Predicting Protein−Ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-Learning Methods

Wei Deng,[†] Curt Breneman,*[,†] and Mark J. Embrechts[‡]

Departments of Chemistry and Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute,
Troy, New York 12180

Received October 30, 2003

Inspired by the concept of knowledge-based scoring functions, a new quantitative structure−activity relationship (QSAR) approach is introduced for scoring protein−ligand interactions. This approach considers that the strength of ligand binding is correlated with the nature of specific ligand/binding site atom pairs in a distance-dependent manner. In this technique, atom pair occurrence and distance-dependent atom pair features are used to generate an interaction score. Scoring and pattern recognition results obtained using Kernel PLS (partial least squares) modeling and a genetic algorithm-based feature selection method are discussed.

INTRODUCTION

Recent advances in molecular biology, X-ray crystallography, and NMR spectroscopy are providing three-dimensional structures of protein−ligand complexes with atomic resolution at an escalating rate. Combined with vast increases in accessible computing power, these advances have facilitated dramatic changes in rational drug design methodologies.

There are two common strategies for approaching structure-based drug design.[1] The first is to start with a set of existing ligands with known binding properties, where the original lead compounds are altered to create diverse families of new candidate molecules. The other method involves the estimation of optimal functional group placements within target binding sites, followed by connection of these functional groups via scaffold structures to generate putative molecules that may be candidates for experimental validation. However, both techniques require a means of rapidly predicting the expected binding energy of the putative structures.

Typically, thousands of virtual candidate molecules are generated for a given binding site and require scoring and rank ordering prior to being considered for synthesis. Therefore, fast protein−ligand scoring functions that can provide useful estimates of ligand binding affinities may be used for synthetic prioritization—a crucial component of efficient drug design.[2,3]

Several methods (force field, empirical, and knowledge-based scoring functions) have been developed to calculate binding free energies. Among these, knowledge-based scoring functions are the most recently developed method and have been used successfully to study protein−ligand interactions.[4−6]

Traditional knowledge-based scoring functions require training sets comprising thousands of protein−ligand complexes. This large training data requirement stems from the statistical definition of a knowledge-based potential as expressed in eq 1:[7]

$$\Delta W_{\mathrm{ij}}(r) = -\,k_{\mathrm{B}}T \ln\left(\frac{g_{ij}(r)}{g_{ref}}\right) \tag{1}$$

where k_{B} is the Boltzmann constant, T is the absolute temperature, and g_{ref} corresponds to an artificially defined reference distribution. The term $g_{ij}(r)$ is a frequency or probability distribution of atom pairs of types i and j at a distance r from each other and is given by[7]

$$g_{\mathrm{i,j}}(r) = \frac{N_{i,j}(r)/4\pi r^2}{\sum_r (N_{i,j}(r)/4\pi r^2)} \tag{2}$$

Here, $N_{i,j}(r)$ is the occurrence of atom pairs i,j at a distance between r and $r + \mathrm{d}r$.

From these definitions, it may be seen that a zero occurrence of any kind of atom pair in the training data cannot be tolerated. Therefore, sufficient data need to be collected and thousands of structures must be examined in order to produce statistically valid results and to ensure the inclusion of some atom types that are rarely seen.[6]

The method described in this work was inspired by the knowledge-based scoring function defined by Gohlke et al.[7,8] and uses the 17 atom type assignments that are listed in Table 1, resulting in 289 possible atom pairs. The algorithm presented in this work counts the occurrence of each atom type pair that incorporates atoms from both ligand and binding site within a certain distance range and uses this information to form descriptors for quantitative structure−activity/−property relationships (QSAR/QSPR) analysis. In our initial approach, all atom pairs with distances between 1 and 6 Å were considered in the analysis.

To achieve better predictive results, distance dependence was also considered within the protein−ligand atom pair interactions. To accomplish this, the descriptor algorithm was

[*] Corresponding author phone: (518) 276-2678; fax: (518) 276-4887; e-mail: brenec@rpi.edu.
[†] Department of Chemistry.
[‡] Department of Decision Sciences and Engineering Systems.

10.1021/ci034246+ CCC: $27.50 © 2004 American Chemical Society
Published on Web 01/24/2004

Table 1. Atom Type Assignments[8]

atom type	atoms included
C.3	carbon sp3
C.2	carbon sp2
C.ar	carbon in aromatic rings
C.cat	carbon in amidinium and guanidinium groups
N.3 (N.4)	nitrogen sp3 (positively charged nitrogen)
N.ar	nitrogen in aromatic rings
N.am	nitrogen in amid bonds
N.pl3	nitrogen in amidinium and guanidinium groups
O.3	oxygen sp3
O.2	oxygen sp2
O.co2	oxygen in carboxylate groups
S.3 (S.2)	sulfur sp3 (sulfur sp2)
P.3	phosphorus sp3
F	fluorine
Cl	chlorine
Br	bromine
Met	Ca, Zn, Ni, Fe

modified to take distances between interacting atom pairs into account. In this modification, the distance range was divided into five bins, 1 Å wide. Each descriptor value was then defined as the occurrence of a particular atom pair within a specified distance bin. From the results obtained in this study, it may be seen that the distance-dependent atom type pair descriptors reflect the characteristics of protein–ligand interactions and show promise for rapid protein–ligand scoring.

Earlier work by DeWitte and Shakhnovich also characterized protein–ligand interactions using the occurrence of specific atom pairs to develop a knowledge-based scoring function with only one distance interval (SMoG).[9] An improved scoring function with two distance intervals, a different reference state, and additional parameters was published later (SMoG2001).[10] In the spirit of traditional knowledge-based potentials, SMoG technology defines an artificial reference state for comparison with the frequencies of occurrence of specific atom pairs. In contrast, the method described in this paper considers only the occurrence and distance range of each atom pair, without the need for comparison against an artificial reference state. Although fatal to some knowledge-based methods, zero occurrences of specific atom pairs are allowed in this method and do not significantly affect the final prediction results.

Therefore, the theory behind the new method is simple but useful. The utilization of QSAR modeling for scoring is flexible and applicable to sparse systems, given that models can be built using various sets of ligand-binding site pairs for specific applications. The statistical learning theory underlying this method is uncomplicated, and a large quantity of training data is not strongly required.

METHODS

Protein–Ligand Complex Data Set. Two data sets were studied, where one contained 61 and the other 105 structurally diverse protein–ligand complexes, respectively. Different types of proteins were present in the data sets. For instance, the 105-complex data set contains aspartic proteases (17), serine proteases (17), metalloproteases (21), human carbonic anhydrases (16), sugar-binding proteases (14), endothiapepsins (10), and other proteins (10). Each of the two data sets contain complete structures for each cocrystallized complex, as well as the unbound protein and ligand structures. All data were obtained from the RSCB Protein Data Bank.[11] The experimental pK_d values for this set of ligand–protein complexes were taken from the literature and were found to range from 1.49 to 11.53.[10]

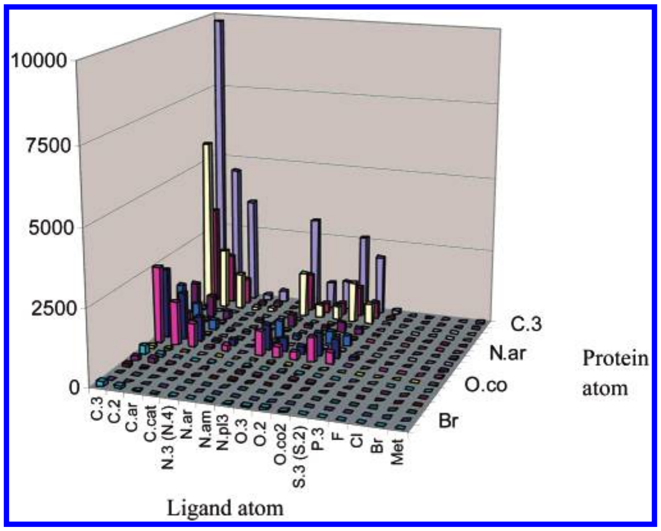

Figure 1. Distribution of atom pairs of all the protein–ligand complexes in the 105-member data set.

Two groups of complexes were randomly selected from the original lists as blind external test sets, one containing 6 complexes for the 61-complex data set and the other containing 10 complexes for the 105-complex data set.

Atom type pair occurrences were then computed for each complex using the simple atom typing and distance algorithm described below. Figure 1 shows the distribution of the occurrences of all atom pairs in the large (105-member) protein–ligand complex data set.

Atom Type Descriptor Generation. All atoms of the proteins and ligands were assigned to one of the 17 Gohlke classes of atom types used in this study. Therefore, each protein–ligand interaction could be characterized as a pattern comprising occurrence counts of atom pairs within a specified cutoff range (6 Å) from ligand atoms.

To take distance dependence into account, the second type of distance-dependent atom type descriptor was developed. The allowed distance range was divided into five bins of 1 Å each, ranging from 1 to 6 Å. Thus, the total number of possible protein–ligand atom pairs grows to 1445, and each descriptor value was then defined as the occurrence of a particular atom type pair within a specified distance bin.

Feature Selection. When the number of descriptors is large, the data set is certain to contain irrelevant and redundant features that increase the dimensionality of the problem and make the model difficult to interpret.[12] This "curse of dimensionality" can also lead to model overtraining.[13] To reduce the number of irrelevant descriptors in the model, a GAFEAT (genetic algorithm feature selection) process is employed to perform objective feature selection before model building.

The genetic algorithm approach is a general optimization method first developed by Holland[14] and involves an iterative mutation/scoring/selection procedure on a constant-size population of individuals.[15] The general process starts with a random or heuristic creation of an initial population, followed by a scoring procedure to determine their "fitness". During every evolutionary generation, the individuals in the

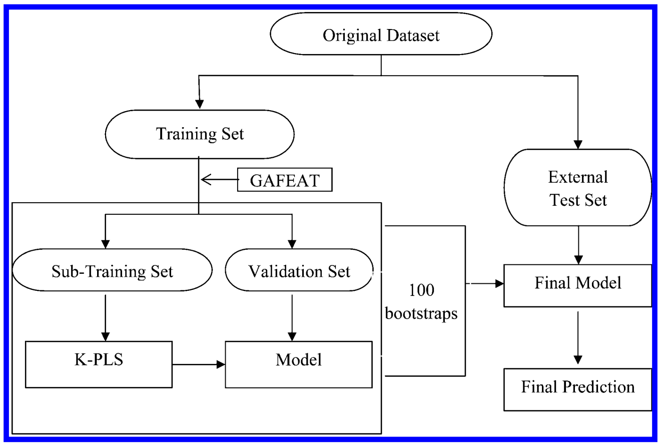

Figure 2. Flow chart of data processing with feature selection.

current population are evaluated according to a predefined set of quality criteria (a fitness function). To form a new population in the next generation, individuals are selected according to their fitness, and new individuals are introduced by crossover and mutation.[16] The following equation shows the fitness function used in the GAFEAT method. In this equation, F_k is the fitness of descriptor k, C_{iR} is the correlation between the descriptor i and the response, C_{ij} is the intercorrelation between descriptor i and j, α is the intercorrelation penalty factor, and β is a "death penalty factor" that is applied if any intercorrelation is more than 0.95. The β factor is assigned a value of 1000.

$$F_{\mathrm{k}} = \sum_{i=1}^{N} C_{iR} + \alpha \sum_{i=1, i\neq j}^{N} C_{ij} + \beta \qquad (3)$$

$$k = 1, 2, 3, ..., \text{population size}$$

Not only can feature selection objectively reduce the number of descriptors without reference to any particular model, but it can also select the most important descriptors that contribute to a given protein–ligand interaction. Although some of the selected descriptors might not be directly attributable to a specific component of a protein–ligand interaction (i.e. hydrogen bonding), they may contain other chemical information relevant to that interaction. An interpretation of the descriptor patterns resulting from subjective (model-driven) feature selection and model building can facilitate an understanding of the fundamental interactions involved in generating a protein–ligand binding score.

Data Processing. The overall model building and testing procedure is illustrated in Figure 2. Following feature selection, the training data are split randomly into a series of training subsets and validation set pairs (in this case, the validation set size if equal to the external test set size). QSAR models are then built using each subset of the training set and evaluated using the corresponding validation set. To generalize the model, this bootstrapping mode is applied to the model-building procedure 100 times (100 bootstraps).[17] Finally, predictions are made on the external test set using an aggregate of all QSAR models from the bootstrapping procedure. The resulting bootstrap aggregate prediction yields a score for each complex.

In this study, the kernel PLS (K-PLS) method was used to produce nonlinear regression models. The K-PLS implementation used in this work is available as part of the RPI Analyze/Stripminer modeling software available on our project website (www.drugmining.com). Figure 3 illustrates how K-PLS works in both training and test modes. K-PLS works in distance space instead of descriptor space and allows a variety of kernel functions to be used to determine distances between descriptor patterns. In the present work, a radial basis function (RBF) kernel was utilized.

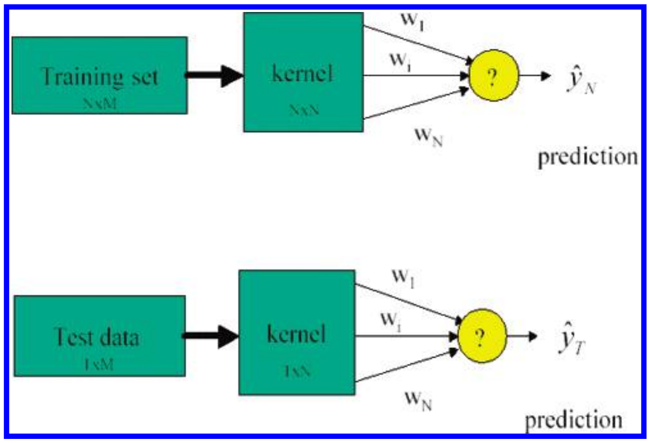

Figure 3. K-PLS prediction in both training and testing modes. N is the number of descriptor patterns (training cases) in the training set, and M is the number of descriptors used in both training and test sets. T is the number of test cases. W_{i} signify kernel input information going into the model, and Y_{r} are the prediction results.

In the training mode, K-PLS internally builds an $N \times N$ kernel matrix from a training set consisting of N compounds and M descriptors. The resulting matrix can be considered a nonlinear transformation of the input data into distance space, after which distances between cases are used as descriptors in a subsequent linear PLS learning process. In the testing mode, a test set with T compounds and M descriptors is multiplied by a modified $T \times N$ kernel matrix, and the results are fed into a previously determined PLS model to make each prediction.[18]

The Analyze/Stripminer shell program was used in this investigation to perform data processing, feature reduction, and predictive modeling. The Analyze/Stripminer program package was developed by one of the coauthors (M.J.E.) and supports several different machine-learning models: kernel partial least squares regression, genetic algorithms, support vector machines, sensitivity analysis, neural networks, and local learning. This program has been utilized successfully for in-silico drug design, economic financially related studies, and nuclear physics applications.[19]

RESULTS

After objective feature selection consisting of the removal of nonchanging features, highly intercorrelated descriptors, and 4-sigma outliers,[20] the remaining descriptors were used in a subjective feature selection and modeling process. Table 2 shows the prediction results (R^2) of the validation sets and external test sets for each data set using both atom type pairs and distance-dependent atom type pair descriptors, both before and after subjective feature selection. Note that the results shown for the validation set predictions are the average values from 100 bootstraps.

Figure 4 and Figure 5 illustrate the K-PLS prediction results for the validation set and test set of model 9, respectively. In Figure 4, the discontinuous vertical bars show

Table 2. Prediction Results (R^2) of the Data Sets Using Atom Pairs and Distance-Dependent Atom Pair Features

model no.	data set size	test set size	descriptor type	no. of features	prediction of validation set after 100 bootstraps (R^2)	prediction of the external test set (R^2)
1	61	6	AT[a]	150	0.30	0.26
2	61	6	AT	25	0.59	0.32
3	61	6	DD[b]	553	0.31	
4	61	6	DD	40	0.71	0.60
5	105	10	AT	130	0.40	
6	105	10	AT	25	0.51	0.60
7	105	10	DD	456	0.43	
8	105	10	DD	40	0.59	0.57
9	105	10	DD	25	0.60	0.60
10	105	10	DD	20	0.60	0.64

[a] AT = atom type descriptors. [b] DD = distance-dependent atom type descriptors.

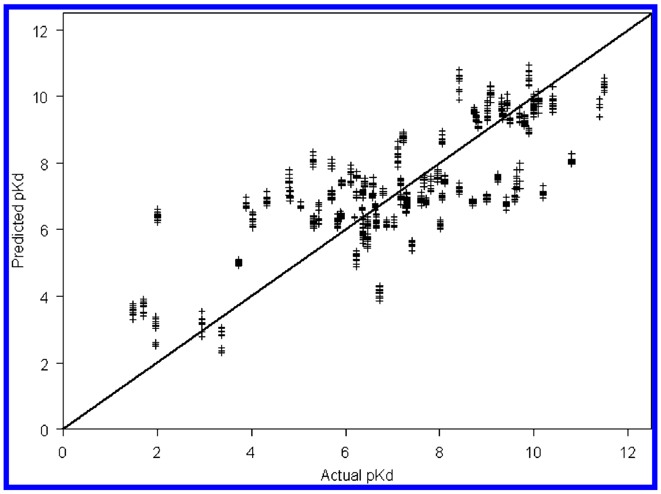

Figure 4. Prediction scatter plot of the validation set of model 9 (K-PLS).

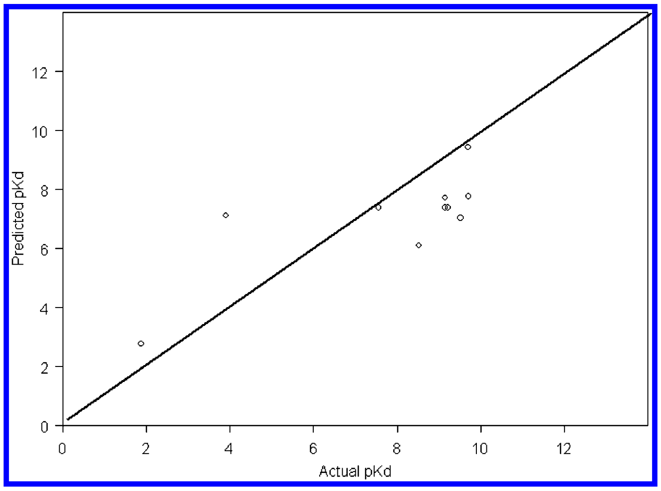

Figure 5. Prediction scatter plot of the external test set of model 9.

the ranges of prediction values for different bootstrap folds in the modeling process.

In an effort to explore an alternative method for nonlinear regression, multiple adaptive regression splines (MARS)[21] predictions were compared to those made by K-PLS models using the same data set. The two methods were found to produce similar results.[22]

A *y*-scrambling process was performed on data sets of models 6 and 9 in which the activity values were randomly shuffled and the data set subjected to the modeling process again. No predictive models could be generated using *y*-scrambled data ($R^2 = 0.01$ for both validation and test sets), indicating that the descriptors contained significant chemical information, and the predictions are not the result of fortuitous correlations or overdetermined models.

DISCUSSION

Drug design is an iterative procedure that starts with a compound displaying an appealing biological profile and ends with a drug candidate with an optimized activity profile. Protein−ligand docking and scoring are challenging but important problems in drug discovery. Instead of studying knowledge-based potentials between ligand and protein atoms using thousands of complexes,[8] the method described in this paper identifies simple descriptors and methods that can be used to build practical QSAR models for scoring.

From Table 2, it can be seen that good predictions of protein−ligand binding affinities can be made and that the feature selection process helps to improve the prediction results. Before objective feature selection, there were 1445 distance-dependent and 289 atom type pair descriptors available for each complex. Models created using full descriptor sets gave predictions that fell below $R^2 = 0.45$. After GAFEAT feature selection, the number of descriptors were fewer than 50, allowing the predictive R^2 values to reach nearly 0.60.

In addition, from Table 2, it could also be seen that the predictions made using distance-dependent atom type descriptors (models 4 and 10) were consistently better than those using only atom type descriptors (models 2 and 6). Therefore, the introduction of distance-dependent features was found to enhance the scoring function. It was shown that the combination of atom type and distance features provides a useful overall representation of topographical and molecular properties of the binding site at the atomic level. Applied as descriptors of protein−ligand interactions, this information helps to characterize the complicated bimolecular system numerically.

One of the advantages of knowledge-based scoring functions over other methods of scoring is that they incorporate physical effects not implicitly presented.[7,8] The feature selection process used in this work identifies descriptors that represent the more important effects in protein−ligand binding and eliminates those which represent insignificant interactions that degrade the efficiency of the scoring functions. In this way, simpler scoring functions are obtained with a great reduction of the computational cost. The bootstrapping mode applied to K-PLS modeling in the present work was found to prevent errors from being introduced due to arbitrary data set partitioning and helps to produce more stable and generalized prediction results.

As seen from the scatter plot of results (Figure 4), the K-PLS model gave good overall predictions for the bootstrap folds in the modeling process. The variability of the modeling results may arise from the relatively coarse characterization of protein−ligand interactions used by this type of scoring method. Because of the moderately small number of protein−ligand complexes studied, the number of descriptors is necessarily limited in this approach. Therefore, even though the introduction of distance-dependent features gave a better geometrical representation of the binding site and improved the prediction results, the number of bins had to be restricted to five in order to prevent a proliferation of descriptors. If larger training sets were available, the number of bins could be increased with less risk of overdetermining the scoring function.

CONCLUSIONS

In this study, a knowledge-based QSAR approach was described that utilizes distance-dependent atom type pair descriptors as a method for generating predictive models of protein−ligand binding affinities. The selection of important descriptors was accomplished using a robust GA-based feature selection algorithm. It was demonstrated that models developed using this technique are applicable to a diverse range of protein−ligand complexes. In the future, this method may prove useful for both the prediction and interpretation of protein−ligand interactions.

In summary, the behavior of protein−ligand scoring functions may be enhanced through the use of geometrical descriptors that incorporate atom type pairs and distances. Machine-learning methods, such as K-PLS, were shown to be effective for the prediction of protein−ligand binding. Extensive cross-validation of the model was accomplished using a bootstrap scheme that involved multiple training and validation sets, as well as a *y*-scrambling routine that indicated that the model was capturing chemical information.

ACKNOWLEDGMENT

The authors would like to thank Professor Alexander Tropsha, School of Pharmacy, University of North Carolina, for providing the crystal structural data. Support from the National Science Foundation (Award No. IIS-9979860) is also greatly appreciated.

Supporting Information Available: Protein−ligand complexes and binding affinities. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Goodford, P. J. Drug Design by the Method of Receptor Fit. *J. Med. Chem.* **1984**, *27*, 557−564.

(2) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(3) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(4) Sippl, M. J. Calculation of Conformational Ensembles from Potentials of Mean Force. *J. Mol. Biol.* **1990**, *213*, 859−883.

(5) Muegge I.; Martin Y. C. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(6) Mitchell J. B. O.; Laskowski R. A.; Alex A.; Thornton J. M. BLEEP-Potential of Mean Force Describing Protein−Ligand Interactions: I. Generating Potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

(7) Gohlke H.; Hendlich M.; Klebe G. Predicting Binding Modes, Binding Affinities and “Hot Spots” for Protein−Ligand Complexes Using a Knowledge-Based Scoring Function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.

(8) Gohlke H.; Hendlich M.; Klebe G. Knowledge-Based Scoring Function To Predict Protein−Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(9) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.

(10) Ishchenko, A. V.; Shakhnovich, E. I. Small Molecule Growth 2001 (SMoG2001): An Improved Knowledge-Based Scoring Function for Protein−Ligand Interactions. *J. Med. Chem.* **2002**, *45*, 2770−2780.

(11) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(12) Moody, J. E.; Hanson, S. J.; Lippmann, R. P. *Advances in Neural Information Processing Systems 4*; Morgan Kaufmann: Denver, CO, 1992; pp 847−854.

(13) Smith, M. *Neural Networks for Statistical Modeling*; Van Nostrand Reinhold: New York, 1993; p 32.

(14) Holland, J. H. *Adaptation in Natural and Artificial Systems*; The University of Michigan Press: Ann Arbor, MI, 1975; Chapter 1, pp 1−19.

(15) Banzhaf, W.; Nordin, P.; Keller, R. E.; Francone, F. D. *Genetic Programming−An Introduction: On the Automatic Evolution of Computer Programs and Its Applications*; Morgan Kaufmann: San Francisco, CA, 1997; pp 125−132.

(16) Mitchell; M. *An Introduction to Genetic Algorithms (Complex Adaptive Systems)*; MIT Press: Cambridge, MA; 1998; Chapter 1, pp 1−31.

(17) Embrechts, M. J.; Arciniegas, F. A.; Ozdemir, M.; Breneman, C. M.; Bennett, K. P.; Lockwood, L. Bagging Neural Network Sensitivity Analysis for Feature Reduction in QSAR Problems. *Proceedings of 2001 INNS−IEEE International Joint Conference on Neural Networks* (Washington, D.C., July 14−19 2001); IEEE Press: Piscataway, NJ, 2001; Vol. 4, pp 2478−2482.

(18) Rosipal R.; Trejo L. J. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *J. Mach. Learn. Res.* **2001**, *12* (2), 97−123.

(19) Embrechts, M. J.; Arciniegas, F.; Ozdemir, M.; Momma, M. Scientific Data Mining with StripMiner^{TM}. In *Proceedings of the 2001 SMCia Mountain Workshop on Soft Computing in Industrial Applications* (Blacksburg, VA, June 25−27, 2001); Embrechts, M. J., VanLandingham, H. F., Ovaska, S., Eds.; IEEE Press: Piscataway, NJ, 2001; pp 13−16.

(20) Breneman, C. M. *Novel Descriptor Selection Methods with Sensitivity Analysis for QSAR,* Book of Abstract, 222nd American Chemical Society (ACS) National Meeting in Chicago, Illinois Aug. 31, 2001.

(21) Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1−141.

(22) The authors thank one of the reviewers for suggesting a comparison of MARS results with those from K-PLS.

CI034246+