	24	0.70/0.93	−8.38/91.0	36.1
33	36.3	1	0.94/0.94	99.7/99.7	99.7
34	72.2	12	0.71/0.93	14.9/99.1	85.2
35	82.1	19	0.71/0.93	54.0/99.7	89.9
36	57.2	2	0.79/0.94	81.0/84.0	82.5
37	65.6	1	0.80/0.80	98.0/98.0	98.0
38	68.9	2	0.74/0.84	99.0/99.7	99.4
39	81	1	0.85/0.85	99.4/99.4	99.4
40	58.7	1	0.82/0.82	99.5/99.5	99.5
41	64.1	9	0.70/0.84	23.6/99.7	85.7
42	17.9	13	0.70/0.88	−6.62/93.1	31.3
43	35.7	20	0.71/0.82	−2.08/96.5	41.1
44	46.5	6	0.74/0.80	54.0/98.4	81.0
45	29.3	4	0.70/0.88	−9.85/97.9	32.6
46	18.3	14	0.70/0.83	−4.72/96.6	26.9
47	14.4	2	0.77/0.84	38.0/58.6	48.3
48	22.8	1	0.83/0.83	98.5/98.5	98.5
49	2.3	20	0.71/0.89	−14.5/93.0	34.7
50	7.6	2	0.74/0.86	50.0/50.0	50.0

is directed toward relatively unexplored and likely high-activity space.

In practice we would run one iteration of the algorithm and submit these combinations of the reagents for follow-up synthesis in the laboratory. In theory, we can run many iterations of the algorithm to identify promising combinations that are regularly identified. To explore commonly identified combinations, we ran the G-SELC algorithm 100 times requesting 50 combinations each time and compiled the results. On average, the algorithm used EI for 19 of the 50 new combinations, while SELC was used for 31 new combinations. Figure 6 displays the distribution of selected combinations by EI and SELC. Data points in the figure are sized by their frequency of selection occurrence. It is clear that EI focuses on combinations that are more likely to produce desirable responses. Nearly all combinations are in the lower-left triangle where the historically highest percent inhibition compounds fall. Hence examining monomers that are frequently selected by the EI part of the algorithm can shed light on the structural characteristics that make a compound desirable. On the other hand, SELC is more divergent in its selection of combinations, exploring nearly the entire space, and enables G-SELC to avoid local optimums associated with the EI part of the algorithm.

It is also important to note that the monomer  $R1 = 1$  is not often selected. This is because the historical set already created many of the  $R1 = 1$  monomer combinations. More selection of  $R1 = 1$  would have appeared in Figure 6 if fewer of the synthesized compounds would have used this reagent in the original data set.

**QSAR Prediction of G-SELC Compounds.** Upon generating 50 new compounds using G-SELC, we evaluated their potency using the final consensus QSAR model. The results are summarized in Table 4. Predictions for the 50 G-SELC compounds correlated very well between subset data and all data training data, using consensus methods (Table S4 and Figure S6, Supporting Information). Twenty-five of these compounds were predicted to inhibit sEH enzyme by more than 50% of its maximum activity at 10  $\mu$ M concentration, which constitutes 50% of the G-SELC selected compounds. This is in keeping with the overall hit rate of the sEH libraries, which upon chemical synthesis and biological assay has yielded a hit rate of 54.4% for the 1D and the 2D libraries combined.<sup>4</sup> It is noted that the 1D library of  $R1 = 1$  was heavily explored upon design and synthesis, for which 383 of the 416 candidate  $R2$ 's were combined with  $R1 = 1$  to form the products, leaving only 8% of the unexplored reagents for  $R1 = 1$  that are subject to G-SELC selection. If we discount  $R1 = 1$  the experimental hit rate for the remaining  $R1 = 2-9$  drops to 38%. In comparison the 50% predicted hit rate for the set of compounds recommended by G-SELC is trended toward higher propensity of sEH activity. The balanced approach of G-SELC delivers chemical diversity as well as biological activity using guidance from prior SAR information, which helps achieve the best outcome in a library design practice.

Furthermore, we examined the percent inhibition range of each of these compounds' nearest neighbors to gauge the predictive validity of the model. Specifically, we selected the nearest neighbors using a Tanimoto similarity cutoff of 0.7. Table 4 lists the consensus model predictions, the number of nearest neighbors for each G-SELC compound, and the average of the sEH activities of the nearest neighbors. All compounds had at least one similar training set compound at the 0.7 similarity cutoff, suggesting a reasonable degree of confidence in model predictions. The

average activities of the nearest neighbor training set compounds are in general consistent with the predicted percent inhibition by the consensus Cubist-GBM-PLS model.

## CONCLUSIONS

We discussed a QSAR approach to model the 1221 chemically diverse compounds from a combinatorial library designed to inhibit human soluble epoxide hydrolase. Six different statistical methods were used in combination with thousands of descriptors independently, out of which Cubist, GBM, and PLS were identified with 88 of the most important descriptors to yield the highest external predictive power. These three models were combined to create a more predictive consensus model. All models were rigorously validated using internal cross-validation and test set prediction as well as external data validation. To expand the chemical space of the original combinatorial library, we applied the G-SELC method to select 50 novel combinations of reagents for the next round of laboratory synthesis. The consensus QSAR prediction was then applied to assess the quality of the suggested lead. The predicted hit rate was 50% as compared to the adjusted historical hit rate of 38%. These results suggest that the G-SELC approach enriches the search toward satisfying the desired optimization criteria, which is to product sEH potency herein. Furthermore, G-SELC can be used retrospectively to help scientists better understand the existing combinatorial space. Examining monomers that are frequently selected by the expected improvement part of the algorithm can shed light on the structural characteristics that make a compound desirable. At the same time, monomers selected by the SELC portion of the algorithm help to explore parts of the combinatorial space that also contain activity but have not been explored otherwise. This study illustrates that the G-SELC method has many important and useful characteristics and should be considered as a useful tool for designing combinatorial libraries.

## ASSOCIATED CONTENT

**S Supporting Information.** Descriptors selected for model building, distribution of training, test and validation data sets, model statistics of multiple regression methods, prediction of G-SELC selection by multiple models, chemical structures of  $R1$ 's, reagent information of the 50 G-SELC selected compounds, and other supplementary tables and figures indicated in the text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [li.xing@pfizer.com](mailto:li.xing@pfizer.com). Telephone: (617) 665-5369.

## ACKNOWLEDGMENT

We thank Drs. Suvit Thaisrivongs and Joseph McDonald for helpful discussion and Dr. Christopher Kibbey for help on isMG. This research is sponsored by Pfizer, Inc.

## REFERENCES

- (1) Inceoglu, B.; Schmelzer, K. R.; Morisseau, C.; Jinks, S. L.; Hammock, B. D. Soluble epoxide hydrolase inhibition reveals novel biological functions of epoxyeicosatrienoic acids (EETs). *Prostaglandins Other Lipid Mediators* **2007**, 82 (1-4), 42-49.

- (2) Node, K.; Huo, Y.; Ruan, X.; Yang, B.; Spiecker, M.; Ley, K.; Zeldin, D. C.; Liao, J. K. Anti-inflammatory properties of cytochrome P450 epoxygenase-derived eicosanoids. *Science* **1999**, *285* (5431), 1276–1279.
- (3) Zhou, J. Z. Structure-directed combinatorial library design. *Curr. Opin. Chem. Biol.* **2008**, *12*, 379–385.
- (4) Xing, L.; McDonald, J. J.; Kolodziej, S. A.; Kurumbail, R. G.; Williams, J. M.; Warren, C. J.; O'Neal, J. M.; Skepner, J. E.; Roberds, S., L. Discovery of potent inhibitors of soluble epoxide hydrolase by combinatorial library design and structure-based virtual screening. *J. Med. Chem.* **2011**, *54* (5), 1211–1222.
- (5) Mandal, A.; Ranjan, P.; Wu, C. F. J. G-SELC: Optimization by sequential elimination of level combinations using genetic algorithms and Gaussian processes. *Annu. Appl. Stat.* **2009**, *3* (1), 398–421.
- (6) Johnson, K.; Mandal, A.; Ding, T., Software for implementing the sequential elimination of level combinations algorithm. *J. Stat. Software* **2008**.
- (7) Mandal, A.; Johnson, K.; Wu, C. F. J.; Bornemeier, D. Identifying promising compounds in drug discovery: genetic algorithms and some new statistical techniques. *J. Chem. Inf. Model.* **2007**, *47* (3), 981–988.
- (8) Mandal, A.; Wu, C. F. J.; Johnson, K. SELC: Sequential elimination of level combinations by means of modified genetic algorithms. *Technometrics* **2006**, *48* (2), 273–283.
- (9) *Molecular Operating Environment*, version 2007.09; Chemical Computing Group: Montreal, Quebec, Canada, 2007; www.chemcomp.com. Accessed March 1, 2009.
- (10) *DRAGON software for molecular descriptor calculations*, version 1.4.2; Milano Chemometrics and QSAR Research Group: University of Milano-Bicocca, Milano, Italy; michem.disat.unimib.it/chm. Accessed February 1, 2009.
- (11) *Daylight Fingerprints*, version 4.83; Daylight Chemical Information Systems, Inc.: Laguna Niguel, CA; www.daylight.com. Accessed February 1, 2009.
- (12) *ClogP*, version 4.3; Biobyte Corporation: Claremont, CA; www.biobyte.com. Accessed August 1, 2008.
- (13) *Cubist*, version 2.04a; Rulequest Research: St. Ives, Australia; www.rulequest.com. Accessed February 1, 2009.
- (14) Friedman, J. H. Multivariate Adaptive Regression Splines. *Annu. Stat.* **1991**, *19* (1), 1.
- (15) Zupan, J.; Gasteiger, J. *Neural networks in chemistry and drug design*. 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999.
- (16) Cramer, R. D. I. Partial Least Squares (PLS): its strengths and limitations. *Perspect Drug Discov* **1993**, *1* (2), 269.
- (17) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5.
- (18) A-Razzak, M.; Glen, R. C. Applications of rule-induction in the derivation of quantitative structure-activity relationships. *J. Comput.-Aided Mol. Des.* **1992**, *6* (4), 349.
- (19) Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38* (4), 367.
- (20) Vapnik, V. *The Nature of Statistical Learning Theory*, Springer-Verlag: New York, 1995.
- (21) Kuhn, M. Package “caret”, *Classification and Regression Training*, version 4.70; The R Project for Statistical Computing: XXX; http://Caret.r-forge.r-project.org. Accessed August 1, 2008.
- (22) Tang, H.; Wang, X. S.; Huang, X. P.; Roth, B. L.; Butler, K. V.; Kozikowski, A. P.; Jung, M.; Tropsha, A. Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J. Chem. Inf. Model.* **2009**, *49* (2), 461–76.
- (23) Wu, C. F. J.; Mao, S. S.; Ma, F. S. SEL: A search method based on orthogonal arrays. *Statistical Design and Analysis of Industrial Experiments*; Marcel Dekker: New York, 1990; pp 279–310.
- (24) Jones, D.; Schonlau, M.; Welch, W. Efficient global optimization of expensive black-box functions. *J. Global Opt.* **1998**, *13*, 455–492.
- (25) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47* (9), 2356–64.
- (26) Shao, L.; Wu, L.; Fan, X.; Cheng, Y. Consensus ranking approach to understanding the underlying mechanism with QSAR. *J. Chem. Inf. Model.* **2010**, *50* (11), 1941–8.
- (27) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J. Med. Chem.* **2005**, *48* (23), 7322–32.
- (28) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J. Med. Chem.* **2006**, *49* (24), 7169–81.
- (29) Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* **2006**, *46* (5), 1984–95.
- (30) Golbraikh, A.; Tropsha, A. Beware of q<sup>2</sup>!. *J. Mol. Graph. Modell.* **2002**, *20* (4), 269–76.
- (31) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48* (4), 766–84.
- (32) Arimoto, R.; Prasad, M. A.; Gifford, E. M. Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. *J. Biomol. Screen.* **2005**, *10* (3), 197–205.