Ensemble Feature Selection: Consistent Descriptor Subsets for Multiple QSAR Models

Debojyoti Dutta,[‡,§] Rajarshi Guha,*[,†,§] David Wild,[†] and Ting Chen[‡]

School of Informatics, Indiana University, Bloomington, Indiana 47406, and Department of Computational Biology, University of Southern California Los Angeles, California 90089

Received December 20, 2006

Selecting a small subset of descriptors from a large pool to build a predictive quantitative structure−activity relationship (QSAR) model is an important step in the QSAR modeling process. In general, subset selection is very hard to solve, even approximately, with guaranteed performance bounds. Traditional approaches employ deterministic or stochastic methods to obtain a descriptor subset that leads to an optimal model of a single type (such as linear regression or a neural network). With the development of ensemble modeling approaches, multiple models of differing types are individually developed resulting in different descriptor subsets for each model type. However, it is advantageous, from the point of view of developing interpretable QSAR models, to have a single set of descriptors that can be used for different model types. In this paper, we describe an approach to the selection of a single, optimal, subset of descriptors for multiple model types. We apply this approach to three data sets, covering both regression and classification, and show that the constraint of forcing different model types to use the same set of descriptors does not lead to a significant loss in predictive ability for the individual models considered. In addition, interpretations of the individual models developed using this approach indicate that they encode similar structure−activity trends.

1. INTRODUCTION

One of the fundamental problems in developing quantitative structure−activity relationship (QSAR) models is to select the relevant chemical descriptors or features that describe the relationship between the compound and its activity. In other words, given a set of n descriptors, we would like to choose k of them to construct our model. This is termed the feature selection problem, and there are two broad classes of methods that can be used, namely, wrapper methods and filter methods.

In traditional QSAR modeling, descriptors are typically selected using a heuristic to maximize some score with respect to a single classifier or regression scheme. This approach is an example of wrapper-based feature selection.[1] A wrapper method essentially consists of two components—the objective function which may be a linear or nonlinear classification (or regression) scheme and an optimization (selection) method to select features for the objective function. Examples of the optimization component include genetic algorithms and simulated annealing. The performance of the classification (or regression) scheme is used to guide the optimization procedure in the selection of descriptors. As a result, the selection procedure is closely tied to the learning algorithm that is used. Thus, for example, we may get one set of descriptors if we are using linear models (such as multiple linear regression or linear discriminant analysis) and another different set if we are using a nonlinear technique (such as a computational neural network, CNN). It is clear that this approach aims to determine the *best* descriptor subset for the modeling technique being used.

Filter methods are also common in QSAR modeling. The difference between filter and wrapper methods is that a filter method does not use any specific classifier or regression scheme to select descriptor subsets. Instead, it only considers the characteristics of the data to perform the selection.[2] The standard procedure of descriptor reduction,[3,4] whereby low-variance and correlated descriptors are removed from an initial, large pool of descriptors to give a smaller, more information rich pool, is an example of a filter-type feature selection method. Other examples include mutual information-based methods[5] and χ^2 methods.[6] In this paper, we focus on wrapper-type methods.

As noted above, wrapper-type feature selection methods try to find the best descriptor subset for the model type that is being used. In general, *best* implies the best predictive performance. However, in many situations, one also requires interpretability. In general, there is a tradeoff between interpretability and predictive ability. For example, a linear regression model is generally more interpretable than a neural network model but is also generally less accurate than a neural network model. This situation results in the common approach, whereby a linear model is developed for its interpretability and a nonlinear model is developed for its predictive ability.[4,7,8] However, recent work on the interpretability of neural network QSAR models[9,10] allows one to provide detailed interpretations of structure−activity trends, in a manner similar to the interpretation of linear regression models,[4,11] though it does involve a number of approximations. Given the ability to interpret both linear and nonlinear models, as well as the superior predictive ability of nonlinear models, it would be advantageous to build both types of models using the same set of optimal descriptors, rather than determining optimal descriptor subsets for the models individually.

* Corresponding author e-mail: rguha@indiana.edu.
† Indiana University.
§ These authors contributed equally to this paper.
‡ University of Southern California Los Angeles.

10.1021/ci600563w CCC: $37.00
Published on Web 04/04/2007

An alternative approach to the development of predictive models is to use an ensemble of models. In this method, one builds more than one model for the same QSAR relationship, either using multiple cases of a specific model type[12] (say, multiple linear regression models) or single instances of multiple model types[13] (such a linear regression and a CNN model). Then, the activities of the unknown compounds are predicted using each of the multiple models, and a consensus is used to arrive at a final prediction. For example, in a classification study, we might train different models such as a linear discriminant analysis (LDA) model or CNN on the training data and then predict the class of unknown compounds using a majority vote from each of the LDA and CNN models. Such methods are known to be more statistically robust.[14] In the scenario where multiple model types are used for ensemble predictions, the question of interpretability is generally ignored. This is understandable since the ensemble approach uses models developed using different descriptor subsets. Clearly, if we are able to develop multiple models of differing types but using the same descriptor subset, we would be able to more easily extract structure−activity trends from the individual models in a consistent manner.

The use of a single subset of descriptors for multiple types of models implies that the subset is not necessarily optimal for any of the individual models. In this paper, we present a simple multiobjective optimization approach that attempts to identify a descriptor subset that tries to minimize the degradation in predictive accuracy of different model types, compared to models built using descriptors selected using traditional feature selection. We term this approach ensemble descriptor selection.

1.1. Related Work. The body of work that is closest to our approach is the very recent literature on multiobjective optimization in chemistry.[15−17] In this technique, several objectives, some of them contradictory, are optimized. Typically, this leads to several solutions that are Pareto optimal.[18] That is, any improvement in one of the objectives will definitely lead to a reduction in some other objective. The set of Pareto optimal solutions may then be examined for selection on the basis of other criteria such as interpretability. Applications include pharmacophore searching,[17] combinatorial library design,[16] and QSAR modeling.[15] In contrast, we use a single composite objective function that does not have contradictory components or terms. Thus, we do not need to study the characteristics of the Pareto optimal set. Also, our approach is closer to a generic framework as we can use any stochastic search method such as simulated annealing or bump hunting.[19]

It should be noted that our goal is not to highlight the use of genetic algorithms for feature selection. The literature abounds with examples that use genetic algorithms for this problem.[20−25] Rather, our goal is to extend the wrapper-based approach to feature selection, which has traditionally focused on identifying an optimal descriptor subset for a single type of model, to the problem of identifying an optimal descriptor subset for multiple types of models simultaneously.

2. METHODS

Consider a data set of n objects with m real-valued descriptors. Also denote the dependent variable as Y. Our goal is to select a subset, S, of the m descriptors such that we maximize an objective function $f(S)$. This objective function could be either to minimize the least-square prediction error of Y or maximize the percent correct classification of Y, depending on the problem. Subset selection is a large mature research area in itself, and for more details, the reader is pointed to Miller.[26] Here, we focus on wrapper methods, and common algorithms for this type of subset selection include some form of stepwise regression or stochastic methods such as genetic algorithms and simulated annealing. However, the feature common to these methods is that they focus on the selection of a descriptor subset for a single type of model. That is, the output of the descriptor selection algorithm is a model (or a set of models) that may be a linear regression, neural network, or some other model type.

On the other hand, we are interested in obtaining a consistent set of descriptors for multiple model types. Thus, we desire to perform descriptor selection such that the selected feature set is optimal for different model types simultaneously. To achieve this, we define ensemble feature selection as follows: Given a $n \times m$ data matrix $\mathbf{X}$, and an optional output column vector $\mathbf{Y}$, and p data models (either regression or classification), choose k out of m descriptors or features such that we jointly maximize either the classification accuracy of all the models within the ensemble or jointly minimize the average least-square error in regression of all the models.

Without a loss of generality, we describe the approach in the context of regression. Our approach is based on a genetic algorithm. Briefly, we start with a set of descriptor subsets, S_i, of specified size. These initial subsets are randomly selected. Each subset is then used to build a neural network model and linear regression model. In contrast to traditional optimization, we define the objective function of the genetic algorithm as

$$f(S_i) = \mathrm{atan}(\mathrm{RMSE}_{\mathrm{CNN}} + \mathrm{RMSE}_{\mathrm{OLS}})$$

where atan is the inverse tangent operator, $\mathrm{RMSE}_{\mathrm{CNN}}$ is the root-mean-square error for the CNN model, and $\mathrm{RMSE}_{\mathrm{OLS}}$ is the root-mean-square error for the linear regression model. The use of the inverse tangent serves to bound the value of the objective function between 0 and $\pi/2$. For the case of classification using, for example, a LDA and a neural network model, we would define the objective function as

$$f(S_i) = \frac{1.0}{\mathrm{atan}[(\mathrm{TC}_{\mathrm{CNN}} + \mathrm{TC}_{\mathrm{LDA}})/2]}$$

where $\mathrm{TC}_{\mathrm{CNN}}$ and $\mathrm{TC}_{\mathrm{LDA}}$ are the percentage correct classifications for the neural network and LDA models, respectively. Once $f(S_i)$ has been evaluated for the whole population, the individual subsets are ranked. The next step involves the selection of pairs of individuals from the population and application of genetic operators, crossover, and mutation. This results in a population of new descriptor subsets, which constitute the next generation. This process is repeated until the value of the objective function converges. The result of this procedure is a set of descriptor subsets that simultaneously optimize the RMSE of the CNN and linear regression model.

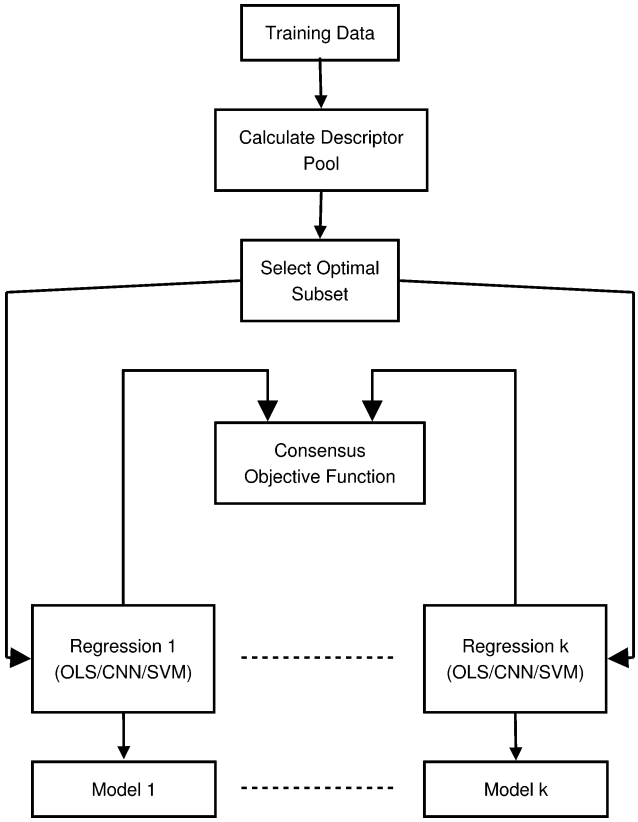

Figure 1. Flow diagram summarizing the ensemble descriptor selection approach.

Given the final, optimal descriptor subset, we can then use the individual models simultaneously (to obtain ensemble predictions) or use them individually for different purposes (such as a linear regression model for interpretation and a neural network model for predictive accuracy) realizing that the individual models now encode the same structure—activity trends.

3. DATA SETS

We tested the ensemble descriptor selection approach on three data sets (see Figure 1 for a diagram summarizing the ensemble descriptor selection approach). The first data set was a collection of 79 platelet-derived growth factor (PDGFR) inhibitors studied by Guha and Jurs.[4] The original data set included observed values of IC_{50} for measurements taken in the presence and absence of human plasma. Since our original study only considered the values for the former case, we used the same values for the current study. Prior to modeling, we converted the observed IC_{50} values to a negative log scale. We calculated a set of 321 topological, geometric, and electronic descriptors using the ADAPT[27,28] software package. This pool was then reduced to a smaller pool of information-rich descriptors by removing low-variance and correlated members, resulting in a reduced pool of 40 descriptors.

The second data set consisted of 435 molecules studied by Fontaine et al.[29] for their ability to inhibit Factor Xa. The reported activity of the compounds was used to divided them into two groups—low-activity and high-activity—and the authors then performed a classification study using the GRiND[30] descriptors to develop a partial-least-squares (PLS) model. Their best PLS model exhibited an accuracy of 88% correct predictions for an external prediction set. Since we did not have access to the GRiND descriptors, we used MOE[31] to evaluate a set of 147 geometric and topological descriptors which were then reduced to an information-rich pool of 62 descriptors.

The third data set consisted of 273 molecules studied by Kauffman ad Jurs[32] for their ability to inhibit cyclooxygenase-2 (COX-2). The original work had evaluated a set of 135 descriptors using the ADAPT toolkit. As described above, we performed descriptor reduction resulting in a reduced pool of 54 descriptors. The study had developed a set of linear regression, CNN, and *k*-NN models. The best reported linear regression model contained eight descriptors and had a RMSE of 0.85 log units for the training set and 0.66 log units for the prediction set.

In all cases, we placed approximately 80% of the data set into a training set and the remainder in an external prediction set. All subsequent calculations were performed using R.[33]

4. RESULTS

Since we did not have access to original predictive models developed for the Factor Xa and COX-2 data sets, we developed a set of models for all the data sets considered in this study. For the PDGFR data set, Guha et al. had developed a three-descriptor linear regression model and a seven-descriptor neural network (with three hidden neurons). In this study, we focused on the three-descriptor case. The original study of the COX-2 data set had developed an eight-descriptor linear regression model, and thus we focused on developing eight-descriptor models for this study. Finally, since we did not have access to the GRiND descriptors used for the Factor Xa study, we considered descriptor subsets of sizes ranging from 4 to 10 descriptors and focused on a seven-descriptor model.

4.1. PDGFR Data Set. Table 1 summarizes the models developed using individual optimization and ensemble optimization. As expected, the CNN model exhibits significantly better performance (0.33 log units) compared to the linear regression model (0.54 log units) for the training set. When we considered the prediction set performance, we observed a RMSE of 0.52 log units for the neural network model and 0.32 log units for the linear regression model. The latter result is surprising since it is significantly lower than the training set RMSE. To investigate the cause of this behavior, we obtained predictions for random subsets of the data set, and we observed the RMSE of such predictions to range from 0.32 to 0.79 log units. Thus, it appears that, due to the small size of the data set, the prediction set performance is sensitive to the molecules placed in it. The linear model was statistically significant, exhibiting an F value of 8.67 on 3 and 59 degrees of freedom ($F_{crit} = 2.76$ at $\alpha = 0.05$). In addition, all the variance inflation factors were greater than 1.0. If we consider the descriptors that were selected for the models, we observe that there is only one descriptor that is common between the CNN and linear regression model, namely, SURR-5, which is the ratio of the weighted hydrophobic surface area to the weighted hydrophilic surface area.[34] This is not surprising since the original work noted that SURR-5 was the most important descriptor according to a random forest model as well as a PLS analysis of the linear model.

Table 1. Summary of the Regression Models Obtained for the PDGFR Data Set Using Traditional Individual Optimization and Ensemble Optimization Schemes

optimization type	model type	descriptors[a]	RMSE[b]	scrambled RMSE
individual	neural network[c]	WTPT-3, SURR-5, FLEX-4	0.33 (0.52)	0.93
	linear regression	SURR-5, RNHS-3, NSB	0.54 (0.32)	0.75
ensemble	neural network[c]	ACHG, NSB, SURR-5	0.37 (0.55)	0.88
	linear regression	ACHG, NSB, SURR-5	0.58 (0.41)	0.83

[a] WTPT-3: sum of path lengths starting from heteroatoms.[36] SURR-5: ratio of weighted hydrophobic SA to weighted hydrophilic SA.[34] FLEX-4: fractional mass of rotatable atoms. RNHS-3: relative hydrophilic SA.[34] NSB: number of single bonds. ACHG: difference between average charge on donatable hydrogens and average charge on acceptor atoms. [b] RMSE: root-mean-square error. The two numbers indicate the RMSE for the training set and prediction set (in parentheses). [c] The neural network architecture was 3-3-1.

If we then consider the CNN and linear regression models developed using the ensemble optimization scheme, we see that, for the neural network model, the training performance has degraded by 12% compared to the model obtained using the traditional descriptor selection approach. The prediction set performance has also degraded by 5%. Similarly, the training set performance of the linear regression model has degraded by 7%, though the prediction set performance experiences a larger degradation. As described previously, such degradation in performance is not surprising and, in general, is to be expected. However, it is clear that the constraint of using the same descriptors in the CNN and linear regression models does not impose significant penalties on their performance. However, even though the linear regression model was slightly degraded in terms of performance, it was still statistically significant, exhibiting an F value of 10.66 on 3 and 59 degrees of freedom ($F_{crit} = 2.76$ at $\alpha = 0.05$).

For both sets of models, we performed y scrambling to ensure that the models did not occur due to chance. For each model, we scrambled the y variable and then rebuilt the model and evaluated the RMSE. This process was repeated 100 times, and the average RMSE of these scrambled models is reported in Table 1. It is expected that if the model encodes a nonrandom relationship between the descriptors and the y variable, the RMSE of a "scrambled'' model will be much higher compared to that of the original model. The results in Table 1 indicate that the models did not encode chance relationships. Another issue that must be considered is the possibility that the models are overfit. This can occur when the model memorizes the features of the training set and consequently exhibits very poor predictive ability. It should be noted that this aspect is linked to the nature of the models built and not the algorithm used to select the features for the models. In Table 1, we see that the prediction set RMSEs are not significantly poorer than the training set RMSEs. Furthermore, the results from the models obtained via ensemble feature selection are quite similar to those obtained using the traditional feature selection approach. Thus, the ensemble approach does not lead to any extra overfitting.

We next consider the descriptors that were selected for the best ensemble model. The three descriptors selected by the ensemble optimization approach were ACHG (the difference between the average charge on donatable hydrogens and average charge on acceptor atoms), NSB (the number of single bonds), and SURR-5. We then performed a PLS analysis of the linear regression model developed using ensemble descriptor selection. The PLS analysis has been used previously for the interpretation of linear regression QSAR models.[4,11] For this study, we did not attempt to perform a full interpretation of the linear model. Instead, Table 2 shows the loadings (X weights) for the three components of the PLS model. The cumulative variance explained by the first components is 61% and by taking all three is 78%. As described in other studies,[4,11] we can assume that the bulk of the structure−activity relationship is described by the first two components. If we then consider component 1, we see that the most important descriptor (as measured by absolute magnitude) is SURR-5, and its weight is negative. This indicates that large values of this descriptor lead to smaller values of activity. This is the same conclusion that was made in ref 4. The next most important descriptor is NSB and has a positive weight, indicating that larger values of this descriptor are correlated with larger values of the activity. Since the number of single bonds is effectively a measure of molecular size, this component indicates that larger molecules are more active. However, all the large molecules in the data set do not exhibit high activity, and thus in the second component, we see that NSB has a negative weight, indicating that small compounds are more active. That is, the second component corrects for the mispredictions made by the second component. Once again, this conclusion is identical to those made in ref 4, except that in the original work molecular size was characterized by the molecular distance edge descriptor.[35] Finally, if we consider the ACHG descriptor, we see that it is the most important descriptor in component 2 and has a negative weight. This implies that smaller values of this descriptor are correlated with higher values of the activity. Since the ACHG descriptor is a difference of charges on H-bond donor and acceptor atoms, a smaller value of this descriptor indicates that the molecule is less polar. Since the activity of the molecules in this data set was measured by a cell-based assay, it is not surprising that molecules which are less polar (i.e., more hydrophobic) will generally exhibit higher activities by virtue of being able to easily pass through the cell membrane. The above discussion indicates that the ensemble feature selection has been able to select a set of descriptors that lead to minor degradation in predictive performance but still encode the structure−activity trends

Table 2. Loadings Obtained from a Partial-Least-Squares Analysis of the PDGFR Data Set Using the Three Descriptors from the Linear Regression Model Obtained Using the Ensemble Descriptors Selection Method

descriptor	component 1	component 2	component 3
ACHG	−2.619	−7.304	4.218
SURR-5	−7.577	−0.542	−4.506
NSB	6.211	−5.060	−3.718

Table 3. Effective Weight Matrix for the Neural Network Model Obtained Using the Ensemble Descriptor Selection Method on the PDGFR Data Set

descriptor	H1	H3	H2
ACHG	245.115	−236.285	−0.597
NSB	50.903	−52.638	4.442
SURR-5	211.904	−199.019	−1.235
SCV	0.52	0.48	0.00

[a] H1, H2, and H3 indicate the hidden neurons. SCV is the squared contribution value.[10]

that were encoded by models obtained using the traditional feature selection method.

Finally, we consider the structure−activity trends encoded in the neural network models developed using the ensemble descriptor selection method. We use the approach described by Guha et al.[10] which linearizes the neural network and develops an effective weight matrix, analogous to the loading matrix used in the PLS-based interpretation of linear models. Table 3 presents the effective weight matrix, where the columns represent the hidden neurons, ordered by their squared contribution values (SCVs) which are shown in the last row. The most important aspect of the weights is the stress on the ACHG and SURR-5 descriptors for hidden neurons 1 and 3. In the first hidden neuron, both ACHG and SURR-5 have positive weights indicating that larger, more polar molecules are predicted to be active. This is in contrast to what we observed for the linear regression model. But if we then consider hidden neuron 3, we see that both these descriptors have large negative weights indicating that smaller, less polar molecules are predicted to be active. Thus, as in the PLS-based interpretations, the third hidden neuron appears to correct for mispredictions made by the second hidden neuron, so that all large polar molecules do not get predicted to be active. However, it is interesting to see that the model does not really focus on molecular size as represented by the NSB descriptor. One possible reason for this is that the SURR-5 descriptor is indirectly a measure of molecular size since it is a function of surface area.

Obviously, we do not expect that a linear regression model and a neural network will embody the exact same structure−activity trends, but it is encouraging to note that, given the same set of descriptors, similar conclusions can be drawn regarding the encoded trends from both types of models.

4.2. COX-2 Data Set. Table 4 summarizes the performance of the neural network and linear regression models developed using traditional and ensemble descriptor selection methods. For the models developed using traditional descriptor selection, the training set RMSE was 0.65 and 0.88 for the CNN and linear regression models, respectively. The linear model was statistically significant with an F value of 15.47 on 8 and 201 degrees of freedom ($F_{crit} = 1.98$ at $\alpha = 0.05$), and all the variable inflation factors were greater than 1.0.

When models were developed using ensemble descriptor selection, there was no change in the training set RMSE for the neural network model and a 1% improvement for the case of the linear regression model. However, the prediction set RMSE for both cases exhibited an improvement: 11% for the neural network model and 16% for the linear regression model. As before, the linear regression model was statistically significant with an F value of 16.34 on 8 and 201 degrees of freedom, with all variable inflation factors greater than 1.0.

We also performed y scrambling to ensure that the models did not occur due to chance factors. As we have described the procedure in the preceding section, we simply note the results in Table 4. The high RMSE on scrambling the y variable, compared to the RMSEs for the original models, indicate that the possibility of chance correlations is low. As before, we note that the difference between the training set and prediction set RMSEs are not significantly large, indicating that the models do not exhibit significant overfitting.

As before, we see that a number of descriptors selected using ensemble descriptor selection are also found in the individually optimized models. More specifically, a number of weighted path descriptors[36] were selected using the ensemble procedure. From Table 4, we see that the CNN and linear regression models developed using traditional descriptor selection both contained a weighted path descriptor (WTPT-5 and WTPT-3, respectively). In addition, both these models used molecular distance edge descriptors[35] (MDE-34 and MDEO-12 for the CNN model and MDEO-22 and MDEO-12 for the linear regression model). The ensemble descriptor selection routine included MDEO-22 in the final descriptor subset. Both the weighted path and molecular distance edge descriptors characterize the number of topological paths between specific atoms. For the MDEO-22 descriptor, these paths are between secondary oxygens. The remaining descriptors selected using ensemble descriptor selection also characterize molecular size, with the exception of MREF (molar refractivity), which in addition to characterizing size can also be considered a measure of the polarizability of the molecule.

Though the original work[32] did not provide an interpretation of the linear regression models that were developed, we provide a brief overview of the relative importance and effects of the descriptors in our linear regression models using the PLS technique used above. Tables 5 and 6 show the loadings (X weights) for the linear regression models developed using traditional descriptor selection and ensemble descriptor selection, respectively. By considering the cumulative X variance, we see that seven components explain 87% of the variance in Table 5, and six components explain 89% of the variance in Table 6. For the purpose of brevity, we only consider the first three components of each table. In Table 5, we see that the two most important descriptors are V7CH (seven-order valence chain χ index[37]) and NCL (number of chlorines), and both have negative weights. In general, the χ chain descriptors are a measure of molecular size. The role of NCL is not entirely clear but could be indicative of polarizability and polarity. In the second component, we see that the two most important descriptors are EMIN (minimum atomic E-state value[38]) and MDEO-12 (molecular distance edge between primary and secondary oxygens), which both have positive weights. The EMIN descriptor characterizes the molecular topology as well as the electronic character of the molecule by combining the valence state electronegativity and the δ or δ^v values. The MDEO-12 descriptor is, again, an indicator of molecular size but specifically considers paths between oxygens. Given the positive weights of these descriptors in this component, we

Table 4. Summary of the Classification Models Obtained for the COX-2 Inhibitor Data Set Using Traditional Individual Optimization and Ensemble Optimization Schemes

optimization type	model type	descriptors[a]	RMSE[b]	scrambled RMSE
individual	neural network[c]	NDB, PND-6, WTPT-5, V6C, V4PC, MDE-11, MDE-34, MDEO-12	0.65 (0.85)	1.23
	linear regression	NCl, V7CH, PND-3, MDEO-22, MDEO-12, EMIN, EMAX, WTPT-3	0.88(0.97)	1.15
ensemble	neural network[c]	WTPT-5, WTPT-4, WTPT-3, NC, MREF, PND-5, PND-3, MDEO-22	0.65(0.76)	1.25
	linear regression	WTPT-5, WTPT-4, WTPT-3, NC, MREF, PND-5, PND-3, MDEO-22	0.87(0.81)	1.14

[a] NDB, number of double bonds; NCl, number of chlorine atoms; NC, number of carbons; MREF, molar refractivity; PND-3, superpendentic index considering nitrogens;[39] PND-5, superpendentic index considering oxygens;[36] PND-6, superpendentic index considering halogens;[39] WTPT-3, sum of path lengths starting from heteroatoms;[36] WTPT-5, sum of path length starting from nitrogens;[36] V6C, sixth-order χ valence cluster;[37] V4PC, fourth-order χ valence-path cluster;[37] V7CH, seventh-order χ chain index;[37] MDE-11, molecular distance edge between primary carbons;[35] MDE-34, molecular distance edge between tertiary and quarternary carbons;[35] MDEO-12, molecular distance egde between primary and secondary oxygens;[35] MDEO-22, molecular distance edge between secondary oxygens;[35] EMIN, minimum atomic E-state value;[38] EMAX, maximum atomic E-state value.[38] [b] The two numbers indicate the RMSE for the training set and prediction set (in parentheses). [c] The architecture of the network was 8-3-1.

Table 5. Loadings Obtained from a Partial-Least-Squares Analysis of the COX-2 Data Set Using the Eight Descriptors from the Linear Regression Model Obtained Using the Traditional Descriptor Selection Method.

descriptor	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6	comp 7	comp 8
WTPT-3	2.771	−6.818	13.952	−2.926	2.332		−1.935	−2.328
EMIN	−4.961	8.593	−7.806	8.379		−1.157	5.550	−3.204
EMAX	−7.400	2.123	8.347	−8.756	1.712	3.050	6.820	2.767
V7CH	−12.077	2.228	−3.052	−0.718	4.835	5.980	−6.721	−2.655
NCL	−11.103	7.811	4.624	−1.903	−3.253	−6.898	−1.575	−1.430
PND-3	−2.510	−8.525	−0.163	−0.665	5.308	−8.757	5.138	−7.829
MDEO-12	7.029	8.687	−0.115	−6.511	5.364	−0.919	−0.311	−8.661
MDEO-22	−0.402	−2.926	1.808	2.050	−9.012	7.709	0.540	−10.717

Table 6. Loadings Obtained from a Partial-Least-Squares Analysis of the COX-2 Data Set Using the Eight Descriptors from the Linear Regression Model Obtained Using the Ensemble Descriptors Selection Method.

descriptor	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6	comp 7	comp 8
WTPT-3	11.390	−6.402	−4.677	−0.573	−6.558	3.058	4.996	−1.331
WTPT-4	1.799	0.545	−12.161	9.476	4.795	1.391	1.372	1.996
WTPT-5	11.553	−7.174	5.132	3.964	−6.596	0.479	1.051	0.412
MREF	−6.284	2.987	−7.697	5.339	−11.072	1.918	−1.418	2.746
NC	−9.361	7.445	−5.090	5.986	−6.681	3.326	−0.573	−3.341
PND-3	1.667	−14.431	1.170	2.903	0.407	1.231	−6.995	−0.755
PND-5	8.351	4.959	−9.145	2.865	0.833	0.151	−9.204	−0.622
MDEO-22	0.533	−1.563	−10.760	8.460	4.703	−5.613	5.127	−1.419

see that they are positively correlated with the observed activity. It is apparent that the focus of the second component is molecular size and, to some extent, polarity. More specifically, larger, more polar molecules will exhibit higher activity. Finally, in component 3, we see that WTPT-3 (sum of path lengths starting from heteroatoms[36]) and EMAX (maximum atomic E-state value[38]) are the most weighted descriptors, and both are positively weighted. The WTPT-3 descriptor is again a measure of molecular size as well as an indicator of polarity. EMAX is a measure of polarity and to some extent size. Thus, this component also broadly indicates that larger more polar molecules are expected to be more active. Overall, from the brief analysis of the descriptors, we see that the linear regression model focuses on molecular size and polarity and indicates that higher values of these descriptors correlate to higher activity.

When we consider the loadings for the descriptors in Table 6, obtained via ensemble descriptor selection, we see many similar aspects. In the first component, we see that the two most weighted descriptors are WTPT-3 (sum of weighted paths starting from heteroatoms[36]) and WTPT-5 (sum of weighted paths starting from nitrogens[36]). Thus, both descriptors characterize molecular size. In addition, since both consider heteroatoms, we can infer that these descriptors also consider polarity. Given that nitrogens are also heteroatoms, WTPT-5 is to some extent redundant, which explains the very similar weights for these descriptors. Both descriptors are positively weighted, indicating that larger and more polar molecules will be predicted active by this component. In component 2, we see that the most weighted descriptors are PND-3 (superpendentic index considering only nitrogens[39]) and NC (number of carbons). Clearly, NC is a measure of molecular size. The superpendentic index is also a measure of molecular size, but focuses on the branched character of a molecular graph[32] and thus characterizes molecular bulk. However, in this component, PND-5 is negatively weighted,

Table 7. Effective Weight Matrix for the Neural Network Model Obtained Using the Ensemble Descriptor Selection Method on the COX-2 Data Set[a]

descriptor	H1	H3	H2
WTPT-5	17.148	−89.547	6.743
NC	68.857	257.125	−2.686
MREF-1	−22.425	−257.601	−0.050
WTPT-4	19.637	55.858	5.751
PND-5	−515.789	152.250	0.981
WTPT-3	30.742	145.523	−3.999
PND-3	8.220	14.565	−6.383
MDEO-22	−8.866	−18.464	−6.091
SCV	0.71	0.29	0.00

[a] H1, H2, and H3 indicate the hidden neurons. SCV is the squared contribution value.[10]

indicating that smaller values of this descriptor (corresponding to less-branched molecules) correlate to higher activity. Since both of the descriptors in this component characterize molecular size, the preceding discussion might appear contradictory. However, if we realize that all "large" molecules may not necessarily be active, then we see that PND-3 balances the effect of NC in component 2 (and the descriptors in component 1) by preventing all large molecules from being predicted as active. That is, component 2 will *correct* for mispredictions made by component 1. Finally, in component 3, we see that the most important descriptors are WTPT-4 (sum of weighted paths starting from oxygens) and MDEO-22 (molecular distance edge between secondary oxygens). Thus, both descriptors characterize molecular size and by taking into account oxygens also characterize polarity. Since both descriptors have negative weights, they indicate that smaller, less polar molecules are likely to be more active. As before, the main effect of this component is to balance the predictions made by the first component, which focuses on large, more polar molecules.

Next, we consider an analysis of the neural network model developed using ensemble descriptor selection. We evaluated the effective weight matrix which is shown in Table 7. For hidden neuron 1, we see that PND-5 is by far the most important descriptor followed by NC. Thus, the focus of this hidden neuron is molecular size and suggests that smaller, less polar molecules will be more active. If we then consider hidden neuron 3, we see that the most important descriptors are now NC and MREF, followed by PND-5 and WTPT-3. It is interesting to note that the descriptors related to molecular size and branching (NC, PND-5 and WTPT-3) now have positive weights, indicating that larger molecules are expected to be active. However, this trend is balanced by the negative weight on MREF, which is a measure of molecular volume.

Obviously, we have not gone into great detail in terms of interpretation, but it is clear that both linear regression models embody the same structure−activity trends. In addition, we see that the structure−activity trends encoded in the neural network are quite similar (mainly differing in relative ordering) to those encoded in the linear regression model with the same descriptors. As noted by Kauffman and Jurs,[32] the difference in size between the central channel of COX-2 and COX-1 has been the focus of designing selective inhibitors. Thus, larger molecules are expected to preferentially enter the COX-2 channel. The discussion of the preceding models clearly indicates that this aspect characterizes the activity of the data set studied here.

4.3. Factor Xa Data Set. Previous work on the Factor Xa data set involved the development of a PLS-classification model using GRiND descriptors. Since we did not have access to these descriptors, we used a set of descriptors obtained from MOE to develop a LDA model and a neural network model. We investigated descriptor subsets of varying sizes and settled on seven descriptors. Table 8 summarizes the statistics of the neural network and LDA models developed using traditional descriptor selection and ensemble descriptor selection. Though not the main focus of this study, it is interesting to note that all the models in Table 8 exhibit similar performance on the prediction set when compared to the original work.[29] When the LDA and neural network models were developed individually, the prediction accuracy for the training set was 99% and 95%, respectively, and 86% and 88% for the prediction set, respectively. When the ensemble descriptor selection procedure was employed, we see that the training set accuracy for the neural network model has decreased by 1% and, for the LDA model, has decreased by a similar value. Clearly, the degradation in accuracy is not significant. For all the models, *y* scrambling was performed as described before. In the case of classification, we can expect that, after scrambling the *y* variable, the resultant classifier should not be able to do much better than random. That is, the percentage correct classification should be around 50%. From Table 8, we see that the classifiers do indeed perform much more poorly when the *y* variable is scrambled, though in each case, the percentage correct classification is above 50%.

If we now consider the descriptors involved in the models developed using traditional descriptor selection, we see that the bulk of the descriptors are VSA descriptors (approximate measures of surface area) weighted either by partial charge or by SlogP.[40] In addition, the LDA model also considers two topological descriptors, Reactive and Diameter. Thus, the main focuses of these models are molecule size, electrostatic interactions, and hydrophobic/hydrophilic interactions. Since we are not aware of any rigorous approach to the interpretation of linear discriminant analysis models, we simply focus on the relative importance of descriptors in the models developed using the ensemble descriptor selection approach. Figure 2 shows a descriptor importance plot for the neural network using the method described by Guha and Jurs.[9] We see that the three most important descriptors for the CNN model's predictive ability are SMR-VSA4, a-nF, and PEOE-VSA.3.1. Both SMR-VSA4 and PEOE-VSA.3.1 measure the surface area of the molecule weighted by molar refractivity (a measure of polarizability) and partial charge. Thus, one many conclude that the neural network model focuses on molecular size and electrostatic effects. If we now consider the linear discriminant coefficients of the LDA model, we see that the three most weighted (in terms of absolute magnitude) descriptors are a-nF, PEOE-VSA.3.1, and Diameter. Though two of the three most important descriptors are common to the three most important descriptors in the neural network model, we see that the LDA model does not appear to focus on polarizability but stresses molecular size moreso.

In the case of this data set, it is not possible to obtain a detailed comparison of structure−activity trends encoded in

Table 8. Summary of the Classification Models Obtained for the Factor Xa Data Set Using Traditional Individual Optimization and Ensemble Optimization Schemes

optimization type	model type	descriptors[a]	% correct[b]	scrambled % correct
individual	neural network[c]	PEOE-VSA-1, PEOE-VSA-5, PEOE-VSA+1, SlogP-VSA9, SlogP-VSA8, PEOE-VSA+3, VSA-POL	99 (86)	59
	linear discriminant analysis	VSA-POL, Reactive, SlogP-VSA2, Diameter, SlogP-VSA4, PEOE-VSA+4, SlogP-VSA1	95 (88)	58
ensemble	neural network[c]	PEOE-VSA+3,a-nF, VSA-OTHER,PEOE-VSA+1, SMR-VSA7, SMR-VSA4, Diameter	98 (86)	58
	linear discriminant analysis	PEOE-VSA+3, a-nF, VSA-OTHER, PEOE-VSA+1, SMR-VSA7, SMR-VSA4, Diameter	94 (86)	65

[a] PEOE-VSA-1, sum of van der Waals surface area for atoms whose partial charges lies between −0.1 and −0.05; PEOE-VSA-5, same as before for atoms whose partial charges lie between −0.30 and −0.25; PEOE-VSA+1, same as before for atoms whose partial charges lie between 0.05 and 0.10; PEOE-VSA+3, same as before for atoms whose partial charges lie between 0.15 and 0.20; PEOE-VSA+4, same as before for atoms whose partial charges lie between 0.20 and 0.25; VSA-OTHER, approximation to the sum of van der Waals surface areas of atoms typed as *other*; VSA-POL, approximation to the sum of van der Waals surface areas of polar atoms; Diameter, largest value in the distance matrix;[41] a-nF, number of flourine atoms. [b] The two numbers indicate the RMSE for the training set and prediction set (in parentheses). [c] The neural network architecture was 7-3-1.

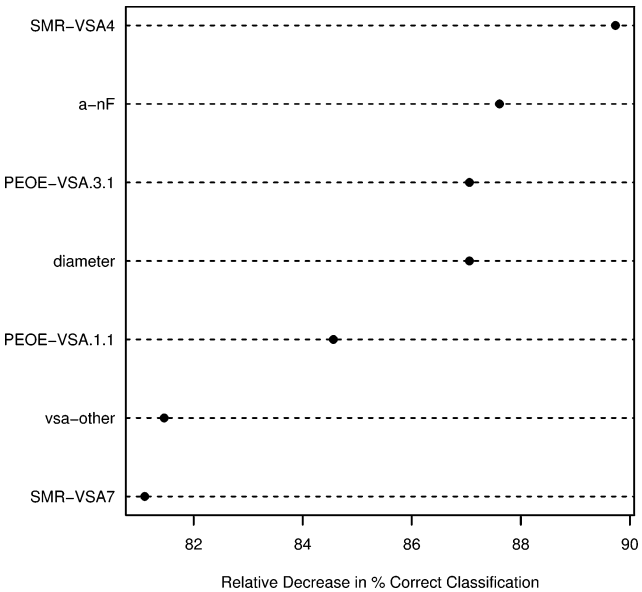

Figure 2. Descriptor importance plot for the neural network model developed for the Factor Xa data set using ensemble descriptor selection.

the different types of models. However, it is apparent that the descriptors obtained using ensemble descriptor selection share many characteristics of the those chosen using traditional descriptor selection. In addition, it appears that the neural network and LDA models developed using ensemble descriptor selection encode similar structure−activity trends. However, it should be noted that, for this data set, given the similar predictive performance of the linear and nonlinear models, one would probably use the LDA model on its own due to its relative simplicity.

5. CONCLUSIONS

In this paper, we have presented an alternative approach to the descriptor selection problem that focuses on the issue of obtaining a descriptor subset that leads to good models of different types. We considered neural network and linear regression models as well as neural network and LDA classification models. Traditional descriptor selection searches the descriptor pool to obtain a descriptor subset that is optimal for a given model type. As a result, different model types often use different descriptors. When such model types are used for ensemble predictions, it can become difficult to derive a comprehensive interpretation of the encoded structure−activity trends.

In contrast, our approach, termed ensemble descriptor selection, searches for descriptor subsets using a genetic algorithm whose objective function is a linear combination of the RMSE or percentage correct classification for the different model types being considered. The net result of this approach is that we get a single descriptor subset that should lead to good models, even though they may be of different types.

One of the main concerns of this approach is that, by constraining the individual model types to the same set of descriptors, their performance would degrade. Our results indicate that, though a degradation in performance is observed, the actual magnitude of the degradation is usually under 12%. However, the models now utilize the same descriptor subset and therefore can be expected to encode similar structure−property trends. Our analysis of the models built for the PDGFR and COX-2 data sets indicates that is indeed the case. It should be noted that the neural network interpretation technique used here does suffer from the linearization step.[10] Though it is possible that some details of the encoded structure−activity trends are not apparent, it is clear that the broad trends are described in a similar manner by the linear and nonlinear models when they use the same descriptor subsets.

Our current implementation employed an objective function which was defined as the inverse tangent of a linear combination of the individual RMSE (or percentage correct

classification) values, in which the individual coefficients were equal. By altering the linear combination coefficients, one could easily *bias* the descriptor selection routine to select descriptors that lead to better (in terms of predictive performance) models of one type at the cost of the other model type. In addition, though we have only considered pairs of model types, there is no reason why more than two types of models could not be considered.

In summary, we believe that this approach to descriptor selection will benefit ensemble-based approaches as well as scenarios where a linear model is used to perform a detailed interpretation and the nonlinear model is used for its higher accuracy.

ACKNOWLEDGMENT

This work was supported by NIH Grant No. NIH-NHGRI/P20 HG 003894-01.

REFERENCES AND NOTES

(1) Kohavi, R.; John, G. H. Wrappers for Feature Subset Selection. *Artif. Intell.* **1997**, *97*, 273−324.
(2) Duch, W. Filter Methods. In *Feature Extraction: Foundations and Applications*; Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., Eds.; Springer: Berlin, Germany, 2006; Vol. 207.
(3) Goll, E.; Jurs, P. Prediction of the Normal Boiling Points of Organic Compounds From Molecular Structures with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974−983.
(4) Guha, R.; Jurs, P. The Development of Linear, Ensemble and Non-Linear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.
(5) Tarca, L.; Grandjean, B.; Larachi, F. Feature Selection Methods for Multiphase Reactors Data Classification. *Ind. Eng. Chem. Res.* **2005**, *44*, 1073−1084.
(6) Liu, Y. A Comparative Study on Feature Selection Methods for Drug Discovery. *J. Chem. Inf. Model.* **2004**, *44*, 1823−1828.
(7) Mattioni, B.; Jurs, P. Prediction of Dihydrofolate Reductase Inhibition and Selectivity Using Computational Neural Networks and Linear Discriminant Analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *21*, 391−419.
(8) Serra, J.; Thompson, E.; Jurs, P. Development of Binary Classification of Structural Chromosome Aberrations for a Diverse Set of Organic Compounds from Molecular Structure. *Chem. Res. Tox.* **2003**, *16*, 153−163.
(9) Guha, R.; Jurs, P. Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance. *J. Chem. Inf. Model.* **2005**, *45*, 800−806.
(10) Guha, R.; Stanton, D.; Jurs, P. Interpreting Computational Neural Network QSAR Models: A Detailed Interpretation of the Weights and Biases. *J. Chem. Inf. Model.* **2005**, *45*, 1109−1121.
(11) Stanton, D. On The Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423−1433.
(12) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R.; Song, Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−799.
(13) Votano, J.; Parham, M.; Hall, L.; Hall, L.; Kier, L.; Oloff, S.; Tropsha, A. QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing Structure-Information Representation. *J. Med. Chem.* **2006**, *49*, 7169−7181.
(14) Dietterich, T. Ensemble Learning. In *The Handbook of Brain Theory and Neural Networks*, 2nd ed.; Arbib, M., Ed.; MIT Press: Cambridge, MA, 2002.
(15) Nicolotti, O.; Gillet, V.; Fleming, P.; Green, D. Multiobjective Optimization in Quantitative Structure−Activity Relationships: Deriving Accurate and Interpretable QSARs. *J. Med. Chem* **2002**, *23*, 5069−5080.
(16) Gillet, V.; Khatib, W.; Willett, P.; Fleming, P.; Green, D. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375−385.
(17) Cottrell, S.; Gillet, V.; Taylor, R.; Wilton, D. Generation of Multiple Pharmacophore Hypotheses Using Multiobjective Optimization Techniques. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 665−682.
(18) Osborne, M.; Rubenstein, A. *A Course in Game Theory*; MIT Press: Cambridge, MA, 1994.
(19) Friedman, J. H.; Fisher, N. I. Bump Hunting in High-Dimensional Data. *Stat. Comput.* **1999**, *9*, 123−143.
(20) Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, in press.
(21) Kah, M.; Brown, C. Prediction of the Adsorption of Ionizable Pesticides in Soils. *J. Agric. Food Chem.* **2007**, in press.
(22) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, in press.
(23) Johnson, S.; Jurs, P. Prediction of the Clearing Temperatures of a Series of Liquid Crystals from Molecular Structure. *Chem. Mater.* **1999**, *11*, 1007−1023.
(24) Deng, W.; Breneman, C.; Embrechts, M. Predicting Protein−Ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-Learning Methods. *J. Chem. Inf. Model.* **2004**, *44*, 699−703.
(25) Fernandez, M.; Tundidor-Camba, A.; Caballero, J. Modeling of Cyclin-Dependent Kinase Inhibition by 1H-Pyrazolo3,4-d. Pyrimidine Derivatives Using Artificial Neural Network Ensembles. *J. Chem. Inf. Model.* **2005**, *45*, 1884−1895.
(26) Miller, A. *Subset Selection in Regression*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, 2002.
(27) Jurs, P.; Chou, J.; Yuan, M. Studies of Chemical Structure Biological Activity Relations Using Pattern Recognition. In *Computer Assisted Drug Design*; Olsen, E., Christoffersen, R., Eds.; American Chemical Society: Washington, DC, 1979.
(28) Stuper, A.; Brugger, W.; Jurs, P. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(29) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687−2694.
(30) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.
(31) *Molecular Operating Environment*, version 2004.03; Chemical Computing Group Inc.: Montreal, Canada, 2004.
(32) Kauffman, G.; Jurs, P. QSAR and *k*-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.
(33) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005.
(34) Stanton, D.; Jurs, P. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer Assissted Quantitative Structure Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(35) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(36) Randic, M. On Molecular Idenitification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(37) Kier, L.; Hall, L. *Molecular Connectivity in Structure Activity Analysis*; John Wiley & Sons: Hertfordshire, England, 1986.
(38) Kier, L.; Hall, L. *Molecular Structure Description. The Electrotopological State*; Academic Press: London, England, 1999.
(39) Gupta, S.; Singh, M.; Madan, A. Superpendentic Index: A Novel Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 272−277.
(40) Wildman, S.; Crippen, G. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(41) Petitjean, M. Applications of the Radius Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.

CI600563W