Development of Quantitative Structure−Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein−Ligand Interfaces

Shuxing Zhang, Alexander Golbraikh, and Alexander Tropsha*

The Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7360

Received March 21, 2005

Novel geometrical chemical descriptors have been derived on the basis of the computational geometry of protein−ligand interfaces and Pauling atomic electronegativities (EN). Delaunay tessellation has been applied to a diverse set of 517 X-ray characterized protein−ligand complexes yielding a unique collection of interfacial nearest neighbor atomic quadruplets for each complex. Each quadruplet composition was characterized by a single descriptor calculated as the sum of the EN values for the four participating atom types. We termed these simple descriptors generated from atomic EN values and derived with the Delaunay Tessellation the ENTess descriptors and used them in the variable selection *k*-nearest neighbor quantitative structure−binding affinity relationship (QSBR) studies of 264 diverse protein−ligand complexes with known binding constants. Twenty-four complexes with chemically dissimilar ligands were set aside as an independent validation set, and the remaining dataset of 240 complexes was divided into multiple training and test sets. The best models were characterized by the leave-one-out cross-validated correlation coefficient q^2 as high as 0.66 for the training set and the correlation coefficient R^2 as high as 0.83 for the test set. The high predictive power of these models was confirmed independently by applying them to the validation set of 24 complexes yielding R^2 as high as 0.85. We conclude that QSBR models built with the ENTess descriptors can be instrumental for predicting the binding affinity of receptor−ligand complexes.

Introduction

The prediction of the protein−ligand binding affinity is a critical component of computational drug discovery. Rapid growth of the Protein Data Bank[1] provides opportunities to enhance current protocols for molecular docking and scoring, which are at the core of structure-based drug design[2−5] and hit identification.[6−8] Accurate estimation of binding affinities, or at least correct relative ranking of different ligands, has proven to be a difficult task due to multiple energetic and entropic factors that must be accounted for.[9,10] The limited accuracy of current scoring functions is one of the problems hampering the broad application of docking and virtual screening in lead optimization.

Many scoring functions have been developed over the years. Force field scoring is based on the classical molecular force field (such as AMBER,[11] CHARMM,[12] MMFF94[13]) to compute nonbonded interaction terms between the receptor and ligand atoms. Additional empirical terms taking into account the effects of solvation and entropy have also been considered.[14] The second family of methods includes so-called empirical scoring functions such as LUDI,[15] VALIDATE,[16] and ChemScore.[17] They are based on the concept that the receptor−ligand interaction energy can be approximated by a multivariate regression of different parameters, e.g., the number of hydrogen bonds, lipophilicity, ionic interactions, entropy penalties. Recently, a third family of methods, based on statistical scoring functions (e.g., DrugScore,[18] SMoG,[19,20] PMF,[21] BLEEP,[22] and distance dependent atom pair descriptors[23]), has become popular. These methods employ the statistical analysis of known receptor−ligand complexes to define the pairwise interatomic pseudopotential of protein−ligand interaction. After the calibration on the training set of complexes, these scoring functions are validated by predicting binding affinities for the complexes of the test sets.

Since the force field based scoring functions are too computationally demanding to allow for efficient virtual screening of large databases,[24] their application in screening is usually limited to small datasets. Of the three approaches outlined above, empirical scoring functions are the most computationally efficient and therefore most widely used in current docking programs.

Knowledge-based scoring functions are based on the compositional analysis of protein−ligand complexes. They derive their origin from protein fold recognition studies in the 1970s.[25] Today the growing sources[1,26−28] of structural information on protein−ligand complexes provide great advantages for the continuing development and enhancement of statistical scoring functions. Studies have shown that in many cases knowledge-based scoring functions surpass both force field based and empirical scoring functions in predicting correct binding modes and affinities of the ligands. At the same time, they are fast, accurate, and at least comparable to empirical scoring functions in the efficiency of virtual screening of large databases and combinatorial lead design.[2−4,8,18,20−22,29]

All methodologies discussed above rely on the availability of structural information about protein−ligand complexes and are classified as structure-based drug design approaches. In contrast, ligand-based approaches rely only on the experimental structure−activity relationships for ligands only. Quantitative structure−activity relationship (QSAR)[30] methods are typically used to find correlations between ligands' binding affinities and their chemical descriptors. Some 3D-QSAR methods such as comparative molecular field analysis (CoMFA) have been developed to find correlation between binding affinities and energetic fields surrounding small molecules, such as steric, electrostatic, hydrophobic.[31−33] The "fields" are thought to

* Corresponding author. Telephone: (919) 966-2955. Fax: (919) 966-0204. E-mail: alex_tropsha@unc.edu.

10.1021/jm050260x CCC: $33.50
Published on Web 04/13/2006

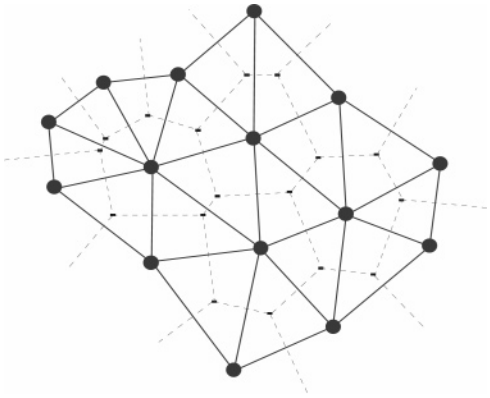

Figure 1. Illustration of Voronoi/Delaunay tessellation in 2D space (Voronoi polyhedra are represented by dashed line and Delaunay simplices by solid line). For the collection of points with 3D coordinates, such as atoms of a protein−ligand complex, Delaunay simplices are tetrahedra whose vertexes correspond to the atoms.

simulate the active site environment but they actually do not consider the receptor geometry or the structural information of the active site (although CoMFA does provide an option to use active site atoms as opposed to a "probe" atom to sample the interaction fields). Several so-called receptor-dependent quantitative structure−activity relationship (RD-QSAR) methods have been developed that rely on the receptor structure information to calculate descriptors.[23,34] Holloway and co-workers[35] have derived a highly significant 3D-QSAR model for HIV-1 protease and its peptidomimetic inhibitors and used it to predict binding affinities for newly designed ligands. Several other authors[16,36,37] have developed new methodologies by considering all of the enthalpic and entropic contributions as well as solvation effects of the receptor−ligand interactions and treated them as independent variables in the RD-QSAR development.

In this paper, we present a hybrid methodology to predict the binding affinities for a highly diverse dataset of protein−ligand complexes using concepts from both structure-based and ligand-based approaches. It is based on a four-body statistical scoring function derived by combined application of the Delaunay tessellation of protein−ligand complexes and the definition of chemical atom types using the fundamental chemical concept of atomic electronegativity. As described in our previous publications,[38−42] Delaunay tessellation naturally partitions a tertiary structure of a protein or a protein−ligand complex into an aggregate of space-filling, irregular tetrahedra, or simplices; the vertices of the simplices are quadruplets of nearest neighbor residues or atoms, respectively (Figure 1). Thus, Delaunay tessellation reduces a complex three-dimensional structure to a collection of explicit, elementary atomic quadruplet structural motifs. Four vertices (atoms) of a simplex form a particular quadruplet composition and the chemical properties of the atom types characterize the type of the tetrahedron.

Atom types can be defined in a number of ways.[16,20−22,43] In general, atoms can be classified into polar and nonpolar carbon atoms, HBA (hydrogen-bond acceptor) and HBD (hydrogen-bond donor), X (halogens), M (metals), cations, anions, and hydrophobic atoms. Herein we present an unconventional way to define atom types using a scale of Pauling electronegativities (EN). To the best of our knowledge, EN has never been used previously to define atom types in a statistical scoring function. We apply atomic EN values to generate descriptors for all frequently observed quadruplet atomic compositions at the interfaces of 517 diverse X-ray characterized protein−ligand complexes. The descriptor value for a specific quadruplet composition in a complex is obtained as a sum of the EN values for the composing atoms. Since these descriptors are based on the constructs from computational geometry (Delaunay tessellation) combined with the fundamental chemical property of composing atom types such as Pauling EN, we term them *geometrical chemical* or *ENTess descriptors*. Herein, we report on the use of the ENTess descriptors as independent variables in multivariate correlation analysis of the experimental dataset of 264 diverse protein−ligand complexes with known binding constants. Following the protocols for developing validated and predictive QSAR models established in the course of our previous studies,[44−47] we have divided these datasets into the training, test, and independent validation sets. We report statistically significant quantitative structure−binding affinity relationships (QSBR) models capable of predicting the binding affinities of ligands in the independent validation set with the R^2 of 0.85.

Materials and Methods

1. Datasets. To develop the ENTess descriptors, we have used two datasets. The first dataset included 517 protein−ligand complexes with high resolution (below 3.0 Å) X-ray crystal structures.[2,4,16,18,20−22,28,48−50] This dataset was used to generate the statistics of quadruplet atomic compositions resulting from Delaunay tessellation of protein−ligand interfaces as discussed below. The second dataset was a subset of the first dataset. It included 264 protein−ligand complexes with known binding affinities (pK_d) ranging between 1.48 (1XLI) and 13.96 (7CPA) log units of molar concentration. The molecular weight of ligands ranged from tens to thousands of daltons. The data were collected from recent publications.[2,4,16,18,20−22,28,48−50] All of the structures in the datasets were prepared for the subsequent analysis as follows: hydrogen atoms and water molecules were removed; ligands were extracted from the protein−ligand complex structures using SYBYL 6.9 and the ligand structures were fixed according to Relibase, which is an online ligand−receptor structure database.[51] We followed the routine that was used by Gohlke and co-workers in their DrugScore development.[18]

2. Structural and Functional Diversity Analysis of the 264 Complexes. To evaluate the structural and functional diversity of this dataset, we have classified the 264 complexes into different families based on their structural and functional annotations using the SWISS-PROT/PDB cross-referencing system.[52] According to this system, each PDB entry is cross-referenced with the SWISS-PROT code, primary gene name (gene expressing that protein), and its source or species of origin. If two proteins have the same primary gene names, they will have very high sequence identity and their structures will be very similar. The family associations of all training set complexes are shown in Table 1. In those cases where no cross-referenced information was available (e.g., PDB entries 1dbb, 1mcf, etc.) the complexes were placed in a group called "MISC".

Based on the SWISS−PROT annotation, the 264 complexes were classified into 71 families reflecting the high functional and structural diversity of this dataset. Some families had multiple members and some had only one member. All of the protein structures within one family were similar, but the ligand structures were different; for different families both protein and ligand structures were dissimilar. We have found that 14 PDB entries were not annotated in the SWISS-PROT/PDB cross-referencing system and they have been classified into the "MISC" family.

3. Atom Type Definitions. To develop simple yet robust chemical geometrical descriptors, we sought some fundamental atomic property that could be attributed to any chemical atom type of either receptor or ligand and could be useful in describing interatomic interactions at the protein−ligand interface. We decided to use the Pauling electronegativity[53] as a parameter to characterize atom types. According to the chemical potential equalization principle as described by Itskowitz and Berkowitz,[54] electronegativity is the first-order term in the energy function of molecules:

Table 1. 264 Protein−Ligand Complexes and the Family Classification Based on Primary Gene Name

family name	no. of complexes	PDB codes of the complexes							family name	no. of complexes	PDB codes of the complexes						
SUBI	1	1sbp							TPIS	5	2ypi	6tim	4tim	7tim	5tim		
ACON	3	8acn	7acn	5acn					FABI	1	2ifb						
6PGD	1	1pgp							CAT3	3	3cla	1cla	4cla				
PHHY	2	1phh	2phh						KAD3	1	2ak3						
F16P	3	1fbc	1fbf	1fbp					HEMA	1	4hmg						
IDH	2	5icd	8icd						RNT1	3	6rnt	1rnt	2rnt				
TRY1	9	1ppc	1pph	3ptb	1tng	1tnh	1tni	1tnj	LDHA	2	1ldm	9ldt					
		1tnk	1tnl						LDHB	1	5ldh						
FKB1	1	1fkf							OPPA	27	1b05	1b0h	1b1h	1b32	1jet	1jeu	1jev
SAV	1	1stp									1b2h	1b40	1b46	1b3f	1b3g	1b3h	1b3l
MDHC	1	4mdh									1b51	1b58	1b4h	1b4z	1b5h	1b5i	1b5j
DAPB	1	1dih									1b6h	1b7h	1b9j	1qka	1qkb	2olb	
RBSB	1	2dri							THRB	4	1etr	1ets	1ett	1tmt			
RBL2	2	1rus	9rub						PRLA	6	8lpr	3lpr	6lpr	9lpr	7lpr	5lpr	
TYSY	2	2tsc	1tlc						MM07	3	1mmp	1mmq	1mmr				
PENP	7	1ppk	1ppl	1ppm	1apt	1apu	1apv	1apw	MM08	3	1mmb	1mnc	1jao				
RENI	1	1rne							PNPH	1	1ulb						
CARP	13	6apr	4er1	4er2	4er4	1eed	2er0	2er6	CATA	1	7cat						
		2er7	2er9	5er2	3er3	1epo	1epp		LYCV	10	181l	182l	1nhb	183l	184l	185l	186l
PYRB	1	8atc									187l	1l83	188l				
XYLA	6	4xia	1xli	2xim	2xis	5xia	8xia		GSHR	1	4gr1						
THER	10	2tmn	5tln	5tmn	3tmn	6tmn	1tlp	1tmn	CATD	1	1lyb						
		4tln	4tmn	7tln					AATM	1	9aat						
AMYG	1	1dog							NRAM	3	1nnb	1nsc	1nsd				
PMG1	1	3pgm							GLNA	1	1lgr						
HISJ	1	1hsl							MYG	1	1mbi						
PLMN	1	2pk4							PRTA	2	4sga	5sga					
ENO1	3	1ebg	5enl	6enl					ARAF	9	1apb	6abp	1abe	1abf	9abp	1bap	7abp
CPXA	4	5cpp	1phf	1phg	2cpp						5abp	8abp					
CAH2	16	1a42	1cil	1cim	1cin	1bn1	1bn3	1bn4	TRY1_TRY2	1	1bra						
		1bnm	1bnn	1bnq	1bnt	1bnu	1bnv	1bnw	RETB	1	1rbp						
		1bcd	1am6						ADHE	2	1adb	1adf					
LDH	1	2ldb							CISY	3	2csc	3csc	1csc				
CBPA	7	2ctc	8cpa	3cpa	6cpa	1cps	1cbx	7cpa	DYR	5	1dhf	4dfr	7dfr	1dr1	1drf		
HV20	1	2mcp							ITHH	3	1dwb	1dwc	1dwd				
NUC	2	1snc	2sns						DGAL	1	2gbp						
TTHY	1	1sta							MALE	1	1mdq						
POL	27	1hih	4hvp	1pro	1dif	2upj	5hvp	1hpv	FLAV	1	3fx2						
		1hpx	8hvp	1hbv	4phv	1sbg	1hsg	1hvk	EL1	4	7est	1ela	1elb	1elc			
		1hvr	1hvs	1hps	9hvp	1hos	1hte	1htf	CONA	1	5cna						
		1htg	1hvi	1hvj	1hvl	1aaq	7hvp		MISC	14	1dbb	1dbj	1dbk	1dbm	2dbl	1mcb	1mcf
RASH	1	5p21									1mch	1mcj	1mcs	1mfe	2cgr	3gap	4fab
SYY	1	4ts1															

$$E(Q_a) = E_0 + \sum_a \mu_a^* Q_a + \frac{1}{2}\sum_a \tilde{\eta}_a Q_a^2 + ... \quad (1)$$

where E is the energy of the molecule, μ_a is the electronegativity of atom a, Q_a is the partial charge on atom a, and $\tilde{\eta}$ is the hardness kernel. E_0 is the collection of terms independent of Q_a, so electronegativity is the main factor determining the atom's polarity and its ability to form a hydrogen bond. For example, oxygen has high electronegativity and high ability to form a hydrogen bond and it is a polar atom type in most cases. Thus, electronegativity could be used to describe the interactions between protein and ligand atoms. Hall et al. have introduced electrotopological state (E-state) indices, which are indirectly related to electronegativity, and successfully used them in QSAR studies of many datasets.[55] Recently Zefirov et al. used an electronegativity equalization scheme as a source of electronic descriptors to study some types of chemical reactivity and obtained good models for thermodynamic and kinetic data such as proton affinity and Taft's inductive σ^* constants.[56]

To collect the most representative statistics for possible ligand atom types, we relied on chemical databases of biologically active organic compounds from the National Cancer Institute (NCI). The first database contains 237 771 compounds[57] and the second one includes 30 000 compounds tested against 60 human cancer cell lines.[58] If an atom type occurred in more than 5000 out of the 237 771 compounds in the first NCI database and in more than 1500 compounds out of 30 000 compounds in the NCI cancer database, we classified it as an independent atom type. For example, O (EN = 3.4), N (EN = 3.0), C (EN = 2.5), and S (EN = 2.4) were classified into independent atom types according to their electronegativity values and their high occurrence in the databases. Although halogens (F, Cl, Br, and I) and P are also important atom types, since each of them occurs less than 5000 times in the NCI database and less than 1500 times in the NCI cancer database, they were classified into the same atom type X [P has a very similar electronegativity value to that of halogens, except for F (between 2.0 and 2.4)]. Similarly, all metal atoms have electronegativity values within 0.6−1.6 and, along with some other rare atom types, were classified into the same atom type M. Atom type definition for proteins is relatively easier, since there are only four atom types, C, N, O, and S, that occur in natural amino acids.

To distinguish ligand vs protein atoms, we have classified the protein and ligand C, N, O, and S as different atom types. Hydrogen atoms were not considered since usually they are not defined explicitly in the X-ray structures. Thus, we have defined four atom types for proteins and six atom types for ligands. In total, there were 554 possible types of interfacial atomic quadruplet compositions, and each of them gave rise to an independent variable (a sum of EN values for composing atom types) for our QSBR studies. Atom type definitions are summarized in Table 2.

4. Delaunay Tessellation of the Protein−Ligand Interfaces. We have developed programs for the protein−ligand complex tessellation based on the *nnsort* method.[59] The protein−ligand interfaces were defined by tetrahedra formed by both protein and ligand atoms. A distance cutoff value of 8 Å was used to exclude Delaunay simplices with long edges (exceeding the physically meaningful interaction distance) between vertices. As shown in Figure 2, we have distinguished three classes of interfacial tetrahedra, i.e., RRRL, RRLL, and RLLL, where each R and L corresponds to a receptor and ligand atom, respectively. For each

Table 2. Atom Type Definitions

ligand atom types		receptor atom types	
O	EN = 3.4	O	EN = 3.4
N	EN = 3.0	N	EN = 3.0
C	EN = 2.5	C	EN = 2.5
S	EN = 2.4	S	EN = 2.4
X	P and halogens, EN = 2.0−2.4, 4.0		
M	metal and all other rare atom types, EN = 0.6−1.6		

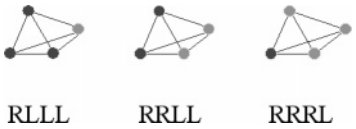

Figure 2. Topological tetrahedral types: RLLL, formed by one receptor atom and three ligand atoms; RRLL, formed by two receptor atoms and two ligand atoms; RRRL, formed by three receptor atoms and one ligand atom.

class we further defined 554 types of quadruplet compositions based on our definition of chemical atom types (cf. Table 2) without taking into account their order in the quadruplet**.** For example, all quadruplets with atom types C_L, C_R, S_L, and X_L were assigned to the same [X_L, S_L, C_L, C_R] composition type.

5. Dataset Division into Training, Test, and Independent Validation Sets. It is generally accepted that the internal validation of the QSAR models built for the training set is sufficient to establish their predictive power.[60−69] However, our previous studies as well as those conducted by other groups have demonstrated that there exists no correlation between leave-one-out (LOO) cross-validated R^2 (q^2) for the training set and the correlation coefficient R^2 between the predicted and observed activities for the test set.[44,70] Our group has advocated the importance of the external model validation, which requires an independent set of compounds.[45,46,71] We have developed a rational approach to dividing the dataset into multiple training and test sets for internal and external validations, respectively.[45,71,72] As described below, we have extended our validation requirements to require not only test sets, but also a second external test set (an independent validation set) for the additional validation.

The dataset of 264 complexes was divided into three subsets in the beginning of the calculations. The first subset of 24 complexes for independent validation was selected randomly. The remaining 240 complexes were divided into multiple chemically diverse training and test sets with the algorithm based on sphere exclusion (SE) developed in our group.[45] SE is a general procedure that is typically applied to databases of organic molecules characterized by multiple descriptors of their chemical structures such that each compound is represented as a point (or vector) in multidimensional descriptor space. The goal of the SE method is to divide a dataset (i.e., a collection of points in multidimensional chemometric space) into two subsets (training and test set) using diversity sampling procedure as follows. SE starts with the calculation of the distance matrix **D** between representative points in the descriptor space. Let D_{min} and D_{max} be the minimum and maximum elements of **D**, respectively. N probe sphere radii are defined by the following formulas. $R_{min} = R_1 = D_{min}$, $R_{max} = R_N = D_{max}/4$, $R_i = R_1 + (i - 1)(R_N - R_1)/(N - 1)$, where $i = 2, ..., N - 1$. Each probe sphere radius corresponds to one division into the training and test set.

In this paper, each protein−ligand complex was characterized with multiple ENTess descriptors as discussed in the first section under Results below. The entire dataset was then treated as a collection of points (each corresponding to an individual protein−ligand complex) in the ENTess descriptor space. Thus, the SE algorithm used in this study consisted of the following steps. (i) Select randomly a point in the ENTess descriptor space. (ii) Include it in the training set. (iii) Construct a probe sphere around this point. (iv) Select points from this sphere and include them alternatively into test and training sets. (v) Exclude all points within this sphere from further consideration. (vi) If no more compounds are left, stop.

Table 3. 24 Randomly Selected Complexes in Three Experiments

experiment 1	experiment 2	experiment 3
188l.pdb	1aaq.pdb	1adf.pdb
1b0h.pdb	1b3l.pdb	1b3f.pdb
1b4h.pdb	1b4z.pdb	1b58.pdb
1b58.pdb	1dbm.pdb	1b5h.pdb
1cim.pdb	1dih.pdb	1cim.pdb
1dbb.pdb	1ebg.pdb	1ebg.pdb
1dbm.pdb	1epo.pdb	1fkf.pdb
1dif.pdb	1hos.pdb	1hte.pdb
1fbc.pdb	1hvj.pdb	1hvl.pdb
1fbf.pdb	1hvr.pdb	1jao.pdb
1hvs.pdb	1mmr.pdb	1phh.pdb
1lgr.pdb	1ppc.pdb	1ppc.pdb
1lyb.pdb	1pph.pdb	1pph.pdb
1mmr.pdb	1qka.pdb	1qka.pdb
1nnb.pdb	1qkb.pdb	1stp.pdb
1nsc.pdb	1rne.pdb	1tmn.pdb
1phg.pdb	1rus.pdb	1tnh.pdb
1tlc.pdb	1sbg.pdb	1tnk.pdb
1tnh.pdb	1stp.pdb	2dri.pdb
2upj.pdb	3fx2.pdb	2sns.pdb
2xim.pdb	3lpr.pdb	3cpa.pdb
5ldh.pdb	4dfr.pdb	4tln.pdb
7dfr.pdb	7abp.pdb	4tmn.pdb
9abp.pdb	7tln.pdb	5ldh.pdb

Otherwise, let m be the number of probe spheres constructed and n be the number of remaining points. Let d_{ij} ($i = 1, ..., m$; $j = 1, ..., n$) be the distances between the remaining points and probe sphere centers. Select a point corresponding to the lowest d_{ij} value and go to step (ii). The random division was repeated three times and the results are summarized in Table 3. The training sets were used to build models and the test sets were used for validation. The independent validation sets of 24 complexes were used for an additional external validation.

6. *k*-Nearest Neighbor (*k*NN) QSBR with Variable Selection. We have described this approach elsewhere[73,74] and present here only its brief overview. *k*NN QSAR is a stochastic variable selection procedure where the model optimization is driven by simulated annealing, as is illustrated in Figure 3 The *k*NN procedure is aimed at the development of the model with the highest leave-one-out (LOO) cross-validated correlation coefficient R^2 (q^2) for the training set.

$$q^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (2)$$

where N and $\bar{y}$ are the number of compounds and the average observed activity of the training set, respectively, and y_i and $\hat{y}_i$ are the observed and predicted activities of the ith compound, respectively.

The procedure starts with the random selection of a predefined number of descriptors from all descriptors. The activity of a compound y_i excluded in the LOO cross-validation procedure is predicted as the weighted average of activities of its nearest neighbors according to the following formula:

$$\hat{y}_i = \frac{\sum_{j=1}^{k} y_j \exp(-d_{ij}/\sum_{l=1}^{k} d_{il})}{\sum_{j=1}^{k} \exp(-d_{ij}/\sum_{l=1}^{k} d_{il})} \quad (3)$$

where d_{ij} are distances between the ith compound and its k nearest neighbors ($j = 1, ..., k$). The optimal number of nearest neighbors that yields the highest q^2 value is defined as part of the LOO cross-

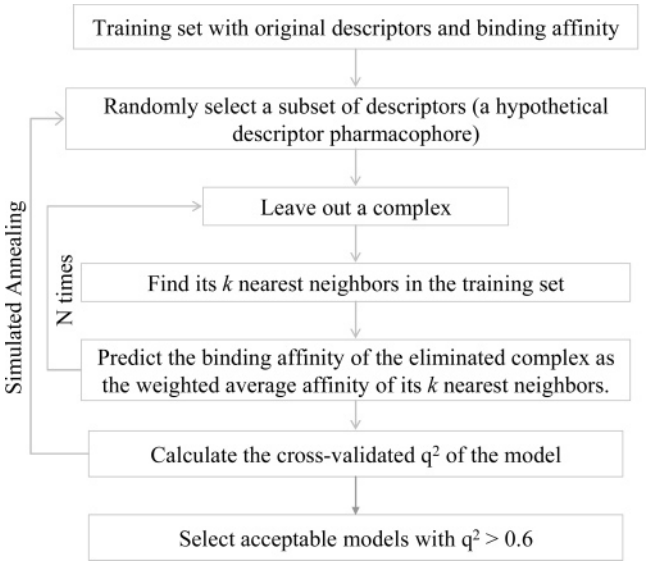

Figure 3. Flowchart of *k*NN-QSAR with variable selection.

validation process as well. After each run of the LOO procedure, a predefined number of descriptors are randomly changed, and the new value of q^2 is defined. If q^2(new) > q^2(old), the new set of descriptors is accepted. If q^2(new) ≤ q^2(old), the new set of descriptors is accepted with probability $p = \exp[q^2(\text{new}) - q^2(\text{old})]/T$ and rejected with probability $(1 - p)$, where T is a simulated annealing "temperature" parameter. During the process, T is decreasing until a predefined value, and when this value is achieved, the optimization process is terminated.

7. *Y*-Randomization Test. The robustness of the models was examined by comparing them to models obtained when using randomized binding affinities of the training set (this procedure is commonly referred to as the *Y*-randomization test). Briefly, we repeated the QSAR calculations with the randomized activities of the training sets. We also compared the q^2 values in the process of the iteration procedure of the simulated annealing for actual and random activities of training sets to see if there is any significant difference. This randomization was repeated five times for each splitting.

8. Model Validation and the Applicability Domain. QSBR models were validated using test sets. They were considered as acceptable if (i) $q^2 > 0.5$ and $R^2 > 0.6$, (ii) $[R^2 - R_0^2]/R^2 < 0.1$ and $0.85 < k < 1.15$ or $[R^2 - R'_0{}^2]/R^2 < 0.1$ and $0.85 < k' < 1.15$, and (iii) $|R_0^2 - R'_0{}^2| < 0.3$,[45] where R_0^2 and $R'_0{}^2$ are the coefficients of determination for regressions through the origin between predicted vs observed, and observed vs predicted binding affinities, respectively, and k and k' are the corresponding slopes. The whole QSBR model validation procedure, as is illustrated in Figure 4, has been successfully used in our laboratory for many datasets and is described in detail elsewhere.[73−75]

The binding affinities of the test set compounds were predicted only if these compounds were within the *applicability domain* of the respective training set models. We define this domain[45] as a threshold distance in multidimensional descriptor space between a test set compound and its k nearest neighbors in the training set. If the distance is beyond the threshold, the prediction is considered unreliable. This threshold distance is calculated as $D^2_{\text{cutoff}} = \langle D^2_{\text{nn}} \rangle + Z \times \text{VAR}$, where $\langle D^2_{\text{nn}} \rangle$ is the squared mean distance between each of the training set compound and its k nearest neighbors, VAR is the variance of D_{nn}, and Z is a user-defined parameter (the default value is 0.5).

Training set models that passed our validation criteria (i)−(iii) were used for the prediction of the independent validation set of randomly selected compounds. For this exercise, we relied on the consensus prediction, which consists of averaging the binding affinities of each compound predicted by all acceptable models.[37]

9. QSBR Model Validation Using Computational Docking Studies. The goal of this component of our studies was to query the QSBR models with respect to their ability to differentiate between native bound conformations of the ligands and their decoys. In addition, we have also questioned whether QSBR models could discriminate known binders from those molecules that are known *not* to bind to the receptors, which is a rigorous test for any docking method. We have randomly selected three complexes from the PDB. They were human dihydrofolate reductase complexed with folate (1DHF),[76] orotidine 5′-phosphate decarboxylase complexed to 6-hydroxyuridine 5′-phosphate (BMP) (1DQX),[77] and human P38 Map kinase in complex with BIRB796 (1KV2).[78] The 1DHF docking study was done with FlexX[79] implemented in SYBYL,[80] while 1DQX and 1KV2 poses were created using Autodock 3.0.[81] In addition, aribinose was docked into dihydrofolate reductase using FlexX[79] and the enzyme coordinates from 1DHF to create an unnatural complex, since it is known that aribinose does not bind to dihydrofolate reductase. We have employed the default docking parameters unless otherwise specified. The ligands were considered flexible, and 50 conformations were docked and scored for each ligand.

Results and Discussion

1. Atom Type Definition and ENTess Descriptor Generation. The nearest neighbor interacting atoms at the protein−ligand interface were defined by the means of Delaunay tessellation as described in Methods. The examples of interfacial tetrahedra are shown in Figure 5 for the complex between HIV protease and acetylpepstatin (PDB code 5HVP). Tetrahedra with edges (i.e., interatomic distances) exceeding 8 Å were excluded. We have applied this procedure to 517 protein−ligand complexes in the training set as described in Methods and counted the number of occurrences of each of the 554 atom quadruplet types. If the number of times a particular type occurred was higher than 50, we considered this quadruplet type significant. Otherwise, this type was discarded, leading to the reduction in the number of descriptors for the subsequent analysis. 132 types of quadruplets were found to occur with sufficiently high frequency (Figure 6). For each type of the tetrahedral composition, the EN values of the four composing atoms were added up, and the resulting sums for all of the tetrahedra belonging to this composition type were then added up again. The result of these calculations represented the value of the descriptor (i.e., one of possible 132 descriptors) for the particular protein−ligand complex (see Figure 7 for the illustration).

All 132 descriptors were initially calculated for the dataset of 264 complexes with known binding constants. We found that 32 out of 132 descriptors had zero values for all 264 complexes, so they were excluded from further consideration. The final descriptor matrix included 264 rows for the protein−ligand complexes and 100 columns for descriptors. We have applied variable selection *k*-nearest neighbor (*k*NN)[74] to this matrix to build models and establish correlations between binding affinities and the ENTess descriptors as described below.

2. Building QSBR Models. To build validated QSBR models, we have divided the dataset of 264 receptor−ligand complexes with known binding constants into training, test, and validation subsets multiple times. Three different subsets of the entire dataset were generated initially by removing 24 randomly selected complexes that constituted the independent validation sets. In each case the remaining subset of 240 compounds was divided into multiple training and test sets using the SE program as described in Methods. For every division, the models were built with the number of descriptors ranging between 10 and 60 with the increment of 5. Five models for each number of descriptors were built. As a result, 55 training set models were generated and then validated by predicting the binding constants of the test sets. Due to the stochastic nature of the corresponding

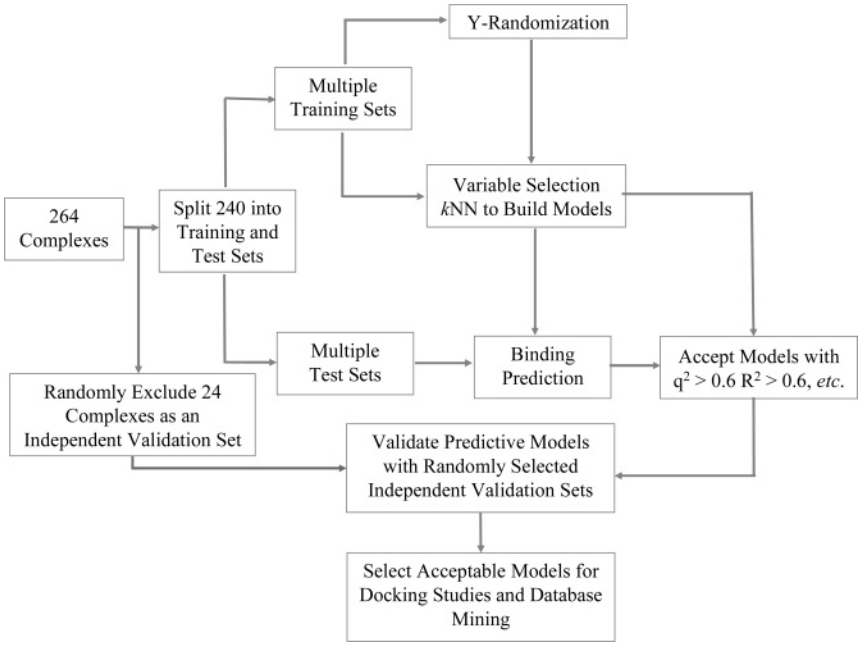

Figure 4. Statistical data modeling and model validation workflow using the *k*NN variable selection approach.

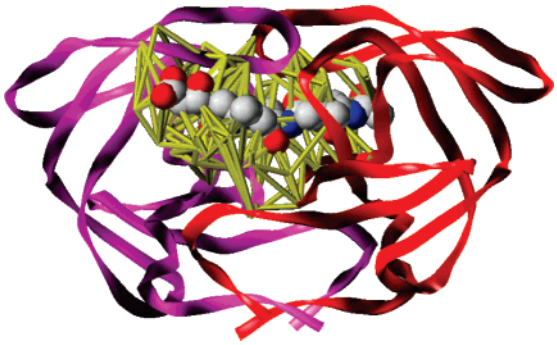

Figure 5. Full atom-based protein−ligand interface tessellation for 5HVP. The magenta and red ribbons are two chains of the protein. The acetylpepstatin ligand is in the spacefill display. Tetrahedra formed by ligand and protein atoms are shown in yellow.

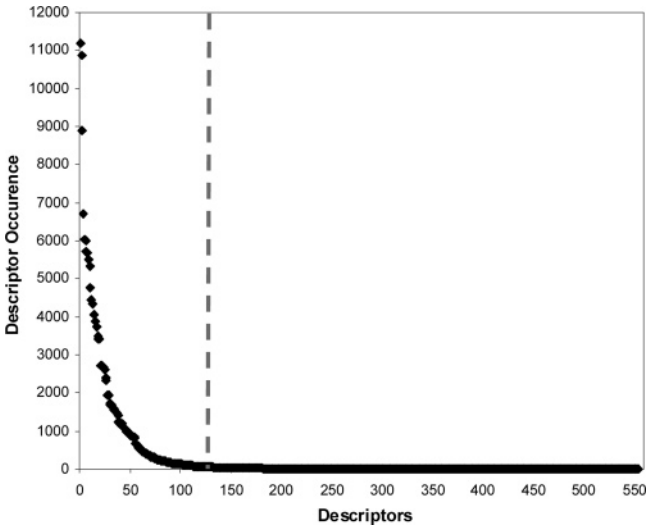

Figure 6. Frequency analysis of 554 composition types for the 517 protein−ligand complex dataset. All of the quadruplets on the left of the dashed line were found more than 50 times.

SE algorithm, the number of divisions was different for different chemically diverse samples selected from the original dataset. In the end, as many as 1155 models for 21 divisions of the first sample of 240 complexes, 1045 models for 19 divisions of the second sample, and 2310 models for 42 divisions of the third sample were built and validated using variable selection *k*NN.

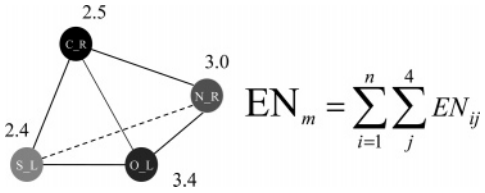

Figure 7. Calculation of the ENTess descriptors. The same atom type from receptor and ligand is treated differently. In the formulas, *m* is the *m*th quadruplet composition type; *n* represents the number of occurrences of this composition type in a given protein−ligand complex, and *j* is the vertex index within the quadruplet.

Application of the acceptability criteria discussed in the Methods section resulted in 354, 515, and 567 models for the three samples described above with $q^2 > 0.50$ and $R^2 > 0.60$. To evaluate the statistically significant predictive power of the training set models, our test sets typically included no less than 15% of the dataset. As could be expected, due to the high diversity of the dataset, the q^2 and R^2 were found to depend on the division of the dataset. For example, we were unable to obtain acceptable training set models for the 173/67 (training/test set complexes) division but were able to generate highly predictive models for the 167/73 division, where the best model had R^2 as high as 0.71 (cf. model 28 in Table 4).

These results could be explained as follows. As a result of the division, some complexes that are potential outliers are included in the test set, which reduces the R^2. On the contrary, if these structures are included in the training set, the test set R^2 could be much higher than the training set q^2. With the criteria described above, an acceptable model was obtained with the test set as large as 118 complexes, i.e., almost half of the entire dataset, with $q^2 = 0.53$ and $R^2 = 0.63$ (cf. model 30 Table 4).

3. Prediction of the Independent Validation Sets. It should be noted that the studies described above rely on the test sets to select the acceptable training set models. So, strictly speaking, the above procedure cannot be regarded as truly external validation. On the contrary, successful prediction of the

Table 4. Best 10 Models for Each of the Three Dataset Divisions[a]

models	k	q^2	n	R^2	S^2	slope	R_0^2	R_{45}^2	R_1^2	R_{comb}^2	R_{cons}^2
Experiment 1											
1	3	0.54	48	0.77	1.04	0.93	0.77	0.76	0.71	0.82	0.85
2	4	0.55	48	0.76	1.13	0.92	0.76	0.75	0.67	0.75	
3	4	0.52	48	0.76	1.05	0.93	0.76	0.75	0.76	0.79	
4	2	0.57	41	0.76	1.46	0.99	0.75	0.75	0.68	0.78	
5	3	0.65	47	0.74	1.27	0.93	0.73	0.72	0.76	0.77	
6	3	0.61	44	0.74	1.25	1.01	0.74	0.73	0.69	0.73	
7	3	0.56	65	0.73	1.44	0.98	0.73	0.73	0.74	0.84	
8	3	0.59	53	0.70	1.24	0.97	0.70	0.70	0.73	0.82	
9	2	0.54	65	0.70	1.56	1.00	0.70	0.69	0.74	0.81	
10	3	0.60	44	0.70	1.44	0.98	0.70	0.70	0.66	0.77	
Experiment 2											
11	3	0.65	40	0.83	0.89	0.95	0.83	0.83	0.77	0.74	0.77
12	3	0.66	40	0.83	0.92	0.97	0.83	0.83	0.57	0.55	
13	3	0.66	41	0.82	0.99	0.95	0.82	0.82	0.68	0.72	
14	3	0.63	47	0.81	0.92	0.97	0.80	0.80	0.6	0.64	
15	2	0.58	51	0.80	0.96	1.02	0.80	0.78	0.64	0.72	
16	3	0.63	51	0.83	0.82	1.05	0.82	0.78	0.61	0.62	
17	3	0.60	47	0.80	0.95	0.98	0.79	0.79	0.72	0.77	
18	3	0.63	47	0.80	0.95	0.97	0.79	0.79	0.58	0.64	
19	3	0.57	44	0.76	1.19	0.98	0.76	0.76	0.8	0.83	
20	2	0.64	50	0.78	0.93	1.00	0.77	0.76	0.62	0.77	
Experiment 3											
21	3	0.55	49	0.78	0.99	0.97	0.78	0.78	0.77	0.81	0.81
22	2	0.52	49	0.77	1.20	0.97	0.76	0.76	0.73	0.74	
23	3	0.52	49	0.75	1.15	0.98	0.75	0.75	0.61	0.74	
24	5	0.51	49	0.75	0.90	0.94	0.72	0.72	0.65	0.69	
25	5	0.52	49	0.74	0.99	0.98	0.72	0.72	0.63	0.65	
26	4	0.52	49	0.74	1.08	0.99	0.73	0.72	0.78	0.83	
27	3	0.55	45	0.70	1.14	0.94	0.70	0.70	0.8	0.83	
28	4	0.53	73	0.71	1.24	0.91	0.68	0.67	0.65	0.84	
29	3	0.55	73	0.68	1.44	0.92	0.68	0.67	0.72	0.74	
30	2	0.53	118	0.63	1.69	0.91	0.57	0.54	0.73	0.74	

[a] *k*, number of the nearest neighbors; q^2, cross validated correlation coefficient for training sets; *n*, number of complexes in the test sets which are within the applicability domain; R^2, correlation coefficient for test sets; S^2, square of standard deviation between predicted and actual pK_i; slope, slope of the regression through the origin; R_0^2, correlation coefficient for test sets for the regression through the origin; R_{45}^2, correlation coefficient for test sets for the line which has slope 45°; R_1^2, correlation coefficient for the external set; R_{comb}^2, correlation coefficient for the external set by using the combination of training and test sets for predictions; R_{cons}^2, correlation coefficient for the external set by consensus prediction with top 10 best models.

randomly selected independent validation set of 24 compounds could be viewed as a realistic test of the models' predictive power. We now discuss the results of this test under different prediction scenarios.

3.1. Prediction with the Best Individual Models. Table 4 presents the 10 best models for each experiment. Model 11 tops the list with R^2 as high as 0.83 and q^2 of 0.65. Figure 8 shows the data fitting of experimental and predicted binding affinities for training and test sets. This model was built with 45 descriptors resulting from variable selection procedures, and three nearest neighbors appeared to be optimal in the leave-one-out (LOO) cross-validation.

Figure 9 shows the trajectory of the SA-driven optimization of the q^2 in developing the best *k*NN models, and Figure 10 shows the relationship between the number of the descriptors and the q^2 for the training set with real vs randomized binding affinities. The latter figure demonstrates that the models built using true binding affinities for the training set afford significantly higher q^2 values as compared to the models generated with the randomized binding energies.

To further validate the models, we made predictions for the independent validation set of 24 randomly selected complexes in three independent experiments (Table 3). For each individual model, we have obtained fairly good correlation between the actual and predicted binding affinity (Table 4), with the exception of models 12 and 18, where R^2 fell below 0.60; all other models had R^2 ranging from 0.60 to 0.80.

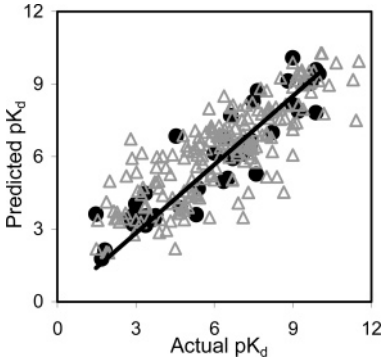

Figure 8. Predictive power of the best model (model 11, cf. Table 4): gray open triangles, prediction for the 200 complexes of the training set ($q^2 = 0.65$); black points, prediction for the 40 complexes of the test set ($R^2 = 0.83$, rmsd = 1.06).

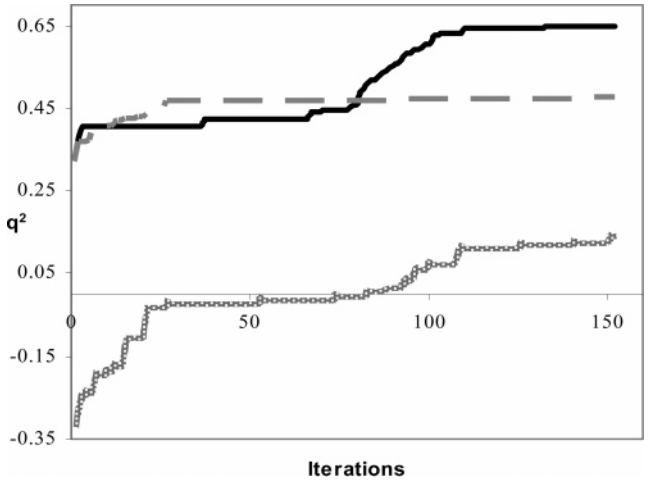

Figure 9. Trajectories for q^2 of the best model (model 11) (solid black) and the model with the lowest q^2 (dashed gray). Trajectory of the model with the highest q^2 (shadowed gray) built with randomized binding energies of the training set.

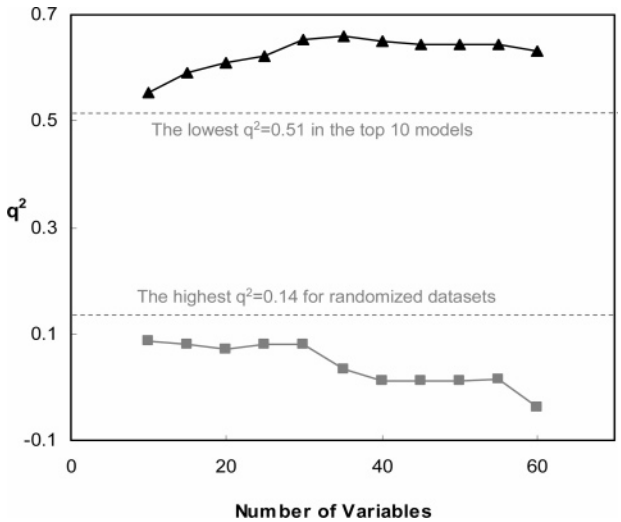

Figure 10. q^2 vs the number of variables selected for the *k*NN QSAR models. The results are for both actual (black) and random (gray) datasets. Every q^2 is the average of 10 independent calculations.

3.2. Predictions Using the Combined Training and Test Sets. All predictions described in the previous section were made using training sets only. Since the dataset of 240 complexes

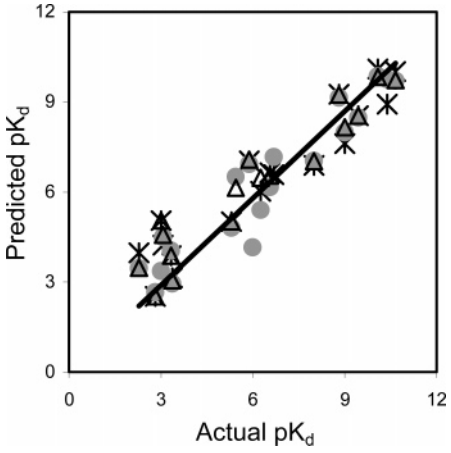

Figure 11. Prediction of binding affinities for the external validation test set (24 complexes) with different approaches (cf. Table 4): asterisks, prediction with model 7 (Table 4), $R^2 = 0.74$ and rmsd = 0.97; black open triangles, prediction with model 7 using the whole dataset of 240 complexes to select *k* nearest neighbors for compounds in the independent test set, $R^2 = 0.84$ and rmsd = 0.90; gray points, consensus prediction by the top 10 best models using the whole dataset of 240 complexes as the training set, $R^2 = 0.85$ and rmsd = 0.98.

was divided into the training and test sets rationally and the test set predictions were used to select acceptable models, it is logical to employ the (re)combined set for the prediction of the independent validation set. Thus, all 240 compounds of the recombined dataset were used for the binding affinity prediction of the independent validation set. We used the descriptors selected and the optimal number of nearest neighbors obtained by the *k*NN training set modeling. Perez et al.[37] have reported previously that using a similar approach improves the prediction accuracy. Following this approach, we made predictions for 24 complexes with the 10 best models for each experiment (cf. Table 3), and the results were significantly better than using only the training set compounds. In addition to R^2, root mean squared deviation (rmsd) between predicted and observed binding is also used to measure the accuracy of the prediction. It is defined in the literature as[15,48]

$$\text{rmsd} = \sqrt{\frac{(\text{p}K_d^{\text{pred}} - \text{p}K_d^{\text{obs}})^2}{N-1}} \qquad (4)$$

where $\text{p}K_d^{\text{pred}}$ and $\text{p}K_d^{\text{obs}}$ are the predicted and observed logarithmic binding affinity, respectively. N is the number of complexes. Gibbs free energy of binding ΔG is related to the binding constant by

$$\Delta G = -RT \ln K_d \qquad (5)$$

For instance, for the predictions made with model 7, R^2 increased from 0.74 to 0.84 and rmsd decreased from 0.97 (5.5 kJ/mol) to 0.90 (5.1 kJ/mol) (cf. Table 4 and Figure 11). Since we only use training set models that have both internal and external high predictive power, every compound in the combined set has nearest neighbors in the selected descriptor space with approximately the same binding affinity. Obviously, combining the training and test sets enriches the structural diversity of the dataset used for prediction such that there is a greater chance for every external compound of finding close nearest neighbors. Furthermore, because we are using the applicability domain threshold, the nearest neighbor relationships translate into similar binding affinities, leading to high values of the external R^2.

Table 5. Comparison of Predictive Power of ENTess Models vs that Obtained with Alternative Scoring Functions

methods	ref	training set size	test set size	R^2 for test sets	consensus R^2 for the external set
BLEEP	22	351	90	0.53	N/A
PMF	21	697	77	0.61	N/A
SMoG96	19	120	46	0.42	N/A
SMoG2001	20	725	111	0.436	N/A
DT2002	*a*	319	67	0.71	N/A
SCORE	49	170	11	0.65	N/A
XSCORE	50	200	30	0.36	N/A
LUDI	15	82	12	0.45	N/A
VALIDATE	16	51	14	0.81	N/A
ChemScore	17	82	20	0.63	N/A
ENTess1		189−200	40−51	0.76−0.83	0.77
ENTess2		199−175	41−65	0.70−0.77	0.85
ENTess3		122−195	45−118	0.63−0.78	0.81

[a] Feng, J., unpublished data.

3.3. Prediction with the Consensus Method. With the consensus approach, the binding affinities for each of the 24 complexes in the independent validation set were predicted as the average of the predicted binding affinities for each complex based on individual models. The results, as shown in Table 4, demonstrate that the consensus prediction is relatively stable with R^2 of 0.85, 0.77, and 0.81, respectively. Figure 11 shows, that the consensus approach predicts more data with higher correlation coefficient than any single model. Notably, as shown in Table 5, model 12 has good q^2 (0.66) and very high R^2 (0.83), but the R^2 for the prediction of the 24 external complexes is below 0.60. This indicates that, even if both q^2 and R^2 are very high, it does not guarantee that the external predictive power of an individual model is acceptable. On the contrary, the consensus prediction usually yields acceptable predictive power. This result is consistent with our previous observations.[44]

4. Analysis of Outliers. For each complex, if the difference between the predicted and experimental binding affinities was greater than three logarithmic units (i.e., $\text{p}K_\text{d}$), we regarded the complex as an outlier. On the basis of this definition, we have observed several outliers in different experiments: 1STP[82] in experiment 1, 1PHG[83] in experiment 2, and 1STP and 7TLN[84] in experiment 3. 1STP is a very interesting complex that was observed as an outlier by several groups working in the area of scoring function development.[17,21,79] The 1STP complex is unique and our predicted affinity with different models underestimated the observed binding affinity by 4−7 $\text{p}K_\text{d}$ units. The biotin−streptavidin complex has the highest known binding constant[82] and it is the only member of the SAV family (Table 1). Consequently, there are no analogues of this complex in the training set. More importantly, Muegge and Martin[21] pointed out that streptavidin functions as tetramer; we only have monomeric complex crystal structures available, whereas the interaction with a second subunit increases the binding of biotin by eight orders of magnitude.

1PHG[83] was predicted to have binding affinity ca. three $\text{p}K_\text{d}$ units lower than the experimental value (for instance, model 7 predicts the $\text{p}K_\text{d}$ value for this complex as 5.52, while the observed binding affinity is 8.66). It is cytochrome P450_cam (camphor 5-monoxygenase) complexed with metyrapone, and it contains the heme group as cofactor. The crystal structure indicates that there is some interaction between the ligand and the heme group that is not taken into account by our scoring function.

7TLN[84] is a metalloproteinase covalently bound to its ligand INC ($CH_2CO(N-OH)$Leu-OCH_3). In addition, there are four Ca^{2+} and one Zn^{2+} ions in the complex. In this case, the

Table 6. Binding Affinity Prediction and the Ranking of Docked Poses Based on Their Predicted pK_d[a]

docking poses	predicted pK_d by ENTess	rmsd (Å)	ranking	docking poses	predicted pK_d by ENTess	rmsd (Å)	ranking
1dqx.pdb	10.694	0	native	abp_1dhf_1.pdb	2.687	0	lowest energy
1dqx_2.pdb	7.696	2.06	1	abp_1dhf_22.pdb	2.687	0.70	1
1dqx_6.pdb	7.685	1.93	2	abp_1dhf_13.pdb	2.686	1.56	2
1dqx_1.pdb	4.813	3.32	3	abp_1dhf_41.pdb	2.685	4.37	3
1dqx_47.pdb	3.786	6.17	4	abp_1dhf_3.pdb	2.668	6.28	4
1dhf.pdb	7.760	0	native	1kv2.pdb	8.702	0	native
1dhf_1.pdb	6.678	1.64	1	1kv2_5.pdb	5.698	1.53	1
1dhf_4.pdb	5.246	1.12	2	1kv2_21.pdb	4.863	1.21	2
1dhf_49.pdb	4.111	2.18	3	1kv2_46.pdb	3.741	7.61	3
1dhf_31.pdb	4.110	2.97	4	1kv2_34.pdb	3.702	4.55	4
1dhf_26.pdb	3.839	7.78	5	1kv2_13.pdb	3.699	2.73	5
1dhf_8.pdb	3.637	6.31	6	1kv2_40.pdb	3.613	6.04	6

[a] The numbers after the pdb codes are the rankings in the original docking methods.

concurrent binding of these ions could affect the prediction of the binding affinity, as was observed with 1LYB.[85] There are too few metal-containing complexes in our training dataset, and our approach may not accurately describe interactions mediated by metal ions.

In addition to the outliers, several complexes were found to be out of the applicability domain in our experiments. This means they are too different from their respective training set complexes in the 100-descriptor space. As described above, most of them have metal ions that may induce large conformational changes upon ligand binding. For example, 1EBG[86] and 4TMN[87] are metal complexes with four magnesium ions and four calcium ions, respectively. Although we have descriptors for quadruplets that contain metal atoms, the representation of the interaction interface is probably insufficient to characterize their metal-mediated large conformational change upon ligand binding. In addition, ligands in these two complexes contain PO_3 and PO_2 groups, respectively, which are not frequent in the entire dataset. Another example is 1FKF,[88] which is an immunophilin−immunosuppressant complex in which the protein conformation changes insignificantly upon binding ascomocin (FK506), but interestingly, the ligand FK506 undergoes a very large conformational change when it binds. FK506 is an antibiotic with a very large molecular weight (804 Da). The drug's association with the protein involves five hydrogen bonds; the protein hydrophobic binding pocket is lined with conserved aromatic residues and contains an unusual carbonyl binding pocket.[88] We suppose that the training set model is incapable of describing these unique interactions accurately. However, despite the small number of outliers, we suggest that the ENTess descriptors as applied in *k*NN QSBR calculations in general led to highly predictive models.

5. Robustness of the Models. As described in Methods, to evaluate the model robustness, we have performed the Y-randomization test. As shown in Figures 9 and 10, q^2 values for models built with real activities of the training set were always much higher than for those built with randomized activities. To exclude the possibility of chance correlations and overfitting, the *Y*-randomization test was repeated five times for each splitting. The highest q^2 for the random datasets was 0.14, while the lowest q^2 for the real datasets was 0.51. In general, if the relationships between binding affinities and descriptors are not random, the models built with randomized affinities of the training sets complexes must have no predictive ability. Indeed, no predictive model built with randomized training set data was found.

6. Comparison with Other Scoring Functions. Our results were compared with those obtained earlier using both knowledge-based and empirical scoring functions, as shown in Table 5. Since there are no standard training and test sets used by different groups, the direct comparison is impossible. Compared to SMoG96,[19] our training sets were a little bigger, but our prediction accuracy was much better, even for a much bigger test set (118 complexes). As compared to other published results, we had test sets of comparable size and much smaller training sets, but nevertheless, our correlation coefficients are much higher. Importantly, we have demonstrated that our method afforded high predictive power for an external structurally diverse dataset. The alternative empirical scoring functions demonstrated comparable results with relatively smaller training sets (except SCORE and XSCORE[48,49]), but the test sets are also small, which highly influences the value of R^2. In summary, our models were rigorously validated using test sets, using the additional external prediction set of 24 compounds to simulate the real application of the models, and by performing *Y*-randomization tests. The results demonstrate the high prediction power of our models and the applicability of our novel geometrical chemical descriptors to binding affinity prediction.

7. Validation Using Docking Studies. For each docking case, the resulting poses were grouped into different bins based on their rmsd against the crystal structure (for 1DQX, 1DHF, and 1KV2) or the lowest energy binding conformation (for the unnatural aribinose−DHFR complex); the bin width was 0.5 Å. The poses with rmsd above 8 Å were not considered. This process led to six nonempty bins for both 1DHF (actual pK_d = 7.4)[76] and 1KV2 (actual pK_d = 10.0)[78] and four nonempty bins for both 1DQX (actual pK_d = 11.05)[89] and the DHFR−aribinose unnatural complex. The poses with the lowest estimated binding free energy were selected as representatives of each bin. Thus, we have obtained six poses for 1DHF and 1KV2 and four poses for 1DQX and DHFR−aribinose complexes.

The pK_d resulting from consensus prediction using the best 30 ENTess models were used to rank the aforementioned poses, and the results are shown in Table 6. These results demonstrate that, in all cases, ENTess predictions could clearly differentiate the native crystallographic bound conformation from the other decoy poses. For instance, our results for 1DHF are consistent with FlexX[79] for the top-ranked poses: ENTess top 1 and 2 were ranked 1 and 4 by FlexX[79] with 1.64 and 1.12 Å rmsd, respectively. Both of them actually belong to the same binding conformation and orientation mode. All of the poses ranked low by FlexX were also ranked low by ENTess. The low binding affinity (ca. 1 mM) predicted by ENTess corresponded to poses with weak binding to the DHFR receptor. Similarly, ENTess estimations were accurate for 1DQX and 1KV2: on the basis of ENTess predictions, all ligand conformations with low rmsd

Table 7. Occurrence of 100 Tetrahedra Types in the Best 30 QSBR Models

descriptor types	occurrence	descriptor types	occurrence	descriptor types	occurrence
CL-CL-CL-NR	27	CL-NR-NR-OR	16	CL-NL-OL-NR	12
CL-OR-OR-OR	24	CL-CL-CL-OR	15	CL-OL-OL-NR	12
CL-CL-NL-NR	22	CL-CL-OL-OR	15	CL-CL-OR-OR	12
CL-NL-OL-OR	22	CL-OL-OL-CR	15	OL-OL-NR-NR	12
CL-CL-NR-NR	22	CL-OL-OL-OR	15	CL-SR-CR-CR	12
CL-NL-CR-CR	22	NL-NL-OL-CR	15	NL-NR-OR-OR	12
OL-OL-CR-OR	22	OL-OL-OL-NR	15	XL-OL-OL-OR	11
OL-OL-OR-OR	22	CL-NL-CR-NR	15	CL-CL-NL-OR	11
NL-SR-CR-OR	22	CL-OL-CR-CR	15	NL-OL-OL-NR	11
NL-NL-CR-CR	21	NL-NL-NR-OR	15	OL-OL-OL-CR	11
XL-CR-CR-CR	21	NL-OL-CR-OR	15	SL-OL-CR-NR	11
CL-NL-NL-NR	20	OL-OL-CR-CR	15	CL-CL-SR-CR	11
XL-CR-CR-OR	20	OL-OL-NR-OR	15	CL-CL-CR-OR	11
CL-SR-CR-OR	20	CL-CR-NR-OR	15	CL-NL-NR-OR	11
OL-OL-OL-OR	19	XL-OL-OL-NR	14	CL-OL-OR-OR	11
CL-OL-NR-NR	18	NL-OL-OL-OR	14	NL-OL-OR-OR	11
NL-NL-OR-OR	18	SL-CL-CR-NR	14	CL-CR-CR-OR	11
NL-OL-CR-CR	18	CL-CL-CR-NR	14	CL-CL-OL-NR	10
XL-CR-NR-OR	18	CL-OL-SR-CR	14	SL-OL-CR-CR	10
SL-CR-CR-OR	18	CL-CR-CR-NR	14	CL-OL-CR-NR	10
CL-CL-NR-OR	17	CL-NR-OR-OR	14	CL-OL-CR-OR	10
CL-NL-CR-OR	17	NL-CR-NR-OR	14	NL-CR-OR-OR	10
CL-NL-OR-OR	17	NL-NR-NR-OR	14	CL-CL-CL-SR	9
SL-CR-CR-NR	17	XL-OL-OL-CR	13	CL-NL-OL-CR	9
CL-SR-CR-NR	17	CL-NL-NL-CR	13	NL-CR-NR-NR	9
NL-CR-CR-CR	17	CL-NL-NL-OR	13	CL-CR-CR-CR	8
NL-CR-CR-NR	17	CL-NL-NR-NR	13	NL-CR-CR-OR	8
SL-CL-CL-CR	16	CL-OL-NR-OR	13	CL-CL-CL-CR	7
SL-CL-OL-CR	16	NL-OL-NR-OR	13	CL-CL-OL-CR	7
NL-OL-OL-CR	16	OL-OL-CR-NR	13	SL-CL-CR-OR	7
CL-CL-CR-CR	16	SL-CR-CR-CR	13	SL-CL-CR-CR	6
NL-NL-CR-OR	16	CL-CR-NR-NR	13	NL-ML-CR-NR	6
NL-OL-CR-NR	16	CL-CR-OR-OR	13		
XL-CR-CR-NR	16	CL-CL-NL-CR	12		

are strong binders, while the low-ranked poses are decoys. Most interestingly, aribinose was successfully docked into the DHFR binding pocket using FlexX,[79] while we knew that the binding did not happen at all. Probably this is the problem of many if not all existing docking programs. In contrast, ENTess suggests that all of the docked poses have very low binding affinity (lower than 1 mM). This observation suggests that binding affinity estimates using ENTess for poses generated with available docking programs can be used to eliminate false positives.

8. Chemical Properties of Descriptors Implicated in Significant QSBR Models. QSBR models generated with the variable selection *k*NN method can be characterized not only by their statistical characteristics but also analyzed in terms of ENTess descriptors that best models are built with. To this end, we have calculated the frequency of occurrence of those selected descriptors found in the 30 best models used for the prediction of external test sets. Table 7 shows the most frequently occurring descriptor types. They demonstrate that frequent quadruplet compositions of atom types include purely hydrophobic (such as four carbon atom tetrahedra), hydrophilic (such as four oxygens or nitrogens or mixed polar atom type quadruplet compositions), as well as tetrahedra with mixed polar and nonpolar atom composition (e.g., including two carbon and two oxygen or nitrogen atoms). These results indicate that variable selection kNN models tend to rely on chemically diverse descriptor types that capture major intermolecular binding interactions such as hydrophobic effect and hydrogen bonds.

9. The Importance of Electronegativity for ENTess Descriptors. ENTess descriptors are very simple; since their values are approximately proportional to the number of quadruplets with certain compositions, it may appear that significant models could be generated without taking into account the electronegativity values at all. To address the importance of EN, we have repeated all calculations described above but using only the numbers of occurrence of different tetrahedra as descriptors. Interestingly, the statistical parameters for training and test set models were comparable with those using the ENTess descriptors, with q^2 ranging from 0.5 to 0.7 and R^2 from 0.6 to 0.8 (data not shown). However, the predictions of the external validation set with these models were much less accurate than using the ENTess descriptors (the consensus prediction R^2 values were always below 0.5). Furthermore, the acceptable training set models, on average, constituted only about 15% of all of the models built, which is far fewer than the 40% obtained when using the ENTess descriptors.

In a separate experiment, we used atomic weights as the property to generate descriptors in place of EN. Similarly, the q^2 and R^2 for training/test set models, respectively, were comparable with those generated with the ENTess descriptors. However, although the prediction of the external validation set gave better results than using the occurrence numbers, the models were not as robust and stable as those built using EN values (the best R^2 value for consensus prediction was 0.63 for only one of the three external validation sets and much lower for the other two validation sets; data not shown). We reason that using electronegativity to calculate the ENTess descriptors affords better models, since EN implicitly incorporates major atomic properties that are important in intermolecular interactions such as polarity, energy, and ability to form hydrogen bond. Including other atomic parameters certainly could further improve our method as we continue its development. In the future studies, we plan to combine charges with EN to derive more sophisticated and perhaps more robust descriptors. Nevertheless, we believe that the simplicity of the approach proposed in this paper and our demonstrated ability to generate reliable QSBR models using ENTess descriptors makes these descriptors attractive for a wide range of QSBR studies.

Conclusions

To the best of our knowledge, our studies represent the first attempt to use electronegativity as a main parameter for the definition of atom types and descriptors for protein−ligand binding affinity prediction based on a QSBR approach. To develop structure-based scoring function, we have combined the atomic EN with the geometrical description of the protein−ligand interface using Delaunay tessellation. Delaunay tessellation is a unique way to represent the geometrical complementarity between receptors and ligands. Electronegativity has been found to define important terms in the molecular energy functions. On the basis of these two concepts, we have developed novel geometrical chemical descriptors. The descriptors have been applied in QSBR studies of binding energies for a dataset of 264 protein−ligand complexes. QSBR models were built with the variable selection *k*-nearest neighbors (*k*NN) algorithm based on simulated annealing.

Using the ENTess descriptors, we have built and validated the QSBR models for protein−ligand binding affinity prediction. Robust and accurate binding affinity predictions with R^2 up to 0.83 for the test sets and 0.85 for the independent validation set have been obtained (Table 4). Compared to the conventional atom type definitions,[16,20−22,43] our method is very simple yet uses fundamental chemical and geometrical principles. Our current analysis relies only on 10 atom types in total and a relatively small number of descriptors, which can be considered as an additional advantage of this methodology. Comparison

with other scoring functions has demonstrated that our approach is accurate and efficient for the prediction of binding affinities for diverse protein−ligand structures. Our QSBR models can be used to predict binding free energy for protein−ligand complexes resulting from experimental studies or docking calculations. We expect that as additional data become available,[90] the accuracy and the range of applicability of our statistical scoring function will increase.

Acknowledgment. Special thanks are due to Dr. M. Karthikeyan for providing the statistics for different atom types in chemical databases and Dr. P. Itskowitz for providing the docking poses from AutoDock and valuable discussions concerning the use of electronegativity in deriving the ENTess descriptors. We also thank Drs. J. Feng, B. Krishnamoorthy, and S.Q. Zong for their help with programming and Mr. R. Shah for discussions concerning the protein family classification. The studies presented in this paper were supported by the NIH research grant GM066940.

References

(1) 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.
(2) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein−ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.
(3) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409−443.
(4) Tame, J. R. Scoring functions: A view from the bench. *J. Comput. Aided Mol. Des.* **1999**, *13*, 99−108.
(5) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **2002**, *16*, 151−166.
(6) Bohm, H. J.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **2000**, *43*, 2664−2674.
(7) Gruneberg, S.; Wendt, B.; Klebe, G. Subnanomolar inhibitors from computer screening: A model study using human carbonic anhydrase II. *Angew. Chem., Int. Ed.* **2001**, *40*, 389−393.
(8) Grzybowski, B. A.; Ishchenko, A. V.; Shimada, J.; Shakhnovich, E. I. From knowledge-based potentials to combinatorial lead design in silico. *Acc. Chem. Res.* **2002**, *35*, 261−269.
(9) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand−receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(10) Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001**, *3*, 231−250.
(11) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force-field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5187.
(12) MacKerell, A. D., Jr.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257−265.
(13) Halgren, T. A. Merck molecular force field: 1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(14) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*, 4−16.
(15) Bohm, H. J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aided Mol. Des.* **1998**, *12*, 309−323.
(16) 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.
(17) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425−445.
(18) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(19) 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.
(20) 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.
(21) 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.
(22) 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.
(23) Deng, W.; Breneman, C.; Embrechts, M. J. Predicting protein−ligand binding affinities using novel geometrical descriptors and machine-learning methods 1. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 699−703.
(24) Kollman, P. A. Free energy calculations: Application to chemical and biochemical phenomenon. *Chem. Rev.* **1993**, *93*, 2395−2417.
(25) Tanaka, S.; Scheraga, H. A. Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins. *Macromolecules* **1976**, *9*, 142−159.
(26) Bader, G. D.; Betel, D.; Hogue, C. W. BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **2003**, *31*, 248−250.
(27) Zhang, S.; Ying, W. S.; Siahaan, T. J.; Jois, S. D. S. Solution structure of a peptide derived from the beta subunit of LFA-1. *Peptides* **2003**, *24*, 827−835.
(28) Roche, O.; Kiyama, R.; Brooks, C. L., III Ligand-protein database: Linking protein−ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592−3598.
(29) Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **1999**, *42*, 2498−2503.
(30) Martin, Y. C. Q*uantiative Drug Design: A Critical Introduction;* Marcel Decker Inc.: New York, Basel, 1978; pp 1−425.
(31) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(32) Kulkarni, S. S.; Gediya, L. K.; Kulkarni, V. M. Three-dimensional quantitative structure activity relationships (3-D-QSAR) of antihyperglycemic agents. *Bioorg. Med. Chem.* **1999**, *7*, 1475−1485.
(33) Kulkarni, S. S.; Kulkarni, V. M. Three-dimensional quantitative structure−activity relationship of interleukin 1-beta converting enzyme inhibitors: A comparative molecular field analysis study. *J. Med. Chem.* **1999**, *42*, 373−380.
(34) Tokarski, J. S.; Hopfinger, A. J. Prediction of ligand−receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: Application to a set of peptidometic renin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792−811.
(35) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; deSolms, S. J. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* **1995**, *38*, 305−317.
(36) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38*, 2681−2691.
(37) Perez, C.; Pastor, M.; Ortiz, A. R.; Gago, F. Comparative binding energy analysis of HIV-1 protease inhibitors: Incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Med. Chem.* **1998**, *41*, 836−852.
(38) Carter, C. W., Jr.; LeFebvre, B. C.; Cammer, S. A.; Tropsha, A.; Edgell, M. H. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.* **2001**, *311*, 625−638.
(39) Sherman, D. B.; Zhang, S.; Pitner, J. B.; Tropsha, A. Evaluation of the relative stability of liganded versus ligand-free protein conformations using simplicial neighborhood analysis of protein packing (SNAPP) method. *Proteins* **2004**, *56*, 828−838.
(40) Zhang, S.; Kaplan, A. H.; Tropsha, A. Unpublished observations.
(41) Singh, R. K.; Tropsha, A.; Vaisman, I. I. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. *J. Comput. Biol.* **1996**, *3*, 213−221.
(42) Tropsha, A.; Singh, R. K.; Vaisman, I. I.; Zheng, W. Statistical geometry analysis of proteins: Implications for inverted structure prediction. *Pac. Symp. Biocomput.* **1996**, 614−623.

(43) Bush, B. L.; Sheridan, R. P. PATTY: A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.
(44) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(45) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* **2003**, *17*, 241−253.
(46) Tropsha, A.; Gramatica, P.; Gomba, V. K. The improtance of being earnest: Validation is the absolute essential for the successful application and interpretaion of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(47) Tropsha, A. Recent Trends in Quantitative Structure−Activity Relationships. In B*urger's Medicinal Chemistry and Drug Discovery;* Abraham, D., Ed.; John Wiley & Sons: New York, 2003; pp 49−77.
(48) Wang, R. X.; Liu, L.; Lai, L. H.; Tang, Y. Q. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−394.
(49) Wang, R. X.; Lai, L. H.; Wang, S. M. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.* **2002**, *16*, 11−26.
(50) Wang, R. X.; Lu, Y. P.; Wang, S. M. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(51) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein−ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607−620.
(52) http://www.imb-jena.de/ImgLibPDB/pages/SWP/index.php. **2005**.
(53) Pauling, L. The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. *J. Am. Chem. Soc.* **1932**, *54*, 3570−3582.
(54) Itskowitz, P.; Berkowitz, M. L. Chemical potential equalization principle: Direct approach from density functional theory. *J. Phys. Chem. A* **1997**, *101*, 5687−5691.
(55) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: Applications to 3D QSAR. *J. Comput. Aided Mol. Des.* **1996**, *10*, 513−520.
(56) Oliferenko, A. A.; Krylenko, P. V.; Palyulin, V. A.; Zefirov, N. S. A new scheme for electronegativity equalization as a source of electronic descriptors: Application to chemical reactivity. *SAR QSAR Environ. Res.* **2002**, *13*, 297−305.
(57) http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html, 2005.
(58) http://dtp.nci.nih.gov/docs/cancer/cancer_data.html, 1999.
(59) Watson, D. F. Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *The Computer J.* **1981**, *24*, 167−172.
(60) Basak, S. C.; Mills, D. Prediction of mutagenicity utilizing a hierarchical QSAR approach. *SAR QSAR Environ. Res.* **2001**, *12*, 481−496.
(61) Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. Quantitative structure−activity relationships of mutagenic and carcinogenic aromatic amines. *Chem. Rev.* **2000**, *100*, 3697−3714.
(62) Cronin, M. T.; Dearden, J. C.; Duffy, J. C.; Edwards, R.; Manga, N.; Worth, A. P.; Worgan, A. D. The importance of hydrophobicity and electrophilicity descriptors in mechanistically based QSARs for toxicological endpoints. *SAR QSAR Environ. Res.* **2002**, *13*, 167−176.
(63) Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. Quantitative structure−antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies. *J. Med. Chem.* **2001**, *44*, 3254−3263.
(64) Girones, X.; Gallegos, A.; Carbo-Dorca, R. Modeling antimalarial activity: Application of kinetic energy density quantum similarity measures as descriptors in QSAR. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1400−1407.
(65) Moss, G. P.; Dearden, J. C.; Patel, H.; Cronin, M. T. Quantitative structure-permeability relationships (QSPRs) for percutaneous absorption. *Toxicol. In Vitro.* **2002**, *16*, 299−317.
(66) Randic, M.; Basak, S. C. Construction of high-quality structure−property-activity regressions: The boiling points of sulfides. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 899−905.
(67) Suzuki, T.; Ide, K.; Ishida, M.; Shapiro, S. Classification of environmental estrogens by physicochemical properties using principal component analysis and hierarchical cluster analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 718−726.
(68) Trohalaki, S.; Gifford, E.; Pachter, R. Improved QSARs for predictive toxicology of halogenated hydrocarbons. *Comput. Chem.* **2000**, *24*, 421−427.
(69) Wang, X.; Yin, C.; Wang, L. Structure−activity relationships and response−surface analysis of nitroaromatics toxicity to the yeast (*Saccharomyces cerevisiae*). *Chemosphere.* **2002**, *46*, 1045−1051.
(70) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-dimensional quantitative similarity−activity relationships (3D QSiAR) from SEAL similarity matrixes. *J. Med. Chem.* **1998**, *41*, 2553−2564.
(71) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput. Aided Mol. Des.* **2002**, *16*, 357−369.
(72) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure−activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **2002**, *45*, 2811−2823.
(73) Hoffman, B.; Cho, S. J.; Zheng, W. F.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure−activity relationship modeling of dopamine D-1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.
(74) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(75) Golbraikh, A.; Bonchev, D.; Tropsha, A. Novel ZE-isomerism descriptors derived from molecular topology and their application to QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 769−787.
(76) Davies, J. F.; Delcamp, T. J.; Prendergast, N. J.; Ashford, V. A.; Freisheim, J. H.; Kraut, J. Crystal-Structures of Recombinant Human Dihydrofolate-Reductase Complexed with Folate and 5-Deazafolate. *Biochemistry* **1990**, *29*, 9467−9479.
(77) Miller, B. G.; Hassell, A. M.; Wolfenden, R.; Milburn, M. V.; Short, S. A. Anatomy of a proficient enzyme: The structure of orotidine 5′-monophosphate decarboxylase in the presence and absence of a potential transition state analog. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2011−2016.
(78) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268−272.
(79) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(80) SYBYL, Version 6.9; Tripos, Inc., St. Louis, MO, 2002.
(81) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195−202.
(82) Weber, P. C.; Ohlendorf, D. H.; Wendoloski, J. J.; Salemme, F. R. Structural origins of high-affinity biotin binding to streptavidin. *Science* **1989**, *243*, 85−88.
(83) Poulos, T. L.; Howard, A. J. Crystal structures of metyrapone- and phenylimidazole-inhibited complexes of cytochrome P-450cam. *Biochemisty* **1987**, *26*, 8165−8174.
(84) Holmes, M. A.; Tronrud, D. E.; Matthews, B. W. Structural analysis of the inhibition of thermolysin by an active-site-directed irreversible inhibitor. *Biochemisty* **1983**, *22*, 236−240.
(85) Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Hosur, M. V.; Sowder, R. C.; Cachau, R. E.; Collins, J.; Silva, A. M.; Erickson, J. W. Crystal structures of native and inhibited forms of human cathepsin D: Implications for lysosomal targeting and drug design. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6796−6800.
(86) Wedekind, J. E.; Poyner, R. R.; Reed, G. H.; Rayment, I. Chelation of serine 39 to Mg2+ latches a gate at the active site of enolase: Structure of the bis(Mg2+) complex of yeast enolase and the intermediate analog phosphonoacetohydroxamate at 2.1-Å resolution. *Biochemisty* **1994**, *33*, 9333−9342.
(87) Holden, H. M.; Tronrud, D. E.; Monzingo, A. F.; Weaver, L. H.; Matthews, B. W. Slow- and fast-binding inhibitors of thermolysin display different modes of binding: Crystallographic analysis of extended phosphonamidate transition-state analogues. *Biochemisty* **1987**, *26*, 8542−8553.
(88) Van Duyne, G. D.; Standaert, R. F.; Karplus, P. A.; Schreiber, S. L.; Clardy, J. Atomic structure of FKBP−FK506, an immunophilin−immunosuppressant complex. *Science* **1991**, *252*, 839−842.
(89) Miller, B. G.; Hassell, A. M.; Wolfenden, R.; Milburn, M. V.; Short, S. A. Anatomy of a proficient enzyme: The structure of orotidine 5′-monophosphate decarboxylase in the presence and absence of a potential transition state analog. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2011−2016.
(90) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M. The PDBbind database: Collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

JM050260X