Hierarchical PLS Modeling for Predicting the Binding of a Comprehensive Set of Structurally Diverse Protein−Ligand Complexes

Anton Lindström,[‡] Fredrik Pettersson,[‡] Fredrik Almqvist, Anders Berglund,[†] Jan Kihlberg, and Anna Linusson*

Organic Chemistry, Department of Chemistry, Umeå University, SE-901 87, Umeå, Sweden

Received August 12, 2005

A new approach is presented for predicting ligand binding to proteins using hierarchical partial-least-squares regression to latent structures (Hi-PLS). Models were based on information from the 2002 release of the PDBbind database containing (after in-house refinement) high-resolution X-ray crystallography and binding affinity (K_d or K_i) data for 612 protein−ligand complexes. The complexes were characterized by four different descriptor blocks: three-dimensional (3D) structural descriptors of the proteins, protein−ligand interactions according to the Validate scoring function, binding site surface areas, and ligand 2D and 3D descriptors. These descriptor blocks were used in Hi-PLS models, generated using both linear and nonlinear terms, to relate the characterizations to p$K_{d/i}$. The results show that each of the four descriptor blocks contributed to the model, and the predictions of p$K_{d/i}$ of the internal test set gave a root-mean-square error of prediction (RMSEP) of 1.65. The data were further divided according to the structural classification of the proteins, and Hi-PLS models were constructed for the resulting subclasses. The models for the four subclasses differed considerably in terms of both their ability to predict p$K_{d/i}$ (with RMSEPs ranging from 0.8 to 1.56) and the descriptor block that had the strongest influence. The models were validated with an external test set of 174 complexes from the 2003 release of the PDBbind database. The overall results show that the presented Hi-PLS methodology could facilitate the difficult task of predicting binding affinity.

INTRODUCTION

In the past decade, the increasing availability of high-quality structural data on macromolecular targets and advances in computational tools have made structure-based design an important component of modern drug discovery. Structures of proteins cocrystallized with ligands are publicly available from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB)[1] and are an important source of data for studies of the interactions between proteins and ligands. The binding affinity is a measure of how strongly a ligand binds to its target binding site and is generally assumed to correlate to the effect that the compound will have on related biological functions. The relationship between molecular structure and affinity is a fundamental concept in modern drug design. The strength of binding can be experimentally determined using techniques such as microcalorimetry,[2,3] ELISA assays,[4] NMR spectrometry,[5] and surface plasmon resonance.[6] However, to perform such analyses, potential ligands need to be synthesized, which is often labor-intensive, time-consuming, and expensive. Therefore, computational methods are often used to predict the binding strength of druglike compounds and thus increase the probability of identifying new ligands that bind to a target of interest. Techniques that can be used in the lead identification process in the search for novel ligands include docking analysis[7,8] and fragment-based de novo design.[9] The success of these methods is largely determined by the accuracy of the identification of the binding modes of the ligands and subsequent estimates of the strength of the protein−ligand interactions. These parameters are usually estimated by applying a scoring function. The use of reliable mathematical models for binding affinity predictions can significantly facilitate the lead discovery process. However, binding affinity is a very complex variable, and the scoring problem, as described by Gohlke and Klebe, highlights the difficulties involved in constructing functions that generate correct solutions.[10] The performance of common scoring functions has been shown to be generally poor by Marsden et al.[11] and Wang et al.,[12] illustrating the difficulties associated with creating a good scoring function, especially a global scoring function for several different types of targets.

There are several classes of scoring functions, the most common classes being free-energy and linear-free-energy perturbation calculations, force field, and knowledge- and regression-based methods.[10] Several regression techniques, based on diverse or homogeneous training sets of complexes (also known as empirical regression methods), have been developed to provide estimates that should theoretically correlate with binding affinity, for examples, SCORE1 and SCORE2 developed by Böhm,[13,14] ChemScore by Eldrige et al.,[15] SCORE by Wang et al.,[16,17] DrugScore by Gohlke et al.,[18] and PLD by Puvanendrampillai et al.[19] In the hybrid empirical/force-field-based scoring function Validate, estimates of the free energy of binding were initially made by calculating physiochemical descriptors of the ligands and protein−ligand complexes.[20] In an attempt to more accurately describe the electrostatic contribution to protein−ligand

* Corresponding author e-mail: anna.linusson@chem.umu.se.

[†] Current address: Computational Biology, Pfizer Global Research and Development, Chesterfield, MO 63017, U. S. A.

[‡] These authors contributed equally to this paper.

10.1021/ci050323k CCC: $33.50
Published on Web 04/08/2006

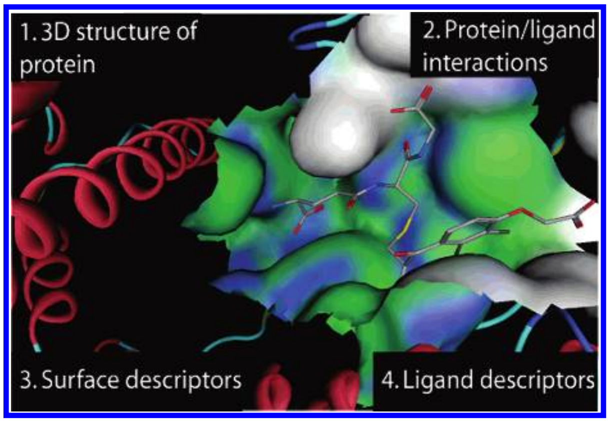


Figure 1. Schematic overview of protein−ligand interactions in the binding site, showing the four different kinds of characterizations used for modeling binding affinity.

binding, 15 additional descriptors were subsequently included to account for hydrogen bonds and highest occupied molecular orbital energy alongside the original Validate descriptors.[21] The expanded Validate II model was applied to 363 ligands to estimate their strength of binding to HIV protease. One major proof of concept that the development of Validate II showed was that scoring functions could be developed based on crystal structures with defined binding modes and used to predict the affinity of ligands that had unknown binding modes and, hence, needed to be modeled.[21] An additional example of molecular descriptors that have been used to model the interactions between ligands and proteins are the VolSurf descriptors used by Zamora et al. to further quantify hydrogen-bond interactions, alongside those previously included in Validate.[22]

The cited scoring functions have been based, at most, on 170 protein−ligand complexes. Using a larger training set for the regression model could provide a way to increase the knowledge exploited by the scoring function.[23] The recently compiled PDBbind database, including data on 1359 diverse complexes with known binding parameters (IC_{50} and K_d or K_i values) and 3D structures,[24] provides an excellent starting point for creating a training set to generate robust predictive regression models.

Here, we present an attempt to predict the strength of protein−ligand interactions using four types of characterization data that could hold important information related to the specificity and strength of binding (Figure 1). These variable blocks were 3D structural description of proteins, protein−ligand interaction terms, surface area descriptors, and physicochemical descriptors of ligands. The first descriptor block was developed to investigate whether characterization of the 3D structure of the proteins, using metrics capturing interatomic distances and secondary structure elements, could help to model protein−ligand interactions. The second block corresponded to the descriptors used in the original Validate scoring function[20] on the basis of our implementation. The third block was an attempt to describe physicochemical properties of the binding site. Here, we used PATTY (Programmable ATom TYper) classifications of atom types[25] to calculate a set of binding site descriptors. The final set of descriptors corresponded to traditional 2D and 3D molecular descriptors of ligands that are commonly used in quantitative structure−activity relationship (QSAR) studies to model biological activity.

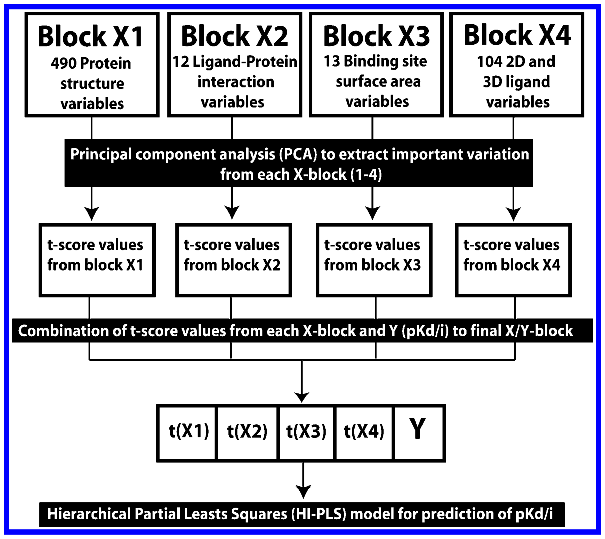

Block X1
490 Protein structure variables
Block X2
12 Ligand-Protein interaction variables
Block X3
13 Binding site surface area variables
Block X4
104 2D and 3D ligand variables
Principal component analysis (PCA) to extract important variation from each X-block (1-4)
t-score values from block X1
t-score values from block X2
t-score values from block X3
t-score values from block X4
Combination of t-score values from each X-block and Y (pKd/i) to final X/Y-block
t(X1)
t(X2)
t(X3)
t(X4)
Y
Hierarchical Partial Leasts Squares (HI-PLS) model for prediction of pKd/i

Figure 2. Schematic overview of a hierarchical modeling procedure.

Hierarchical partial-least-squares regression to latent structures (Hi-PLS) was used to relate the characterization of the protein−ligand complexes by the four different blocks to their $pK_{d/i}$ values (Figure 2). This regression technique offers a good way of investigating the importance and influence of different blocks of data, through the compression of each individual block by principal component analysis (PCA) before the regression.[26] This projection method condenses the data set into its main properties and reduces the noise levels before modeling. The Hi-PLS method has been previously applied to analyze 897 G-protein-coupled receptors.[27] In the more commonly applied regression methodology PLS,[28] the data are directly correlated to a response without the use of PCA and, thus, are less easy to interpret, especially if large numbers of variables are used in the model.[26]

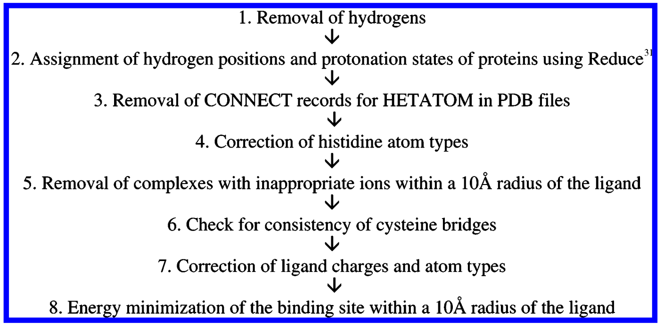
In this study, Hi-PLS regression was used to estimate $pK_{d/i}$ values of a refined set of protein−ligand complexes from the PDBbind database. The complexes were divided into their respective structural classification of proteins (SCOP) classes[29,30] to see whether this division could yield improved models. The resulting models were compared with models derived using a standard PLS regression technique and were subsequently evaluated using internal and external test sets.

MATERIALS AND METHODS

The PDBbind Database. The PDBbind database is the largest collection of data on protein−ligand complexes, with information on both binding parameters and known 3D structures.[24] The 2002 release of the PDBbind database included data on 1359 protein−ligand complexes with K_d, K_i, and IC_{50} values and was based on PDB release number 103. The PDBbind database is also available as a further refined database containing data on 800 protein−ligand complexes, filtered to include proteins that bind only one known ligand with druglike properties, excluding (for instance) compounds with a molecular weight higher than 1000 and both carbohydrates and nucleic acids. Complexes with cofactors (e.g., NAD or heme) and those for which the X-ray structure has not been determined at a resolution of at least 2.5 Å are also excluded.

Experimental determination of the binding data is essential for the modeling procedure and subsequent predictions. The

Scheme 1. Stepwise Procedure for Cleaning Up the Protein−Ligand Complexes Used in this Work



1. Removal of hydrogens
↓
2. Assignment of hydrogen positions and protonation states of proteins using Reduce[31]
↓
3. Removal of CONNECT records for HETATOM in PDB files
↓
4. Correction of histidine atom types
↓
5. Removal of complexes with inappropriate ions within a 10Å radius of the ligand
↓
6. Check for consistency of cysteine bridges
↓
7. Correction of ligand charges and atom types
↓
8. Energy minimization of the binding site within a 10Å radius of the ligand

modeling in this paper was based on data compiled from many different sources and laboratories that often supplied no details of quality control procedures. Therefore, one should bear in mind that there may be significant errors in the measured values, and the scale of the errors is likely to vary between the different methods used. To minimize the variability, it was decided to exclude complexes for which the only reported binding data were IC_{50} values and, hence, perform the modeling on the remaining set, for which K_d and K_i values were available. K_d and K_i values are equilibrium constants and, thus, depend solely on the temperature, while IC_{50} values are highly dependent on other assay conditions, for example, the amount of ligand available to the receptor.[23] The K_d and K_i values in the PDBbind database have generally been recorded at room temperature and neutral pH.[24] In our modeling process, we tested the effects on the performance of the models basing them solely on either the K_d or K_i values. The resulting models did not show any improvement compared to models based on the total set (data not shown), so complexes for which K_d or K_i values were available were included in the modeling procedure.

Refinement of the PDBbind Database. In this study, the refined PDBbind database, consisting of 800 complexes, was further filtered and reduced to a final database of 612 complexes (Scheme 1) with $pK_{d/i}$ binding values ranging from 0.6 to 13.96. The proteins and ligands were checked for inaccuracies using both manual and automated methods. Water was included using the representation presented in the PDBbind database as an input throughout the pretreatment of the protein−ligand complexes. Complexes comprising ligands with a molecular weight larger than 700 were removed.

Protein Pretreatment. All hydrogens in the proteins were first removed from the PDB file, as recorded in the PDBbind database, and then added again using the software Reduce,[31] after which the side chains ASN, GLN, and HIS were flipped and the OH and SH groups were rotated. Thus, protonation states were set according to the program, based on the likelihood that the residues would be charged at physiological pH together with the surrounding microenvironments' opportunity to form hydrogen bonds. It should be noted that this procedure is by no means exhaustive or complete, and the assignment of protonation states will be wrong in some cases. Nevertheless, we decided to perform this procedure for the 612 complexes used in this initial modeling study, knowing that the errors in protonation state of the binding site could be a source of inaccuracy in the modeling performance. The next refinement step was to remove the CONNECT records for metals from the PDB text files to avoid including potentially incorrect or irrelevant covalent bonds between, for example, Zn^{2+} ions and the proteins in the modeling. The surrounding microenvironment for the particular heteroatom was also subsequently corrected by text editing, for example, making sure that the sp^2 hybridized nitrogen atom of histidines were facing the Zn^{2+} ion. Calculation of the descriptors involved energy minimization of the proteins up to a distance of 10 Å from the ligands, so it was vital to evaluate parts of the protein within this range. Complexes were searched for metals and counterions within 10 Å of the ligand and subsequently excluded from the database if the following species were identified: SO_4^{2}, PO_4^{3-}, Cd^{2+}, and Hg^{+}. The original references were studied in cases where the ions Ca^{2+}, Cl^{-}, Cu^{2+}, Fe^{2+}, Fe^{3+}, K^{+}, Mg^{2+}, Mn^{2+}, Na^{+}, and Zn^{2+} were identified, and complexes were included only if it could be determined that the ions were present during the binding experiment.

The ideal length of the S−S bond of a cysteine bridge is considered to be 2.1 Å.[32] The PDB files of the PDBbind database were searched for SSBONDS records with a distance greater than 2.1 Å. Ten such structures were found, and on the basis of information in the original references, one complex was removed (1M21[33]). For the remaining nine structures, the S−S bond was manually inserted.

Ligand Pretreatment. Protonation states of the ligands originated from the PDBbind database, which provided data on the ligands in the form of mol2 files. The ligand files were screened for missing atom types, and charges were calculated using MMFF94s as implemented in the Molecular Operating Environment (MOE)[34] and corrected using Support Vector Language (SVL) programming in MOE.[34] A subset of ligands was visually examined to confirm that the representation was correct. Special attention was paid to the types of nitrogen atoms known to cause modeling problems.

Energy Minimization of the Binding Site. The protein−ligand complexes were energy minimized using molecular mechanics (MM) in the 10 Å core surrounding their respective ligands. MM calculations (in vacuo) were performed in MOE[34] using a sequential stepwise procedure of 100 steepest descent (SD) iterations, 100 conjugate gradient iterations (CG), and 100 truncated Newton (TN) iterations while keeping chiral constraints. For MM calculations, the MMFF94s force field was used.

Division of the Data Set into Subgroups. In the SCOP database,[29,30] proteins are classified according to their 3D structure. Proteins are annotated in a hierarchy through the levels *Class*, *Fold*, *Superfamily*, *Family*, and *Species* down to the structural domains of individual PDB entries. In this work, we investigated the possibility that better models could be obtained if the data set consisting of the 612 complexes was divided according to the top-level classification of SCOP. This level consists of the following seven classes: (A) proteins whose secondary structure consists mainly of α helices (all α), (B) proteins whose secondary structure consists mainly of *β* sheets (all *β*), (C) proteins with interspersed α helices and *β* sheets (α/*β*), (D) proteins with segregated α helices and *β* sheets (α + *β*), (E) multidomain proteins, (F) membrane and cell surface proteins, and (G) small proteins. Of these, four classes (A−D) had sufficient representatives among the 612 complexes to enable subset modeling, and the classification was used to divide the refined PDBbind database into subgroups.

Variable Blocks. The protein−ligand complexes in the filtered and refined database were characterized by four blocks of variables, describing the 3D structure of the protein, the interaction between the ligand and protein, surface areas of the protein binding site, and physicochemical properties of the ligand (Figure 1). Water was included, using the output from the refinement of the database as an input, including the energy minimization throughout the calculations of the descriptors, if not explicitly stated otherwise.

Block 1: Multivariate Characterization of Protein Structure. New descriptors were developed for describing the 3D structure of the proteins. The proteins were characterized in a multivariate way with two sub-blocks of descriptors based on (*i*) C-α coordinates and (*ii*) ϕ/ψ angles. Autocovariance (AC) and autocross-covariance (ACC) transformations[35] were applied to the C-α descriptor block and the ϕ/ψ descriptor block, respectively, to deal with protein sequences of varied length and alignment problems.[36−38] In AC and ACC, the relationship between pairs of amino acids at specific distances (lags) is established and averaged over the sequence, thus generating the same number of variables for each protein, regardless of its original sequence length. The lag lengths need to be smaller than the minimum size of the shortest protein but large enough to capture relevant information. Here, the selected lag lengths were 1−49 amino acids, on the basis of the loadings from the PCA of the protein structure variables, which indicated that these lags describe most of the structural variation of the proteins (data not shown).

The C-α distance variables were generated by applying the AC transformation to calculated Euclidean distances between every pair of C-α atoms separated by a specific number of amino acids in the sequence, as given by eq 1, where *x*, *y*, and *z* are the coordinates for the C-α atoms, lag is the number of amino acids spanning the distance, and *n* is the total number of amino acids in the sequence. This calculation was repeated for lags 1−49, resulting in 49 variables, one for each lag.

$$\text{Euclid(lag)} = \frac{\sum_{i=1}^{n-\text{lag}} \sqrt{(x_i - x_{i+\text{lag}})^2 + (y_i - y_{i+\text{lag}})^2 + (z_i - z_{i+\text{lag}})^2}}{n - \text{lag}} \qquad (1)$$

For the second set of descriptors, ϕ/ψ angles were calculated for every residue. The ϕ and ψ angles refer to rotations of the two rigid peptide units around the same C-α atom, and most combinations produce steric collisions. Each of the secondary structures (β sheet, left-handed α helix, and right-handed α helix) can be characterized by different allowed combinations of ϕ/ψ angles, as pioneered by Ramachandran and Sasisekhran.[39] Every residue was characterized as being part of one or none of these secondary structures on the basis of the allowed combination of angles (Table 1) and was given a binary description (2−4% of the residues, depending on the scope classification, were not assignable). For example, if a residue was part of a β sheet, it was given the description β sheet = 1, left-handed α helix = 0, and right-handed α helix = 0. The final ϕ/ψ angle descriptors, called structural element combinations (SECs), were generated by applying the ACC transformation to the binary secondary structure propensities, separated by a specific number of amino acids in the sequence, as given by eq 2, where *j* and *k* represent the three different structural element classes (β sheet, left-handed α helix, and right-handed α helix) of each of the amino acids in the combined pairs, *a* is the binary classification value of the amino acids, lag is the number of amino acids between the classified ϕ/ψ angles, and *n* is the number of amino acids in the sequence. This will give nine SECs for each amino acid pair (3·3 = 9), and for each SEC, there will be 49 lags, resulting in 49·9 = 441 variables in total. A high value of a SEC for a lag indicates that there is a systematic trend for that specific structural element combination at that specific amino acid distance. This will give information about the overall secondary structure of the protein. More details of these new descriptors are being considered in a separate study currently being performed by our group.

Table 1. Limits Used for Assigning Secondary Element Structure to All Residues of Proteins in the PDBbind Database

structural element	ϕ angle range	ψ angle range
β sheet	$-180° < \phi < -30°$	$60° < \psi < 180°$ $-150° < \psi < 180°$
right-handed α helix	$-140° < \phi < -30°$	$-90° < \psi < 45°$
left-handed α helix	$20° < \phi < 125°$	$-45° < \psi < 90°$

$$\text{SEC}(j,k,\text{lag}) = \frac{\sum_{i=1}^{n-\text{lag}} a_{j,i} a_{k,i+\text{lag}}}{n - \text{lag}} \qquad (2)$$

Block 2: Protein−Ligand Interaction Descriptors. Protein−ligand interactions were described through implementation of the calculations performed in Validate[20] using the SVL programming language in the MOE[34] platform. The following six key features of the protein−ligand complexes were calculated as described by Head et al.[20]: steric and electrostatic complementarity, the ligand strain energy, log *P* (partition coefficient), steric fit, complementary surface area interactions, and the number of rotatable bonds in the ligand. The strain energies of the ligands were calculated by isolatation and energy minimization of the ligands in a vacuum using MMFF94s in MOE[34] via a sequential stepwise procedure of 100 SD iterations, 100 CG iterations, and 200 TN iterations while keeping chiral constraints.

Block 3: Binding Site Surface Area Statistics. The solvent-accessible surface area (SASA) of the binding site in each energy-minimized protein structure was calculated for five different types of atom at an empirically determined distance from the atomic coordinates. The five atom types (Table 2) were classified using PATTY,[25] which is an automatic atom classifier implemented in MOE.[34]

The distance from the atomic coordinate that defined the SASA was empirically determined by examining several complexes where protein−ligand interactions involved the five different atom types. This resulted in SASA distances of 1.5 Å for the hydrophobic, 2.0 Å for the negative, 2.5 Å for the positive, 2.1 Å for the H-bond acceptor, and 2.6 Å for the H-bond donor atom types. The SASA calculations were automated using SVL programming in the MOE platform.[34]

Table 2. Atom Types Defined by PATTY[25] Rules as Applied in MOE[34]

atom type	description
cation (+)	an atom with a basic functionality, i.e., protonated
anion (−)	an atom with an acidic functionality, i.e., deprotonated
donor	an atom that is neither of the above and has a polar H
acceptor	an atom that is none of the above and has a lone pair of electrons
hydrophobic	an atom that is none of the above and is an accessible atom with hydrophobic functionality

In addition to the five atom-type SASAs, defined in Table 2, the SASA of polar atom types was calculated as the sum of negative and positive SASAs together with the total SASA of the binding site based on the atom types listed in Table 2. The six different types of SASA were divided by the total SASA to reduce size dependencies, yielding 13 descriptors, in total, characterizing the binding site SASA (six atom-type SASAs, one total SASA, and six SASA ratios).

Block 4: Physical Chemical Properties of Ligands. To account for the ligands' properties and their contributions to the specificity and strength of their binding to the proteins, 104 2D and 3D descriptors were calculated using the MOE software, for example, molecular weight, solvent-accessible surface areas, van der Waals surface areas, and Kier indexes.[34] For a complete list of descriptors, see the Supporting Information. These descriptors represent a characterization of ligands that is commonly used in QSAR modeling, and it was decided to include a large number of descriptors without a thorough selection of variables to ensure that no relevant information was omitted.

Multivariate Methods. PCA was applied to condense the variation in the data sets into its principal properties and, hence, obtain new information-rich orthogonal latent variables with reduced noise levels. Hi-PLS and standard PLS were applied to relate the descriptor blocks to the biological activity and subsequently predict the strength of the protein−ligand interactions. The K_i and K_d values from the PDBbind database were used in the regression. The binding data were transformed using the negative logarithm expressed as p$K_{d/i}$. The procedure is schematically described in Figure 2, and the multivariate analysis was performed using Simca-P 10.5[40] and Evince[41] software.

Principal Component Analysis. PCA is a projection method in which systematic variation in a data set is extracted into a few variables, so-called principal components.[42] The principal components are linear combinations of the original variables and are uncorrelated to each other as described by

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \mathbf{t}_3\mathbf{p}_3' + ... + \mathbf{t}_A\mathbf{p}_A' + \mathbf{E} = \mathbf{TP}' + \mathbf{E} \quad (3)$$

where **X** is the original data matrix, A is the total number of extracted principal components, and **E** is the residual matrix. The new latent variables, **t** scores, show how the objects and experiments relate to each other, while the **p** loadings reveal the importance of the original variables for the patterns seen in the scores.

Partial-Least-Squares Regression to Latent Structures. PLS is a multivariate regression method that relates a data matrix (**X**) to a response (**Y**).[28] PLS has proven to be a powerful tool for finding relationships between descriptor matrices and biological responses, for example, in QSARs, where many collinear variables are often used. As in PCA, latent variables are constructed to reduce the dimensions of **X**. To obtain the principal components, PLS maximizes the covariance between the response matrix **Y** and a linear combination of the original variables according to

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \mathbf{t}_3\mathbf{p}_3' + ... \mathbf{t}_A\mathbf{p}_A' + \mathbf{E} = \mathbf{TP}' + \mathbf{E} \quad (4)$$

for each **t** so that

$$\mathbf{Y} = \mathbf{t}_1\mathbf{c}_1' + \mathbf{t}_2\mathbf{c}_2' + \mathbf{t}_3\mathbf{c}_3' + ... + \mathbf{t}_A\mathbf{c}_A' + \mathbf{F} = \mathbf{TC}' + \mathbf{F} \quad (5)$$

$$\mathbf{t} = \mathbf{Xw} \quad (6)$$

where **t** is the score vector, **p** is the loading vector for the data matrix **X**, **c** is the loading vector for **Y**, A is the number of PLS components, **E** is the residual matrix for **X**, and **F** is the residual matrix for **Y**. **w** is the weight vector and describes the importance of the variables in **X** to **t** and the response variables **Y**. **w** is selected to maximize the covariance between **t** and **Y**. For a more detailed description of PLS, see Wold et al.[28] The first PLS component weight vector (**w1**) has been used to interpret the obtained models (see below for more details).

Hierarchical PLS. In Hi-PLS, the data matrix **X** is divided into sub-blocks, to each of which PCA is applied prior to the PLS regression. The main advantage with this approach is that it facilitates interpretation when large amounts of data of different kinds are involved.[26] In a standard PLS model, the influence of a block of data with many variables would overwhelm that of a block with few variables, making it difficult to distinguish important information from a small subset of descriptors. In addition, the PCA offers the possibility to reduce the noise level in **X** before the regression modeling and, hence, decreases the risk of chance correlations.[26] The Hi-PLS models were based on the **t** scores from the underlying PCA models performed on the four descriptor blocks (Figure 2).

The number of components for the PCA models of the four blocks was selected so that 85% of the variation in the data was explained, with an eigenvalue larger than 2.0. Manual inspection and interpretation of the PCA scores from the protein structural descriptors (block 1) revealed that only the first four score vectors contained systematic information related to the higher-level SCOP classifications. Hence, for this variable block, only four score vectors were included in subsequent modeling. To account for nonlinearity in the data, the new condensed data sets of PCA scores were further expanded by adding quadratic terms and interaction terms. The AC and ACC transformations performed to obtain the protein structure variables already included internal nonlinear information, so no internal interaction terms were generated for block 1.

On the basis of the expanded hierarchical data set, PLS models were calculated for the data related to all of the complexes (denoted ABCD) as well as for each of the subsets formed by the four SCOP classes: A, B, C, and D. The number of significant PLS components was determined by calculating leave-one-out cross-validation Q^2 values according to

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i^p)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (7)$$

where $\hat{y}_i^p$ is the *i*th predicted p$K_{d/i}$ using the models, y_i is the *i*th experimentally determined p$K_{d/i}$ value, and $\bar{y}$ is the average of these values.[43] The leave-one-out procedure for determining the Q^2 value was selected for its reproducibility and the fact that some of the classes included few objects because the data sets were divided into training and test sets for validation purposes.

Variable Selection. Expanded variables, quadratic or interaction terms, were excluded in an iterative manner if they were shown to lack influence on the model according to the leave-one-out cross-validated Q^2 values (eq 7). The nonlinear variables were removed if their absolute coefficient values were $\leq$ 0.01, and a new PLS model was generated. In the following rounds in this iterative procedure, the cutoff value for excluding terms was progressively increased by 0.01 until the Q^2 value was optimal, and the remaining nonlinear terms were considered to be important for the relationship between the protein–ligand characterizations and their biological responses. All of the complexes from our refined version of the 2002 release of PDBbind were included in this procedure.

Interpretation of the Hi-PLS Models. Interpretation of the influence of the original variables in Hi-PLS proceeds through two levels, first, through the PLS models and, second, through the PCA loading vector **p**.

Our analysis of the influence of the variables in the Hi-PLS models was based on the first PLS component weight vector, **w1**. According to Trygg and Wold, **w1** provides the best estimate of the importance of a variable for describing the response when only one response is used.[44] Variables with small **w1** values do not contribute to the model, so only coefficients with substantial weight vectors were analyzed when interpreting the models. A high, positive **w1** value for a variable indicates that this PCA score vector is positively correlated with the binding affinity, while a high negative value shows that it is negatively correlated with the response. Plots of the PLS weight vectors of the PCA scores of the four descriptor blocks for the Hi-PLS models are described in the Results section.

The PCA score vectors' corresponding loading vectors **p** need to be investigated to further understand the influence of the original descriptors. The PCA loading vectors for the four blocks are given in the Supporting Information, and important notations are discussed in the Results section.

Model Validation. The models were derived using a training set and validated by an internal test set and (subsequently) an external test set that became available after the modeling had been finalized. The performance of the models was estimated by calculating the explained variation in **Y**, expressed as R^2, the root-mean-square error of estimation (RMSEE), the root-mean-square error of prediction (RMSEP), and the objects' distance to the model in **X** space (DModX). In addition, the significance of the models was investigated by comparing them to models based on randomized **Y**.

Selection of Training and Test Sets. Space-filling designs were used to select training and internal test sets from each of the structural classes A–D and for the total set of complexes. The selections were based on data sets composed of the response (p$K_{d/i}$) and the principal properties for each of the four variable blocks. The number of PCA components used was the same as the number used in the Hi-PLS modeling. The resulting data matrices were centered and scaled to unit variance prior to applying the design. The space-filling designs were performed in Matlab[45] according to the Max-min code described by Marengo and Todeschini.[46] Two-thirds of the complexes were selected as training sets, and the remainder formed the internal validation set.

External Test Set. The Hi-PLS models' ability to predict p$K_{d/i}$ values of complexes outside of the selected training sets was investigated using an external test set. New entries were released in the 2003 version of the PDBbind database,[47] which were available to us after the models had been validated by the internal test set. These entries were filtered and refined in the same way as described previously for the PDBbind 2002 release.

Model Permutations. The order of **Y** was randomly permuted 20 times, and models were constructed for each of the randomized **Y**'s. The plot of the correlation coefficient between the original and permuted **Y**'s versus the cumulative R^2 and Q^2 gives a regression line where the intercept (R^2 and Q^2 when the correlation coefficient is zero) is an estimate of the significance of the model.[40] This procedure was applied for all of the classes and the total set of complexes. The plots are given in the Supporting Information, while R^2 and Q^2 values are reported in the Results section below.

RESULTS

The refinement of the 2002 release of the PDBbind database left information on 612 protein–ligand complexes, including biological data, representing four SCOP classes with sufficient representatives for appropriate multivariate modeling. Two-thirds of the complexes were selected as training sets, while the remainder constituted internal test sets. New entries in the 2003 release were used as the external test set. The number of complexes in each set (and class) and the p$K_{d/i}$ ranges can be seen in Table 3.

Hi-PLS modeling was applied to all of the classes to investigate the method's applicability to predict the strength of binding using the four different descriptor blocks. Separate PCA models (one for each of the four descriptor blocks) were calculated for classes A–D, as well as for the total set (referred to as ABCD). The resulting PCA scores were used as input variables in five Hi-PLS models with p$K_{d/i}$ as the response (Table 4). The PCA loadings (**p**) were used to interpret the Hi-PLS weights (**w1)** for the score vectors of blocks 1–4; that is, score vectors with strong Hi-PLS weights were traced back to the original descriptors through their **p** values. Please note that the signs of the **w1** and **p** values are multiplied to give the actual influence on the p$K_{d/i}$, for example, if a score vector with a negative **w1** is traced back to an original descriptor with a negative **p**, this will have a positive contribution to the p$K_{d/i}$. The physicochemical meaning of this interpretation is described for each of the five groups, and the **p**'s are listed in the Supporting Information. The initial models showed a clear nonlinear

Table 3. Description of the Different Classes

			number of complexes in		
class	secondary structure	p$K_{d/i}$ range	training set	internal test set	external test set
ABCD	all α, all β, α/β, and α + β	0.60−13.96	411	201	174
A	all α	1.28−12.00	19	10	16
B	all β	0.60−12.00	185	88	70
C	α/β	1.66−11.11	133	66	61
D	α + β	1.36−13.96	74	37	27

Table 4. Number of Variables Used as X Blocks in the Hi-PLS Modeling

	number of PCA scores used in Hi-PLS				number of
class	block 1	block 2	block 3	block 4	nonlinear terms[a]
ABCD	4	5	4	5	4
A	4	5	4	4	14
B	4	5	4	6	35
C	4	5	4	5	16
D	4	5	4	6	28

[a] Square and cross terms included.

relationship, so the data were expanded by including selected cross and square terms of the PCA scores (see Materials and Methods for details). A summary of the statistical data is shown in Table 5.

Standard PLS, where the original descriptor variables were used directly without extracting the **t** scores, was also performed for comparison (Table 6).

The DModX analyses detected no observations that deviated significantly from the model space. The results from the hierarchical modeling approaches are summarized separately for each class below. As discussed in more detail below, estimates and predictions of p$K_{d/i}$ were poor for a number of deviating complexes, some of which have proved difficult to model in previous studies.[12]

Hierarchical Modeling of All Protein Classes (Class ABCD). The multivariate characterization of protein structures (block 1) using the Euclidian distance between C-α coordinates and SEC descriptors based on ϕ/ψ angles resulted in four principal components. The first and third components, in particular, agreed well with the SCOP classifications, as can be seen in Figure 3, where classes A and B form elongated groups with classes C and D between them.

The Hi-PLS model based on the 22 **X** variables and p$K_{d/i}$ as the response resulted in a single-component model describing 25% of the variation in **Y** with a cross-validated Q^2 value of 0.22 (Table 5). In comparison, models based on randomized noncorrelated **Y** from the permutation analysis gave an $R^2 < 0.05$ and negative Q^2 values (see the Supporting Information). The RMSEE of the model was 1.95, and the p$K_{d/i}$ of the internal test set was predicted with an RMSEP of 1.65 (Table 5 and Figure 4a,b).

All four descriptor blocks contributed to the model, as the Hi-PLS weights (**w1**) for each block differed considerably from zero (Figure 4c). It can also be seen that the majority of the extracted PCA scores added value to the model. However, for the ligand descriptors (block 4), the first PCA score seems to dominate. This component is correlated with the size of the ligands, larger molecules, as expected, being predicted to give stronger binding. In this global model, it can also be seen that the number of nonlinear terms included was fairly small, although those that were included made a notable contribution.

Hierarchical Modeling of Class A. The secondary structure of proteins in this class consists predominantly of α helices. A single-component Hi-PLS model based on 31 variables explained 73% of the variation in p$K_{d/i}$ and had a cross-validated Q^2 value of 0.27 (Table 5). Models based on randomized noncorrelated **Y** from the permutation analysis gave an R^2 of 0.5 and negative Q^2 values (see the Supporting Information). The RMSEE of the model was 1.20, and the p$K_{d/i}$ of the test set was predicted with an RMSEP of 0.80 (Figure 5a,b and Table 5).

Inspection of the **w1** weights revealed that the protein structural descriptors (block 1) were of minor importance for class A, while the binding site surface and ligand descriptor blocks (blocks 3 and 4) appeared to have the strongest impact on the model (Figure 5c). The most important underlying original variables of the binding site were its hydrophobic surface area (which showed a positive correlation to p$K_{d/i}$) and polar surface area descriptors (which showed negative correlations to the binding strength). The underlying original descriptors related to ligand properties (block 4) that appeared to be important were the molecular weight and other size-related variables, for example, the van der Waals surface area and volume.

Hierarchical Modeling of Class B. β sheets predominate in the secondary structure of proteins belonging to this class. A single-component Hi-PLS model based on 54 variables explained 41% of the variation in the p$K_{d/i}$ values and had a cross-validated Q^2 value of 0.30 (Table 5). The permutation study of randomized p$K_{d/i}$ values resulted in a model with an R^2 of 0.1 and negative Q^2 (see the Supporting Information). The calculated and predicted values versus the experimental p$K_{d/i}$ values can be seen in Figure 5d and e, where the RMSEE of the training set was 1.60 and the test set was predicted with an RMSEP of 1.56 (Table 5).

The interpretation of the weight plot (Figure 5f) traced back to the original descriptors through the PCA **p** vectors revealed that the strongest binders in this protein class have a large, hydrophobic binding site with both hydrogen donors and hydrogen acceptors (as indicated by the positive correlations between p$K_{d/i}$ and the total surface area, the hydrophobic surface area, and the acceptor and donor surface areas; block 3). The interpretation of the influence of block 2 (protein−ligand interaction descriptors) also showed that hydrophobic interactions between the ligand and protein were important for strong binding. In addition to a positive correlation between p$K_{d/i}$ and ligand size, the model also showed a negative correlation for ligands with a large fraction of polar-water-accessible surface and high mass density. Another notable feature is that a large number of nonlinear terms was included in the model (Figure 5f).

Hierarchical Modeling of Class C. The structure of these proteins includes interspersed α helices and β sheets. A single-component Hi-PLS model based on 34 variables

Table 5. Summary of Statistical Data for the Hi-PLS Models[a]

class	number of PLS components	R_{tra}^2	Q^2	R_{int}^2	R_{ext}^2	RMSEE	$RMSEP_{int}$	$RMSEP_{ext}$
ABCD	1	0.25	0.22	0.40	0.17	1.95	1.65	1.92
A	1	0.73	0.27	0.67	0.21	1.20	0.80	1.66
B	1	0.41	0.30	0.50	0.31	1.60	1.56	1.86
C	1	0.36	0.15	0.50	0.16	1.97	1.43	2.52
D	1	0.64	0.50	0.77	0.29	1.19	1.13	1.45

[a] tra = training set; int = internal test set; ext = external test set.

Table 6. Summary of Statistical Data for the Standard PLS Models[a]

class	number of PLS components	R_{tra}^2	Q^2	R_{int}^2	R_{ext}^2	RMSEE	$RMSEP_{int}$	$RMSEP_{ext}$
ABCD	2	0.22	0.19	0.32	0.12	1.98	1.75	2.06
A	0	*b*	*b*	*b*	*b*	*b*	*b*	*b*
B	2	0.28	0.19	0.41	0.27	1.77	1.69	1.85
C	1	0.14	0.01	0.23	0.13	2.28	1.59	2.1
D	2	0.57	0.30	0.70	0.28	1.31	1.34	1.50

[a] tra = training set; int = internal test set; ext = external test set. [b] No significant model could be calculated.

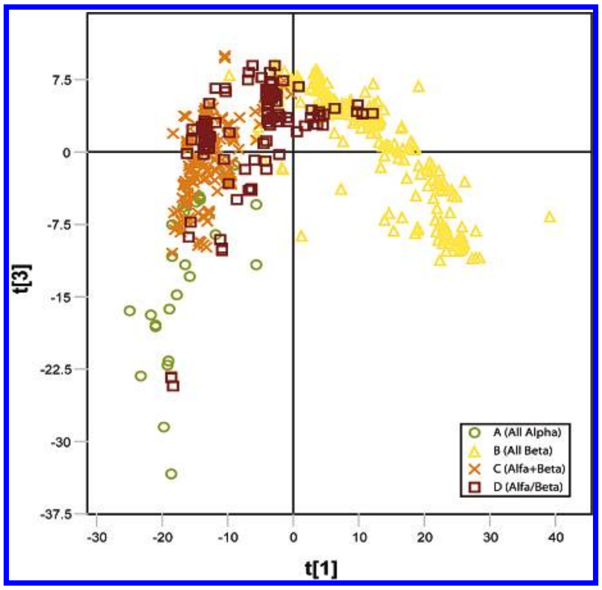


Figure 3. Score plot of the first and third principal components after compression of the multivariate characterization of the protein structure (block 1) of the ABCD class.

explained 36% of the p$K_{d/i}$ variation, with a cross-validated Q^2 value of 0.15 (Table 5). Models based on randomized noncorrelated **Y** from the permutation study gave an R^2 of 0.1 and negative Q^2 (see the Supporting Information). The calculated p$K_{d/i}$ values of the model were estimated with an RMSEE of 1.97, and the internal test set was predicted with an RMSEP of 1.43 (Figure 5g,h). According to the statistical data (Table 5), the model for the class C proteins was weaker than the models for any of the other subsets considered in this paper.

All four descriptor blocks were important for predicting the binding strength. The PLS weight plot (Figure 5i**)** shows that the structural variables of the protein (block 1) had a strong influence on the model, which may reflect the importance of protein structure for predicting the p$K_{d/i}$ values in this class. In addition, interpretation of the influence of the protein−ligand interaction descriptors (block 2) revealed that the potential van der Waals energy of the system and the steric fit of the ligands correlated positively with the biological response. Similar aspects of the underlying surface and ligand descriptors (blocks 3 and 4) seemed to be important for strong binding interactions in this class of proteins to those found for classes A and B—including hydrophobic binding sites and large ligands.

Hierarchical Modeling of Class D. The proteins belonging to this class contain segregated α helices and β sheets. A single-component Hi-PLS model based on 47 variables explained 64% of the variation, with a cross-validated Q^2 value of 0.50 (Table 5). The permutation analysis of randomized **Y** resulted in a model with an R^2 of 0.2 and negative Q^2, that is, a model with a noncorrelated response (see the Supporting Information). The p$K_{d/i}$ values of the training set were estimated with an RMSEE of 1.19, and the internal test set was predicted with an RMSEP of 1.13 (Figure 5j,k and Table 5). These results indicate that the Hi-PLS methodology modeled the interactions of the class D proteins more successfully than those of any other classes considered here.

The interpretation of the PLS weight plot (Figure 5) focused on blocks 2 and 3 (protein−ligand interactions and binding site surface area), showing that complementary hydrophobic interactions favored strong binding, together with the presence of hydrogen-bond-donating and -accepting areas in the binding site. In addition to the ligand size, two other PCA loading vectors correlated strongly to the p$K_{d/i}$ values, and interestingly, these vectors did not reflect the weight of the ligands. Instead, they were dominated by fractional ligand surface areas, fractions of hydrophobic and negative surface area having a positive correlation with p$K_{d/i}$, while fractions of polar and positive surface areas had a negative effect. In addition, increased numbers of aromatic atoms and bonds, and a high lipophilicity, should, according to the model, increase the binding strength.

External Validation of the Hi-PLS Models. The 2003 release of PDBbind[47] was made available to the public after the models in this paper had been developed. Hence, it gave an opportunity to use new entries as an external test set to assess the ability of our Hi-PLS models to predict the p$K_{d/i}$ values, with a total set of 174 proteins (after the cleanup procedure). The proteins were divided into 18 class A

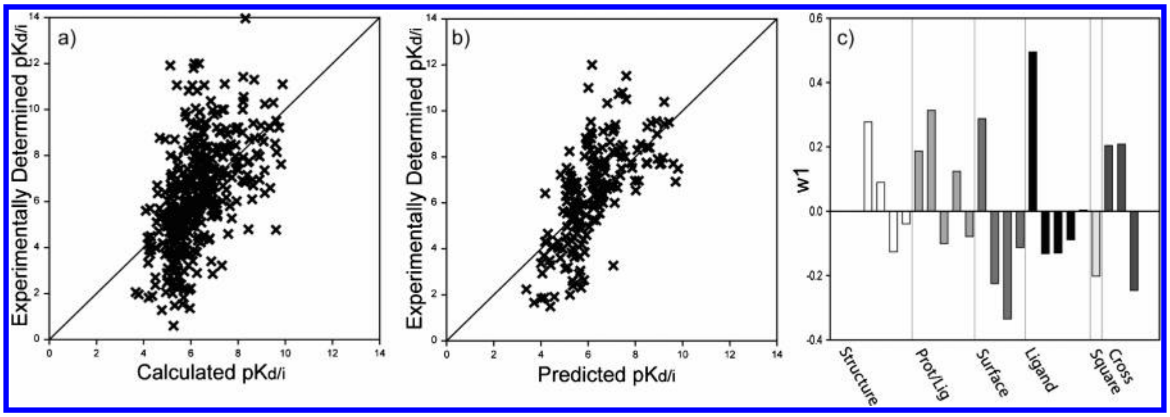


Figure 4. Results of the Hi-PLS modeling based on the total set of complexes (ABCD), showing (a) the calculated p$K_{d/i}$ values of the training set versus the experimental data, (b) the predicted p$K_{d/i}$ values of the internal test set versus experimental data, and (c) the Hi-PLS weight vectors (**w1**) of the four condensed descriptor blocks (Structure = block 1; Prot/Lig = block 2; Surface = block 3; Ligand = block 4).

proteins, 70 class B proteins, 61 SCOP class C proteins, and 27 class D proteins, and their p$K_{d/i}$ values were predicted using the corresponding models in Table 5 (Figure 6a−e and Table 5). The p$K_{d/i}$ values of the total set of complexes were predicted with a RMSEP of 1.92, and the best predictions were for classes A and D, with RMSEPs of 1.66 and 1.45, respectively. The p$K_{d/i}$ values of classes B and C were predicted with RMSEPs of 1.86 and 2.52, respectively.

DISCUSSION

Hi-PLS Modeling for Predicting Binding Strength. Generally the Hi-PLS models gave better results than standard PLS models for the data presented in this work (Tables 5 and 6). This was probably because Hi-PLS can extract more important trends from complex data sets than standard PLS. Another key feature of the Hi-PLS modeling was the inclusion of nonlinear terms, which were not included in the standard PLS. The high number of original variables would have led to an explosion of expanded terms in standard PLS, approximately 71 500 square and interaction terms. The collinear nature of the original descriptors and the risk of chance correlation with the high number of variables made it inappropriate to expand the standard PLS with nonlinear terms.

The models based on all of the complexes (class ABCD) and the subclass sets (classes A−D) using four descriptor blocks and p$K_{d/i}$ values as the biological response showed statistical significance when the models were validated with a permutation analysis. On the basis of the results presented in Table 5, we can conclude that the models were moderately successful for classes A (all α helices) and D (α helices and β sheets segregated), while the models for class B (all β sheets), class C (α helices and β sheets interspersed), and class ABCD (all complexes) gave rather weak predictions. Clearly, although the models were significant, more work needs to be done before they can be generally used as scoring functions. Follow-up work will include a closer look at different descriptor sets as well as modifications of the regression technique using more refined techniques such as O-PLS.[44,48] In addition, further attention to the pretreatment of the protein−ligand complexes (e.g., the protonation states) and force field parametrization (see model deviators below) may improve the predictions.

Interpretation of the Hi-PLS Models. Hierarchical modeling facilitated identification of the complexes' characteristics that were important for binding. Four descriptor blocks were used in the modeling, representing physicochemical characterizations of the protein structures, the binding sites, the interactions between the proteins and ligands, and the ligands themselves. Originally these matrices were of different sizes (Figure 2), which would have biased their influence on the models if they had been used directly. This potential problem was circumvented by compressing the separate data sets using PCA, yielding similar numbers of variables for all blocks (Table 4). The fact that all four blocks were shown to be important in the modeling procedure strengthens the hypothesis that features believed to be important for binding, for example, protein and ligand properties, can be characterized separately and merged to constitute a larger matrix.

Investigation and comparison of the PLS weight plots (Figure 5c,f,i,l) and the underlying descriptors (Supporting Information) allowed detailed information about important factors to be obtained. Large lipophilic ligands and lipophilic binding sites were generally important for all of the classes and showed positive correlations with the p$K_{d/i}$ values. These results are not surprising, but some more subtle information was also obtained. For instance, binding strength among the class A protein−ligand complexes was clearly favored by large ligands and hydrophobic binding sites. If generally applicable, these findings could help medicinal chemists to identify promising new targets for drug discovery programs. The model based on the structures belonging to class B had the most complex curvature pattern, reflected in its high number of nonlinear terms (Figure 5f and Table 4), indicating that there are intricate nonlinear relationships between the structural descriptors and the strength of ligand binding in these complexes. It may be helpful for researchers with specific interests in this class of proteins to refine the nonlinear modeling or further divide the subgroup. For class D, interpretation of the variables showed that the protein structure block was not as important as the other blocks of descriptors, indicating that complexes in this structural class may be fairly homogeneous with respect to ligand binding. The influence of the ligand descriptors for this protein class was less dependent on the size and weight of the ligands than it was for the other structural classes, suggesting that there may be relatively high potential to develop small, soluble high-affinity ligands for proteins of this class.

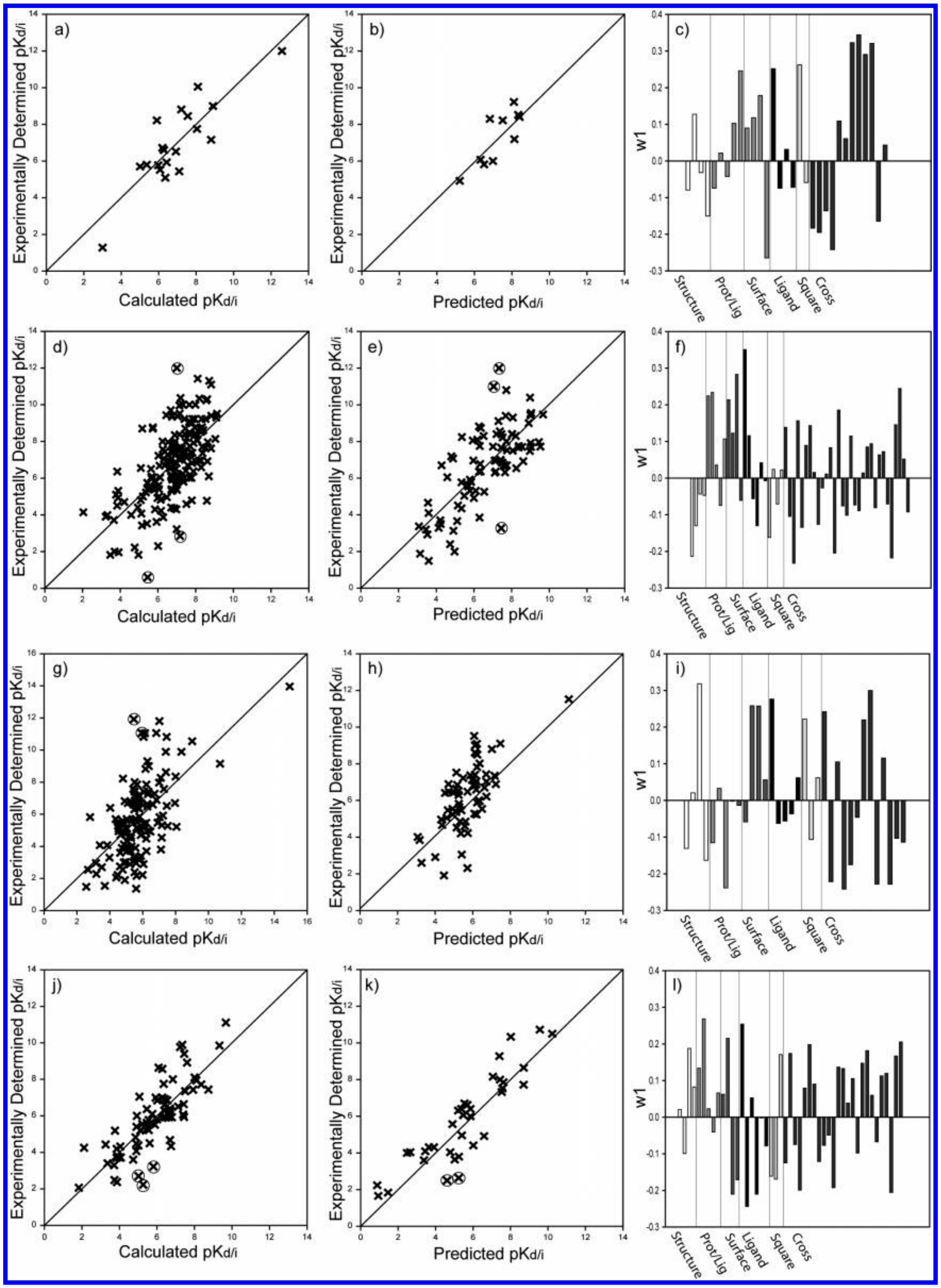


Figure 5. Results of the Hi-PLS modeling based on the subsets, classes A (a−c), B (d−f), C (g−i), and D (j−l). Parts a, d, g, and j show the calculated p$K_{d/i}$ values of the training set versus the experimental data; parts b, e, h, and k show the predicted p$K_{d/i}$ values of the internal test set versus experimental data, and parts c, f, i, and l show the Hi-PLS weight vectors (**w1**) of the four condensed descriptor blocks. Encircled data points denote deviating observations, discussed in the text.

Model Deviators. In both the training and test sets of our extended Hi-PLS models, there were several deviating observations with respect to the estimated and predicted p$K_{d/i}$ values. The deviation of a complex may be due to model errors or to a lack of data needed to correctly describe all of the observations in the data set. We chose not to exclude any objects and instead tried to highlight possible causes for the deviations because we think that this information could be valuable for further improvement of the models and the development of scoring functions.

Many of the deviators seen in this study were complexes that have previously proved difficult to model using linear approaches, such as standard PLS, or nonlinear regression modeling.[12] We anticipated that the use of Hi-PLS modeling with nonlinear terms could improve the predictions for these known deviators. A comparison of the estimation errors of

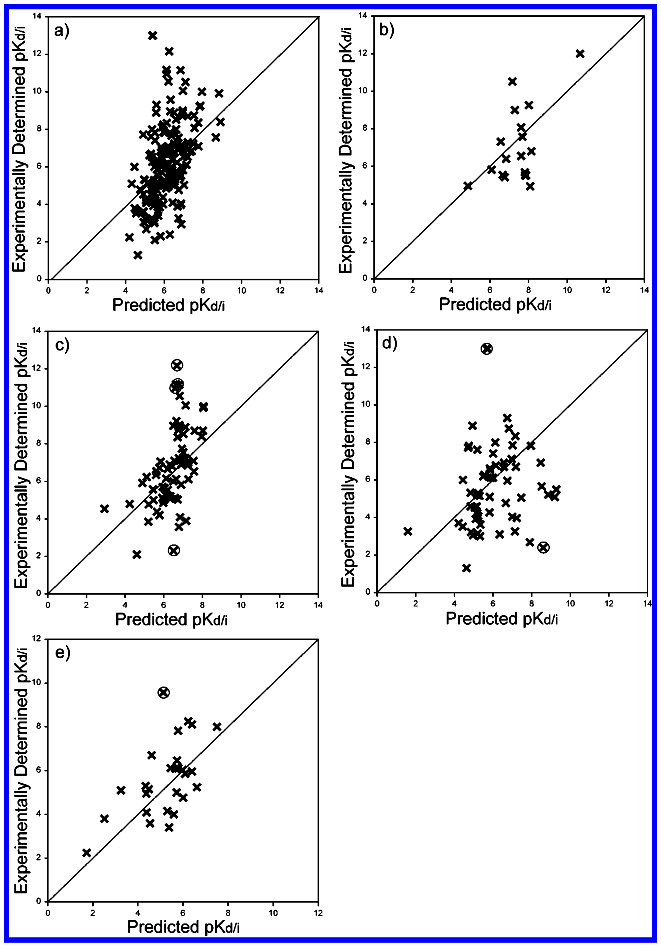


Figure 6. Experimental versus predicted p$K_{d/i}$ values for the external test set using the Hi-PLS models for (a) class ABCD (all complexes), (b) class A (all α helices), (c) class B (all β sheets), (d) class C (α helices and β sheets interspersed), and (e) class D (α helices and β sheets segregated). Encircled data points denote deviating observations discussed in the text.

the p$K_{d/i}$ values of the 19 outliers reported by Wang et al.[12] and the corresponding errors in our Hi-PLS models is shown in Table 7.

Three of these 19 protein−ligand complexes were excluded by our refinement procedure and, hence, were not included in our models. Our Hi-PLS models gave better p$K_{d/i}$ estimations (improvement of more than one log unit) for six complexes, while the errors for the remaining 10 complexes were in the same range as those obtained by the commonly used scoring functions. However, it should be noted that most of the values of the deviators were estimated rather than predicted values, because they were included in the space-filling design to increase the diversity of the training set.

In addition to the deviators previously presented by Wang et al.[12] (Table 7), we found large errors in the estimated biological responses of several other complexes. All of the deviators were investigated with great care using the original reference for the crystallization procedure and the determination of p$K_{d/i}$. Some interesting findings that may explain the weak modeling results are discussed below.

Large proportions of the errors in the p$K_{d/i}$ estimates are probably due to conformational changes in the proteins upon binding to their respective ligands, resulting in the estimated p$K_{d/i}$ values being higher than the experimentally determined values. For example, the main deviators in class D were the complexes 1AI4 (p$K_{i(exptl)} = 2.50$, p$K_{i(calcd)} = 4.61$), 1AJN (p$K_{i(exptl)} = 2.64$, p$K_{i(calcd)} = 5.24$), and 1AJP (p$K_{i(exptl)} = 2.23$, p$K_{i(calcd)} = 5.26$), complexes of the enzyme penicillin acylase, which catalyses the cleavage of an amide bond in benzylpenicillin. Done et al. have shown that two subsets of substrates of this enzyme bind via two different modes.[49] The binding of one of the subsets of ligands (comprising the deviating complexes 1AI4, 1AJN, and 1AJP) causes a major conformational change in the protein. The complexes which do not induce this movement (1AJQ and 1AI7) were not deviators in the models (1AJQ p$K_{i(exptl)} = 4.31$, p$K_{i(calcd)} = 3.91$; 1AI7 p$K_{i(exptl)} = 4.09$, p$K_{i(calcd)} = 3.45$). For the deviant complexes 1BR5 (p$K_{i(exptl)} = 2.70$, p$K_{i(calcd)} = 5.02$) and 1BR6 (p$K_{i(exptl)} = 3.22$, p$K_{i(calcd)} = 5.81$), there is also a large conformational change following binding, in a tyrosine residue in the active site.[50] The change in conformation causes a difference in entropic energy between the bound and unbound state that is not captured at all by the descriptors in our model, explaining the poor predictions. Other weakly deviating binders were the class B ligands in complexes 4SGA (p$K_{i(exptl)} = 3.20$, p$K_{i(calcd)} = 7.20$) and 5SGA (p$K_{i(exptl)} = 2.85$, p$K_{i(calcd)} = 7.17$). The incorrect predictions in these cases may be due to a very large conformational movement of an imidazole ring upon binding and the displacement of 16 water molecules which occupy the active site in the unbound state.[51]

Another source of estimation errors is believed to be the poor quality of the molecular mechanic calculations when metal chelations and unusual transition states are present, because the force field parametrization is most likely insufficient in such cases. Errors of this type may have been responsible for several deviators among the class C complexes. The high-affinity binding complex 1CTU (p$K_{i(exptl)} = 11.92$, p$K_{i(calcd)} = 5.46$) has a binding mode that is enhanced by the complexation between Zn^{2+} and a water molecule. This complexation leads to an “entropy trap mechanism” in which the entropic contribution to the process is highly unfavorable (−10.5 kcal/mol at 25 °C) but is compensated by strong binding enthalpies in the covalently hydrated binding state.[52] The trapping of the water molecule was probably not correctly accounted for in the molecular mechanics calculations, and the resulting error indicates that water molecules in the binding site need to be appropriately included in calculations of the protein−ligand interaction descriptors. For the structure 1LOR (p$K_{i(exptl)} = 11.06$, p$K_{i(calcd)} = 5.93$), the transition state in the enzymatic reaction is thought to be highly favorable compared to other, similar protein−ligand complexes,[53] which could explain its weak predictions. At the lower end of the p$K_{d/i}$ values, 1ZSB (class B; p$K_{d(exptl)} = 0.60$, p$K_{d(calcd)} = 5.46$) was poorly modeled, possibly because the polarity of a histidine in the binding site may be reversed in a protein−ligand−water−Zn^{2+} complex,[54] and the resulting histidine anion was not parametrized in the force field used prior to our modeling. In the predictions for the external test set of the PDBbind 2003 release proteins, there were also several deviating observations. The complexes 2ADA[55] (class C; p$K_{i(exptl)} = 13.00$, p$K_{i(calcd)} = 5.67$), 1A4M[56] (class C; p$K_{i(exptl)} = 13.00$, p$K_{i(calcd)} = 5.73$), and 1O86[57] (class D; p$K_{i(exptl)} = 9.57$, p$K_{i(calcd)} = 5.13$) all include a chelated metal ion (Zn^{2+}), which may account for the poor predictions, and the active site of 1CRU[58] (class B; p$K_{i(exptl)} = 2.28$, p$K_{i(calcd)} = 5.72$) includes

Table 7. Comparison of the Estimation Errors of the $\mathrm{p}K_{\mathrm{d/i}}$ Values for the 19 Outliers Reported by Wang[12] Obtained Using the Subclass Hi-PLS Models in Table 5 and Five Common Scoring Functions[12]

PDB code	SCOP class	experimental $\mathrm{p}K_{\mathrm{d/i}}$	calculated $\mathrm{p}K_{\mathrm{d/i}}$[a]	error of calculated value	mean error in scoring function[b]
1N4K	A	10.05	8.10[c]	−1.95	4.33
1B8O	A	10.64	*d*	*d*	4.29
1SWN	B	12.00	7.00[e]	−5.00	5.49
1SWK	B	12.00	7.00[c]	−5.00	5.20
1ZSB	B	0.60	5.46[c]	4.86	4.57
1IF7	B	10.52	*f*	*f*	4.47
1BNN	B	10.00	*f*	*f*	4.09
5SGA	B	2.85	7.17[c]	4.32	3.98
7CPA	C	13.96	14.94[c]	0.98	6.24
1CTU	C	11.92	5.46[c]	−6.46	5.95
1QPB	C	1.36	5.60[c]	4.24	5.46
1ELS	C	10.82	7.45[c]	−3.37	5.41
1DUV	C	11.80	7.01[c]	−4.79	4.98
1DQX	C	11.05	6.14[c]	−4.91	4.80
1LOR	C	11.06	5.93[c]	−5.13	4.70
1RBO	C	10.55	9.00[c]	−1.55	4.36
1XLI	C	1.48	2.57[c]	1.09	4.34
1M0N	C	2.22	5.99[c]	3.77	4.25
1M0O	C	2.31	5.71[e]	3.40	4.09

[a] Value obtained using the Hi-PLS models in Table 5. [b] The mean error of the values reported by Wang obtained using five popular scoring functions.[12] [c] Training set estimation. [d] Excluded because of the presence of $\mathrm{PO_4}^{3-}$ in the binding site. [e] Test set prediction. [f] Excluded because of the presence of a Hg^{+} ion in the binding site but not included in the assay.

a Ca^{2+} that may not be correctly accounted for by the force field calculations. For complex 1NWl[59] (class C; $\mathrm{p}K_{\mathrm{i(exptl)}} = 2.39$, $\mathrm{p}K_{\mathrm{i(calcd)}} = 8.61$) in the external test set, the ligand binding is affected by a Ca^{2+} ion located outside the 10 Å radius that defined the active site in the force field calculations and, hence, was not captured in the calculations.

In addition to the above-mentioned sources of error, we concluded that our models could not completely account for the tight binding of the streptavidin−biotin complexes (class B; e.g., 1SWK $\mathrm{p}K_{\mathrm{d(exptl)}} = 12.00$, $\mathrm{p}K_{\mathrm{d(calcd)}} = 7.00$; 1SWN $\mathrm{p}K_{\mathrm{d(exptl)}} = 12.00$, $\mathrm{p}K_{\mathrm{d(calcd)}} = 7.00$; and 1SWP $\mathrm{p}K_{\mathrm{d(exptl)}} = 11.00$, $\mathrm{p}K_{\mathrm{d(calcd)}} = 6.66$). These complexes are known to have extremely hydrophobic binding sites and extensive hydrogen-bonding networks that may not have been correctly parametrized.[60] Similar deviations were again seen when such complexes (class B; e.g., 1NDJ $\mathrm{p}K_{\mathrm{d(exptl)}} = 12.10$, $\mathrm{p}K_{\mathrm{d(calcd)}} = 6.69$; 1N43 $\mathrm{p}K_{\mathrm{d(exptl)}} = 10.55$, $\mathrm{p}K_{\mathrm{d(calcd)}} = 6.83$; 1N9M $\mathrm{p}K_{\mathrm{d(exptl)}} = 10.96$, $\mathrm{p}K_{\mathrm{d(calcd)}} = 6.60$; and 1NC9 $\mathrm{p}K_{\mathrm{d(exptl)}} = 11.17$, $\mathrm{p}K_{\mathrm{d(calcd)}} = 6.64$) were included in the external test set.

CONCLUSIONS

The Hi-PLS models of the binding strength of a large, diverse set of protein−ligand complexes, on the basis of thorough characterization of the ligands and proteins, had a moderate ability to model and predict the $\mathrm{p}K_{\mathrm{d/i}}$ values of internal and external test sets of complexes (the latter obtained from the 2003 release of the PDBbind database). The results showed that all four descriptor blocks (protein structure, protein−ligand interaction, binding site, and ligand descriptors) contributed to the models. An inspection of the weight vectors of the four different blocks allowed chemical features that are important for binding (both generally and at a more-detailed level) to be identified.

Division of the protein−ligand complexes into four groups according to the SCOP classification scheme, based on the distribution of α helices and β sheets in the 3D structure, gave improved models for the class A (all α helices) and class D (α helices and β sheets segregated) proteins in terms of the estimates and predictions for the internal and external test sets. The global model based on all of the complexes (class ABCD) did not have satisfactory predictive ability, nor did the models for the class C (α helices and β sheets mixed) or class B (all β sheets) complexes.

Many of the deviating observations in the resulting models were also common deviators in analyses by several popular scoring functions, and the errors in the modeled or predicted values of $\mathrm{p}K_{\mathrm{d/i}}$ in our Hi-PLS models were in the same range or better than previously reported values. Generally, the deviators in the models were complexes in which a major conformational change occurs upon binding or complexes that include unusual geometries due to ligand−metal chelation that have not yet been sufficiently parametrized in the force field applied.

The Hi-PLS modeling presented in this paper gave significant, but not satisfactory, predictions of $\mathrm{p}K_{\mathrm{d/i}}$ values. The presented methodology should be seen as an initial attempt to use different descriptor blocks together with Hi-PLS in the difficult field of estimating binding affinity. Further developments of relevant descriptors together with a refinement of the Hi-PLS regression method offer promising approaches to develop more focused scoring functions.

ACKNOWLEDGMENT

The authors thank AstraZeneca R&D Mölndal, Sweden, and Stiftelsen J.C. Kempes Minnes Stipendie Fond for financially supporting this project. The authors are grateful for the valuable support and ideas given by Doctor Johan Trygg at Umeå University.

Supporting Information Available: Tables of the 2D and 3D descriptors used for multivariate modeling, tables of the variable loadings for each score vector for each Hi-PLS model presented in the work, and the permutation study made for validating the models. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(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) Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L. N. Rapid Measurement of Binding Constants and Heats of Binding using a New Titration Calorimeter. *Anal. Biochem.* **1989**, *179*, 131−137.

(3) Raffa, R. B.; Porreca, F. Thermodynamic Analysis of the Drug−Receptor Interaction. *Life Sci.* **1989**, *44*, 245−258.

(4) Porstmann, T.; Kiessig, S. T. Enzyme Immunoassay Techniques. An Overview. *J. Immunol. Methods* **1992** *150* (1−2), 5−21.

(5) Villar, H. O.; Yan, J.; Hansen, M. R. Using NMR for Ligand Discovery and Optimization. *Curr. Opin. Chem. Biol.* **2004**, *8*, 387−391.

(6) Löfås, S. Optimizing the Hit-to-Lead Process Using SPR Analysis. *Assay Drug Dev. Technol.* **2004**, *2*, 407−415.

(7) Lewis, R. M.; Leach, A. R. Current Methods for Site-Directed Structure Generation. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 467−475.

(8) 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.

(9) Caflisch, A.; Miranker, A.; Karplus, M. Multiple Copy Simultaneous Search and Construction of Ligands in Binding Sites: Application to Inhibitors of HIV-1 Aspartic Proteinase. *J. Med. Chem.* **1993**, *36* (15), 2142−2167.

(10) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.

(11) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B. O.; Glen, R. C. Predicting Protein−Ligand Binding Affinities: A Low Scoring Game? *Org. Biomol. Chem.* **2004**, *2*, 3267−3273.

(12) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions using the PDBbind Refined Set of 800 Protein−Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.

(13) Böhm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein−Ligand Complex of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(14) Böhm, 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.

(15) 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.

(16) Wang, R.; Lui, L.; Lai, L.; Tang, Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein−Ligand Complex. *J. Mol. Model.* **1998**, *4*, 379−394.

(17) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

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

(19) Puvanendrampillai, D.; Mitchell, J. B. O. Protein Ligand Database (PLD): Additional Understanding of the Nature and Specificity of Protein−Ligand Complexes. *Bioinformatics* **2003**, *19*, 1856−1857.

(20) Head, R.; 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.

(21) Oprea, T. I.; Marshall, G. R. Receptor-Based Prediction of Binding Affinities. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 35−61.

(22) Zamora, I.; Oprea, T.; Cruciani, G.; Pastor, M.; Ungel, A. L. Surface Descriptors for Protein−Ligand Affinity Prediction. *J. Med. Chem.* **2003**, *46*, 25−33.

(23) Ajay; Murcko, M. A. Computational Methods to Predict Binding Free Energy in Ligand−Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.

(24) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(25) 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.

(26) Eriksson, L.; Johansson, E.; Lindgren, F.; Sjöström, M.; Wold, S. Megavariate Analysis of Hierarchical QSAR Data. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 711−726.

(27) Gunnarsson, I.; Andersson, P.; Wikberg, J.; Lundstedt, T. Multivariate Analysis of G Protein-Coupled Receptors. *J. Chemom.* **2003**, *17*, 82−92.

(28) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(29) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chotia, C. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* **1995**, *247*, 536−540.

(30) Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data. *Nucleic Acids Res.* **2004**, *32*, D226−D229.

(31) *Reduce 2.21 for Linux*; The Richardson Laboratory, Duke University: Durham NC. http://kinemage.biochem.duke.edu (accessed Sept 2004).

(32) Petersen, M. T. N.; Jonson, P. H.; Petersen, S. B. Amino Acid Neighbours and Detailed Conformational Analysis of Cysteines in Proteins. *Protein Eng.* **1999**, *12*, 535−548.

(33) Labahn, J.; Neumann, S.; Büldt, G.; Kula, M. R.; Granzin, J. An Alternative Mechanism for Amidase Signature Enzymes. *J. Mol. Biol.* **2002**, *322*, 1053−1064.

(34) *MOE 2004.03 for PC and Linux*; Chemical Computing Group: Montreal, Quebec, Canada.

(35) Box, G. E. P.; Jenkins, G. M. *Time Series Analysis;* Holden-Day: Oakland, CA, 1976.

(36) Wold, S.; Jonsson, M.; Sjöström, M.; Sandberg, M.; Rännar, S. DNA and Peptide Sequences and Chemical Processes Multivariately Modelled by Principal Component Analysis and Partial Least-Squares Projections to Latent Structures. *Anal. Chim. Acta* **1993**, *277*, 239−253.

(37) Edman, M.; Jarhede, T.; Sjöström, M.; Wieslander, Å. Different Sequence Patterns in Signal Peptides From Mycoplasmas, Other Gram-Positive Bacteria, and *Escherichia coli*: A Multivariate Data Analysis. *Proteins* **1999**, *35*, 195−205.

(38) Long, I.; Andersson, P. E. S.; Lundstedt, T. Multivariate Analysis of Five GPCR Receptor Classes. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 95−104.

(39) Ramachandran, G., N.; Sasisekhran, V. Conformation of Polypeptides and Proteins. *Adv. Protein Chem.* **1968**, *28*, 283−437.

(40) *Simca-P 10.5 for PC*; Umetrics: Umeå, Sweden.

(41) *Evince for PC and Linux*; Umbio: Umeå, Sweden.

(42) Jackson, J. E. *A Users Guide to Principal Components*; Wiley: New York, 1991.

(43) Wold, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397−405.

(44) Trygg, J.; Wold, S. Orthogonal Projections to Latent Structures (O-PLS). *J. Chemom.* **2002**, *16*, 119−128.

(45) *Matlab for Windows*; MathWorks, Inc.: Boston, MA.

(46) Marengo, E.; Todeschini, R. A New Algorithm for Optimal Distance-Based Experimental Design. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 37−44.

(47) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48* (12), 4111−4119.

(48) Trygg, J. O2-PLS for Qualitative and Quantitative Analysis in Multivariate Calibration. *J. Chemom.* **2002**, *16*, 283−293.

(49) Done, S. H.; Branningan, J. A.; Moody, P. C. E.; Hubbard, R. E. Ligand-Induced Conformational Change in Penicillin Acylase. *J. Mol. Biol.* **1998**, *284*, 463−475.

(50) Yan, X.; Hollis, T.; Svinth, M.; Day, P.; Monzingo, A. F.; Milne, G. W. A.; Robertus, J. D. Structure-Based Identification of a Ricin Inhibitor. *J. Mol. Biol.* **1997**, *266*, 1043−1049.

(51) James, M. N.; Sielecki, A. R.; Brayer, G. D.; Delbaere, L. T.; Bauer, C. A. Structures of Product and Inhibitor Complexes of *Streptomyces Griseus* Protease A at 1.8 Å Resolution. A Model for Serine Protease Catalysis. *J. Mol. Biol.* **1980**, *144*, 43−88.

(52) Xiang, S.; Short, S. A.; Wolfenden, R.; Carter, C. W., Jr. Transition-State Selectivity for a Single Hydroxyl Group during Catalysis by Cytidine Deaminase. *Biochemistry* **1995**, *34*, 4516−4523.

(53) Wu, N.; Pai, E. F. Crystal Structures of Inhibitor Complexes Reveal an Alternate Binding Mode in Orotidine-5′-monophosphate Decarboxylase. *J. Biol. Chem.* **2002**, *227*, 28080−28087.

(54) Huang, C.; Lesburg, C. A.; Kiefer, L. L.; Fierke, C. A.; Christianson, D. W. Reversal of the Hydrogen Bond to Zinc Ligand Histidine-119 Dramatically Diminishes Catalysis and Enhances Equilibrium Kinetics in Carbonic Anhydrase II. *Biochemistry* **1996**, *35*, 3439−3446.

(55) Wilson, D. K.; Rudolph, F. B.; Quiocho, F. A. Atomic Structure of Adenosine Deaminase Complexed with a Transition-State Analog:

Understanding Catalysis and Immunodeficiency Mutations. *Science* **1991**, *252*, 1278−1284.

(56) Wang, Z.; Quiocho, F. A. Complexes of Adenosine Deaminase with Two Potent Inhibitors: X-ray Structures in Four Independent Molecules at pH of Maximum Activity. *Biochemistry* **1998**, *37*, 8314−8324.

(57) Natesh, R.; Schwager, S.; Sturrock, E.; Acharya, K. Crystal Structure of the Human Angiotensin-Converting Enzyme−Lisinopril Complex. *Nature* **2003**, *421*, 551−554.

(58) Oubrie, A.; Rozeboom, H. J.; Dijkstra, B. W. Active-Site Structure of the Soluble Quinoprotein Glucose Dehydrogenase Complexed with Methylhydrazine: A Covalent Cofactor−Inhibitor Complex. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11787−11791.

(59) Peisach, D.; Gee, P.; Kent, C.; Xu, Z. The Crystal Structure of Choline Kinase Reveals a Eukaryotic Protein Kinase Fold. *Structure* **2003**, *11*, 703−713.

(60) Freitag, S.; Le Trong, I.; Chilkoti, A.; Klumb, L. A.; Stayton, P. S.; Stenkamp, R. E. Structural Studies of Binding Site Tryptophan Mutants in the High-Affinity Streptavidin−Biotin Complex. *J. Mol. Biol.* **1998**, *279*, 211−221.

CI050323K