Interaction Model Based on Local Protein Substructures Generalizes to the Entire Structural Enzyme-Ligand Space

Helena Strömbergsson,[†] Pawel Daniluk,[‡,§] Andriy Kryshtafovych,[||] Krzysztof Fidelis,[||] Jarl E. S. Wikberg,[⊥] Gerard J. Kleywegt,[*,#] and Torgeir R. Hvidsten[†,▽]

The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden, Department of Biophysics, Faculty of Physics, University of Warsaw, Warsaw, Poland, Department of Biophysics and CoE BioExploratorium, Faculty of Physics, University of Warsaw, Warsaw, Poland, UC Davis Genome Centre, UC Davis, Davis, California, Department of Pharmaceutical Pharmacology, Uppsala University, Uppsala, Sweden, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden, and Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden

Received June 12, 2008

Chemogenomics is a new strategy in *in silico* drug discovery, where the ultimate goal is to understand molecular recognition for all molecules interacting with all proteins in the proteome. To study such cross interactions, methods that can generalize over proteins that vary greatly in sequence, structure, and function are needed. We present a general quantitative approach to protein−ligand binding affinity prediction that spans the entire structural enzyme-ligand space. The model was trained on a data set composed of all available enzymes cocrystallized with druglike ligands, taken from four publicly available interaction databases, for which a crystal structure is available. Each enzyme was characterized by a set of local descriptors of protein structure that describe the binding site of the cocrystallized ligand. The ligands in the training set were described by traditional QSAR descriptors. To evaluate the model, a comprehensive test set consisting of enzyme structures and ligands was manually curated. The test set contained enzyme-ligand complexes for which no crystal structures were available, and thus the binding modes were unknown. The test set enzymes were therefore characterized by matching their entire structures to the local descriptor library constructed from the training set. Both the training and the test set contained enzyme-ligand complexes from all major enzyme classes, and the enzymes spanned a large range of sequences and folds. The experimental binding affinities (pK_i) ranged from 0.5 to 11.9 (0.7−11.0 in the test set). The induced model predicted the binding affinities of the external test set enzyme-ligand complexes with an r^2 of 0.53 and an RMSEP of 1.5. This demonstrates that the use of local descriptors makes it possible to create rough predictive models that can generalize over a wide range of protein targets.

INTRODUCTION

Experimental characterization of molecular interactions on a proteome-wide basis is not realistic, considering that there are over 30000 proteins in the human genome that can theoretically interact with an astronomic number ($\sim 10^{62}$)[1] of small organic molecules (ligands or drugs). Chemogenomics is a new strategy in drug discovery that attacks this problem by attempting to fully describe protein−ligand space and ultimately identify all ligands for all protein drug targets.[2] However, existing computational approaches to interaction modeling typically apply to single protein−ligand complexes of one or a few related proteins interacting with several ligands.[3−5] However, computational models for predicting protein−ligand interactions on a proteome-wide scale should apply to proteins that vary greatly in terms of sequence, structure, and function and over ligands that vary greatly in structure, chemical composition, and physicochemical properties. Such models could have a huge impact on drug discovery, allowing for prediction of novel cross-interaction side-effects and rapid screening of entire proteomes and compound databases.

The most common approach to protein−ligand interaction prediction is computational docking of ligands into the 3D structure of a protein target.[4,6] This requires the use of empirical scoring or energy functions to rank different poses of the same ligand and to rank the fit of different ligands to a protein. Although the calculations are limited to the protein active site, docking is rather computationally expensive, in particular when the algorithm allows for flexibility not only in the ligand conformation but also in the protein backbone and side chains (induced fit). Scoring functions are often trained on and adjusted for targets and ligands of a particular type and thus do not necessarily generalize to diverse high-throughput docking data sets.[7] The development of faster and more accurate scoring functions is currently a particularly

* Corresponding author phone: +46 18 471 4870; fax: +46 18 53 03 96; e-mail: gerard@xray.bmc.uu.se. Corresponding author address: Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751, 24 Uppsala, Sweden.
[†] The Linnaeus Centre for Bioinformatics, Uppsala University.
[‡] Department of Biophysics, University of Warsaw.
[§] Department of Biophysics and CoE BioExploratorium, University of Warsaw.
[||] UC Davis Genome Centre.
[⊥] Department of Pharmaceutical Pharmacology, Uppsala University.
[#] Department of Cell and Molecular Biology, Uppsala University.
[▽] Umeå Plant Science Centre, Department of Plant Physiology, Umeå University.

10.1021/ci800200e CCC: $40.75
Published on Web 10/21/2008

active area of research.[8] Docking is well suited for assessing whether ligands fit into a protein but are in general not able to accurately predict protein−ligand binding affinities.[6]

Approaches to generate models that predict protein−ligand binding affinity are based on the classical machine learning setup: complexes with experimentally measured binding affinities are used as training examples in a regression or classification method in which ligands are encoded by various descriptors. Traditionally, quantitative structure−activity relationship (QSAR)[3,9] approaches have been applied to accurately predict how strongly one series of chemically related ligands binds to a single protein target using ligand descriptors such as molecular weight and number of hydrogen bond donors/acceptors.[10,11] Unfortunately, QSAR requires large training sets for every single protein target since it does not take advantage of affinity data available for related proteins. Consequently, QSAR methods are unable to predict cross interactions between ligands and "unseen" proteins and are thus not suitable for proteome-wide modeling. Proteochemometrics (PCM) was introduced to remedy this by extending QSAR modeling to include protein descriptors.[12] Protein and ligand descriptors, together with experimentally determined binding affinity values, are used as training examples to statistical learning methods such as partial least-squares,[5,13] or machine learning methods such as rough set-based rule learning,[14] to induce predictive models. It has been shown that the PCM approach can yield models that predict the binding affinity of a series of ligands to various targets with an accuracy comparable to that of QSAR modeling.[15] Both traditional QSAR descriptors and 3D-structure-based GRIND descriptors have been used to describe ligands in PCM.[15,16] As protein descriptors both simple binary attributes representing the occurrence of specific amino acids[5] and z-scales[17] representing principal components of amino-acid properties have been used in PCM.[13,18] Both approaches require proteins to be related so that a multiple sequence alignment can be constructed and descriptors computed for specific positions in that alignment.

A quantitative approach very similar to PCM has been applied by Lindström et al. to predict and separate low- and high-binding affinity complexes.[19] That study used protein descriptors that were calculated from the active site such as the binding site surface area. This limits the prediction range of the model to those cases where there is some knowledge of the ligand-binding site of the protein. In another quantitative approach, the binding affinities of zinc-containing metalloprotein ligand complexes were predicted.[20] The model was based on traditional QSAR ligand descriptors and the properties of the amino acids in the binding site. A virtual screening application with the aim to find ligands to orphan G-protein coupled receptors (GPCR) was applied by Bock and Gauge.[21] Various sequence properties were used as receptor descriptors, and the ligands were described by a connectivity matrix. The model was able to predict binding affinities as shown by cross-validation on the training data set with a coefficient of determination (r^2) of 0.43 and a Root-Mean-Square Error of Prediction (RMSEP) of 1.2. The coefficient of determination measures the proportion of variability in the binding affinity data accounted for by the model and is computed as the squared correlation coefficient between predicted and experimentally measured binding affinity values. A set of ligands was matched with orphan GPCR receptors, and a number of high-affinity interactions were predicted. However, the results are awaiting experimental validation.

Here we introduce a general PCM modeling approach in which proteins are represented using local descriptors of protein structure. A local descriptor is a discrete structural entity encompassing the complete local neighborhood around an amino acid.[22] A library of commonly occurring local descriptors has been created, describing the binding pockets of protein−ligand complexes regardless of sequence similarity or global structural similarity. The local descriptors in this library can be matched to previously unseen protein structures without knowledge of the binding site of those proteins. In principle, this means that models can be generated that span the entire enzyme-ligand space, containing proteins that vary greatly in terms of sequence, structure, and function. A recent pilot study showed that the PCM approach incorporating local descriptors can yield ligand-binding models that generalize over a small set of hydrolase and lyase enzymes.[23] In the present study we employed a comprehensive training and test set consisting of enzymes from all the major enzyme classes (EC)[24] to investigate if this PCM approach generalizes over the entire structural enzyme-space. The test set does not include information on the binding site of the ligands and was manually curated to obtain a mapping between experimental binding affinity values and protein sequence (Uniprot accession codes). To our knowledge, this mapping is not available in any public enzyme-ligand database, and both data sets are therefore made available as Supporting Information (S1 and S2).

The PCM model was induced using support vector machine (SVM) regression and yielded an r^2 of 0.53 ($p <$ 1.4E-85) and a RMSEP error of 1.5 on the diverse external test set. The generality of the model is further demonstrated by its superiority over a nearest neighbor (NN) predictor. We also show that in many cases, the model can correctly rank different ligands that bind to the same enzyme. The model does not require sequence alignment or prior knowledge of the ligand binding site location and can be used for rapid prediction of binding affinities of new protein−ligand combinations provided that the protein structure is known. To our knowledge, an interaction model of this generality and scope has not been reported previously.

MATERIALS AND METHODS

Selection of Druglike Ligands. This study only considered druglike ligands. They were selected using a combination of the Lipinski rules[25] and the criteria of druglikeness defined by Boström et al.[26] The ligands included in the training and test sets had the following properties: (a) 80 < molecular weight (Da) < 750, (b) calculated lipophilicity (log P) less than 5, (c) fewer than 5 H-bond donors, and (d) fewer than 10 H-bond acceptors.

Collection of Training Set Data. We collected structural and binding affinity information on all available enzymes stored in the following databases: AffinDB,[27] PDB Bind,[28] Binding MOAD,[29] and Protein Ligand Database[30] as of March 30, 2007. Only enzyme-ligand pairs for which a binding affinity constant was available were included in the training set. This resulted in 3531 triplets, where each triplet consisted of a crystallized enzyme, a cocrystallized ligand,

and a binding affinity constant expressed as the negative logarithm of the inhibition constant K_i or the dissociation constant K_d. There was a significant overlap between the databases and inhibition constants for the same complex were occasionally reported from multiple literature sources. To obtain one single inhibition measure for each unique enzyme-ligand pair, the mean inhibition value was computed for every set of identical pairs in the data set. The training data set was thus reduced to 1421 triplets. Protein−ligand crystal structures were obtained from the Protein Data Bank (PDB),[31] and their idealized ligand structures were downloaded as coordinate files from the MSDchem[32] database. Triplets containing a nondruglike ligand were removed from the data set. The final training data set was thus composed of 826 triplets (Supporting Information S1).

Collection and Manual Curation of Brenda Test Set Data. All available enzyme-ligand binding affinity data were obtained from the Brenda database[33] (as of January 1, 2007). The collection of data contained 15002 entries, each consisting of a binding affinity value linked to a ligand, an enzyme classification (EC) number, an organism, and (in two-thirds of the cases) a PubMed identification number (PMID). To select entries associated with enzymes of known 3D structure, each EC number in the original data set was checked against the PDBsum EC→PDB mapping database[34] for the existence of crystal structures. This resulted in 4948 entries that had a binding affinity value linked to a PMID, a ligand structure, and an EC number that is associated with one or more 3D protein structures. Since EC numbers in many cases cover proteins that differ both in sequence and structure, each entry had to be curated manually by reading the associated primary literature. The enzyme/isozyme name, the EC number, and the species name were used as input to the Biothesaurus database[35] in order to find the amino-acid sequence of each isozyme/enzyme. Enzymes for which it was not possible to find a unique protein sequence were removed from the data set. Due to the limited online availability of older publications, the curation was limited to entries published from 1995 to 2006. With these constraints, the final data set consisted of 1621 entries for which the primary structure of the enzyme, a 3D structure of the ligand (downloaded from Brenda), and the K_i value (extracted from Brenda) of the protein−ligand interaction were available. Each enzyme sequence was searched against the PDB[31] to identify structures of proteins with more than 95% sequence identity. The retrieved protein structures for each entry were ranked by protein structure quality using information obtained from ProCheck.[36] A simple empirical quality measure, q, was defined as follows

$$q = \frac{f + g - d + 10a + 10p + 10c}{r} \quad (1)$$

where f is the percentage of residues in the most favored regions of the Ramachandran plot, g is the percentage in the generously allowed regions, d is the percentage in disallowed regions, a is the overall average G-factor,[37] p is the G-factor of the φ/ψ distribution, c is the G-factor of the χ_1-χ_2 distribution, and r is the resolution of the crystal structure in Å. The highest scoring crystal structure was selected as the enzyme representative for computation of local descriptors. In this study, only "druglike" compounds were selected. This resulted in 542 entries that were used as a test set.

Ligand Descriptors. The coordinate files for the training set ligands were downloaded from the MSDchem database,[32] and the test set ligands were taken from the Brenda database.[33] The 3D structures of the Brenda ligands and the MSDchem ligands were low-energy conformers computed with Corina.[38] The training and test set ligand descriptors were computed with Dragon 2.1.[39] Dragon allows the calculation of 0D-3D QSAR descriptors. 0D descriptors are related to overall properties such as molecular weight, 1D descriptors include local properties such as fragment counts and number of functional groups, 2D descriptors are mostly related to topology and molecular walk counts, and 3D descriptors are geometrical descriptors and include properties such as radius of gyration and 3D Wiener index.[10] This resulted in 1481 0D-3D ligand descriptors. Of these, only 27 descriptors were used as input to the model (Supporting Information S3). These descriptors were selected based on their interpretability and their ability to describe physiochemical properties important for putative drugs.

Local Descriptors of Protein Structure. A local descriptor of a protein 3D structure is a collection of continuous short backbone fragments centered around a particular amino acid.[22,40] It is derived in a series of steps: (a) The central amino acid of the local descriptor is represented by the point in space on the vector $[C_\alpha, C_\beta]$ that lies 2.5 Å away from C_α. (b) All amino acids within a radius of 6.5 Å of the central amino acid are identified. For each such amino acid, four sequence neighbors, two on each side, are added to obtain a continuous backbone fragment of five amino acids. (c) All overlapping fragments are merged resulting in a set of nonoverlapping backbone fragments. A local descriptor is named, e.g., 1coma#94; the local neighborhood in PDB entry 1com, chain A, centered on residue number 94.

A database of local descriptors had previously been constructed based on all structures in ASTRAL version 1.63.[41] This is a representative subset of the PDB that contains domains with less than 40% mutual sequence identity. To build a library of binding-pocket descriptors for this study, we matched the structure of all local descriptors from enzymes in our training set to the ASTRAL-derived similarity database. Two local descriptors are considered to be similar if a number of criteria are met including an overall rmsd (root-mean-square distance) threshold.[22] This resulted in a large number of similarity groups with ASTRAL-based local descriptors as centers. We then removed groups that did not contain at least five different enzymes in the training set or that did not describe amino acids in contact with the ligand for at least five different enzymes. A nonredundant library of 405 binding pocket descriptor groups was finally constructed by iteratively selecting the group describing the largest number of contacts not already described by previously selected groups. A contact was defined if two atoms, one from the amino acid and one from the ligand, were within 5 Å of each other. The enzyme structures in the test set were matched to the final local descriptor library and were therefore not used in the process described above.

Enzyme Class Diversity. The EC number of each entry in the training and test set, obtained from the PDB, was used as a basis for obtaining an estimate of the EC coverage in the data sets. The EC distribution in the PDB was obtained from the EC→PDB mapping database.[34] The percentage of

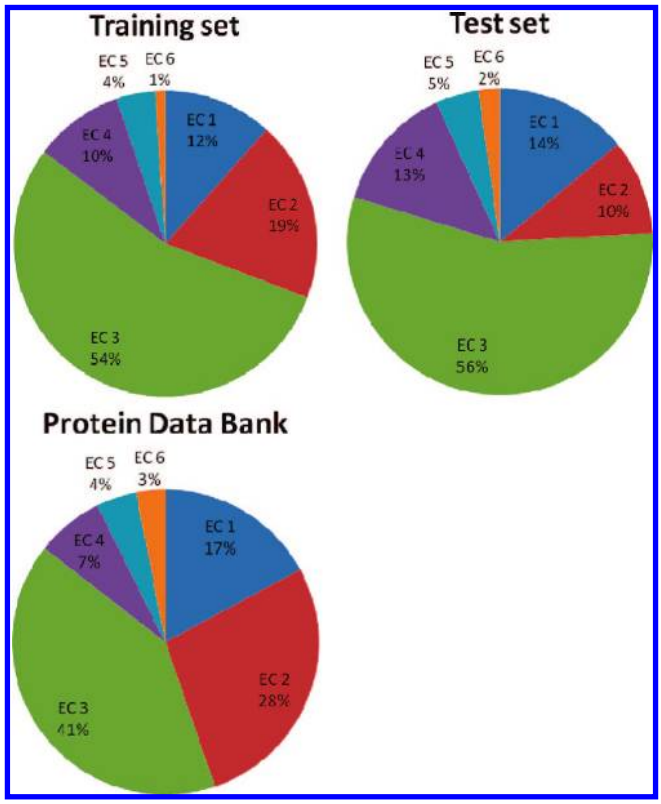


Figure 1. The enzyme class (EC) coverage of the training and test sets. All six major ECs are represented in both sets. For comparison, the EC coverage of the PDB is also shown.

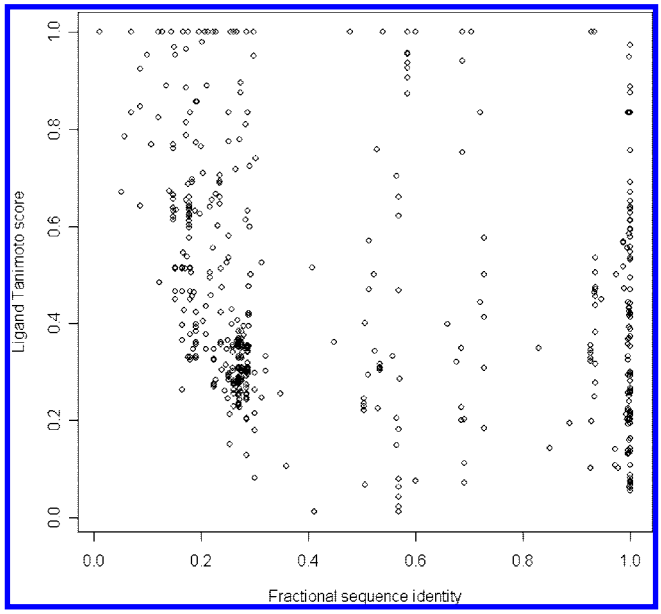


Figure 2. For each complex in the test set, the most similar complex in the training set was identified. The plot shows the ligand Tanimoto scores and the protein sequence identities of each such pair of complexes. The plot shows that the test set contains many complexes that have a very low similarity to any of the training set complexes (see text for details).

entries at the highest level of the EC tree (e.g., class 1.-.-.-) in the training set, the test set, and the PDB are shown in Figure 1.

Enzyme-Ligand Complex Similarity Calculations and Clustering. Enzyme similarity was calculated by aligning sequences with the Needleman-Wunsch algorithm,[42] as implemented in the European Molecular Biology Open Software Suite (EMBOSS,[43] program "needle"). Ligand similarity was calculated as a fingerprint-based Tanimoto coefficient with OpenBabel.[44] Test set complexes that displayed both sequence identity of at least 95% and ligand similarity of at least 0.95 to any entry in the training set were inspected manually. This was done by inspection of the sequence alignment and by comparison of the ligand structures.

A similarity score s can be calculated between two enzyme-ligand complexes by combining the sequence identity and ligand similarity scores

$$s = \frac{1}{1+\alpha}\left(\alpha t + \frac{p}{100}\right) \quad (2)$$

where t is the Tanimoto coefficient of the ligands, p is the percentage sequence identity, and α is a weight factor (in this study $\alpha = 1$).

The enzyme-ligand similarity score s (eq 2) was computed for all possible pairs of enzyme-ligand complexes between the training set and the test set. The computed $1-s$ scores were compiled into a distance matrix that was used as an input to the Ward hierarchical clustering method, available from the R suite of statistics software.[45] The Ward clustering algorithm defines the distance between clusters as the increase in the error sum of squares (*ESS*). The value of *ESS* is given by

$$ESS = \sum_{j=1}^{N} (x_j - \bar{x}_j)'(x_j - \bar{x}_j) \quad (3)$$

where x_j is the multivariate measurement associated with the j^{th} complex, and $\bar{x}_j$ is the centroid of the cluster of which x_j is a member. At each step of the clustering analysis, a union of every possible pair of clusters is considered, and the two clusters whose union results in the smallest increase in *ESS* are joined.[46] The results are shown in Figure 3.

Model Induction by Support Vector Machines. We used the support vector regression method as implemented in the LIBSVM software[47] to induce models. Support vector machines have a number of advantages over the standard linear regression analysis used in most interaction studies, in particular the robustness to overfitting and large number of features as well as the possibility to use nonlinear kernels.[48] We used a radial basis kernel function after experimenting with different kernels on the training set.[47] Parameters were fitted by a grid search followed by hill-climbing using 2-fold cross-validation within the external 10-fold cross-validation or on the entire training set for the final model.[47]

RESULTS AND DISCUSSION

An Enzyme-Wide Interaction Data Set. To properly validate our approach to generalized PCM modeling, we created one training set and one test set that both encompassed enzymes from all major EC enzyme classes (EC1-EC6). The training set was compiled from the databases AffinDB,[27] PDB Bind,[28] Binding MOAD,[29] and Protein Ligand Database[30] (see Materials and Methods). Binding affinity data expressed as inhibition (K_i) or dissociation constants (K_d) was obtained from these databases. Ligands that were not druglike according to a slightly modified

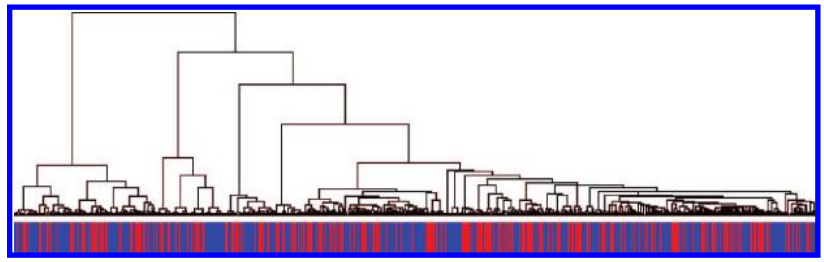


Figure 3. A hierarchical cluster dendrogram of the training (blue) and test set (red) protein−ligand similarity scores (see text for details). The blue-red bar-code-like pattern shows that the complexes in the test set are representative of the complexes in the training set.

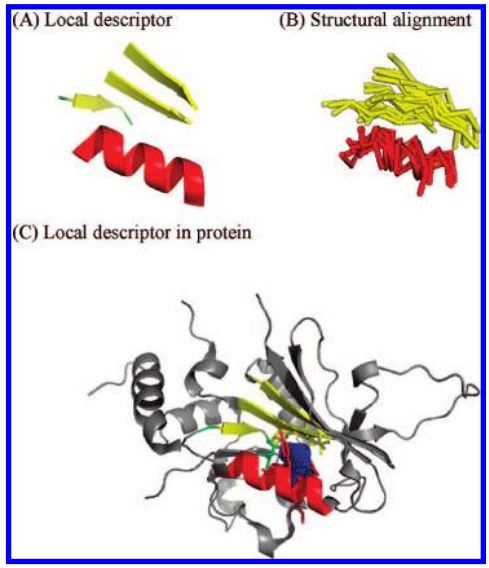


Figure 4. (A) The local descriptor centered on amino acid 14 of chain A in *Plasmodium falciparum* dihydrofolate reductase (PDB code 1j3j; the descriptor is labeled 1j3ja#14). (B) Structural alignment of a set of local descriptors similar to 1j3ja#14 in the PDB. (C) The local descriptor 1j3ja#14 contains 9 out of 16 amino acids in chain A of PDB entry 1j3j that are in contact with the ligand 5-(4-chlorophenyl)-6-ethylpyrimidine-2,4-diamine. The descriptor is colored according to its secondary structure. Contacting residues are drawn as sticks. The ligand is in blue.

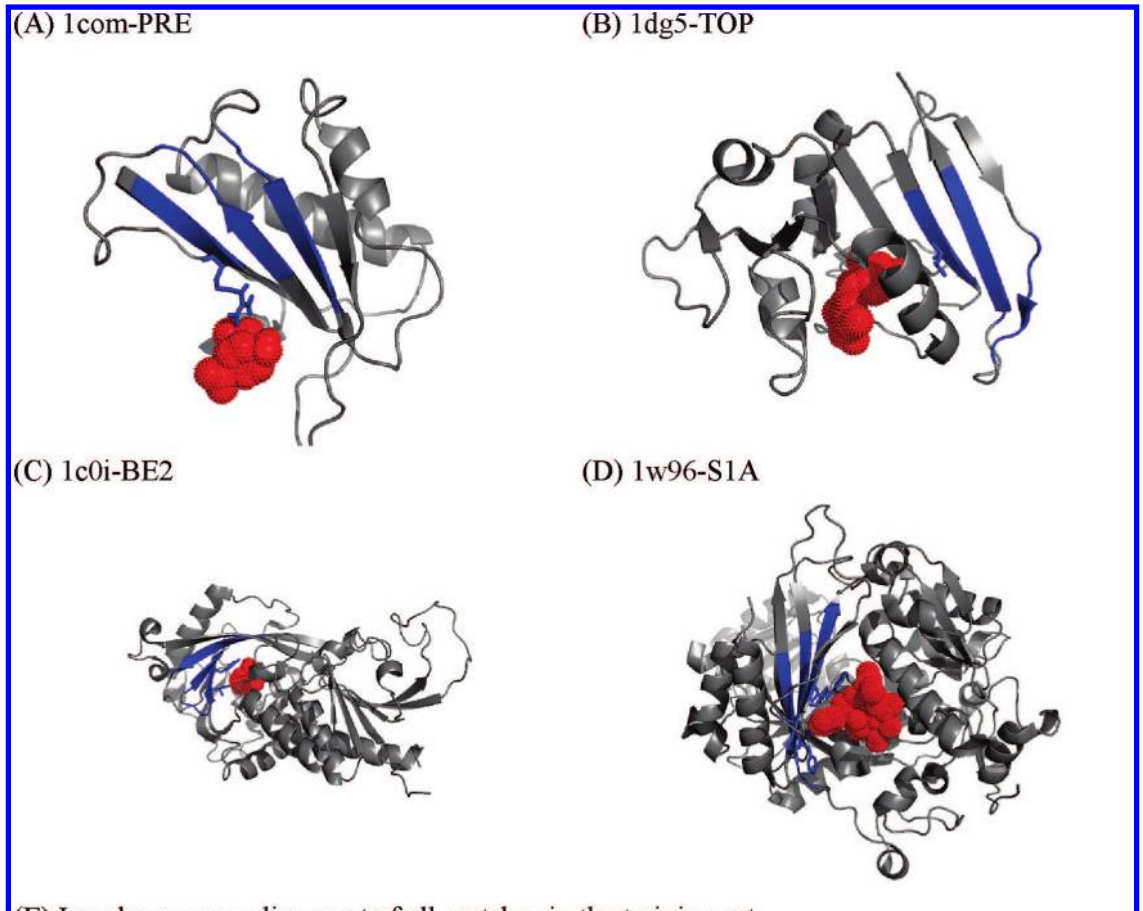
version of the Lipinski rules were removed (see Materials and Methods). This resulted in a training set that consisted of 826 enzyme structures cocrystallized with their ligands.

The test set was compiled from the Brenda enzyme database[33] and curated manually. The Brenda database is one of the most comprehensive sources of enzyme data and contains more than 15000 K_i values linked to primary literature references, EC numbers, and ligands. For this study, it was necessary to link K_i values to specific enzyme crystal structures. Therefore, the references for K_i values published from 1995 to 2006 were amassed which resulted in 1621 entries each comprising one K_i value associated with an amino acid sequence and a ligand (see Materials and Methods). To obtain a mapping between binding affinity values and enzyme crystal structures, each amino acid sequence was searched against the Protein Data Bank (PDB).[31] The test set eventually contained 542 entries that consisted of a K_i value associated with one enzyme structure and an associated but not cocrystallized ligand. The binding sites of the ligands in the test set were thus unknown. Both data sets are available as Supporting Information (S1 and S2).

Figure 1 shows the enzyme class coverages of the training and test set and confirms that they are similar to that of the entire PDB. For instance, the largest group in both the training and test set is the hydrolase class (3.-.-.-) which is also the largest class in the PDB. This indicates that the data in the training and test sets are representative subsets of the available enzyme crystal structure data and that both data sets include information from all major enzyme classes.

To assess similarities between the training set and the test set, all test set complexes were compared with all training set complexes by computing the sequence identities of the enzymes and a fingerprint-based Tanimoto score for the ligands (see Materials and Methods). The fractional protein sequence identity and the Tanimoto score were then averaged to obtain an *enzyme-ligand similarity score*. This score ranges between 0 and 1, where a value of 1 indicates that the two complexes are identical. For each complex in the test set, we found the closest complex in the training set using this enzyme-ligand similarity score and plotted the corresponding ligand similarity against the protein sequence identity. Figure 2 shows that the test set contains a large number of target complexes that are potentially difficult to predict. 60% of the test complexes contain an enzyme with less than 30% sequence identity to the closest training complex, and 64% of the ligands has a Tanimoto score of less than 0.5. 36% of the test set complexes are located in the lower left quadrant of the plot and have both less than 50% sequence identity and a Tanimoto score of less than 0.5 to the closest training complex. Only 13% of the test set is located in the corresponding upper right quadrant. A relatively small number of complexes were deemed to be very similar to complexes in the training set. All pairs that had both a sequence identity of at least 95% and a ligand Tanimoto score of at least 0.95 were inspected manually. This resulted in the removal of 17 complexes from the test set. To assess how representative the test set is for the training set, we constructed a matrix of all pairwise similarity scores in both the training and the test set. The complexes were then clustered by hierarchical clustering (Figure 3; see Materials and Methods). The bar-code-like pattern strongly indicates that the enzyme-ligand complexes in the test set are representative of the training set.

Local Descriptors Describe Binding Pockets. Local descriptors of protein structures describe the entire neighborhood of amino acids around one specific residue (Figure 4A).[22] The initial neighborhood is extended to a set of nonoverlapping continuous backbone fragments at least five amino acids long. Only neighborhoods with at least three nonoverlapping fragments are considered. We used this concept to build a library of local descriptors that describe amino acids in contact with ligands in the training set (Figure 4B,C, details in Materials and Methods). This library thus contains stable, local conformations of protein structure that commonly occur in and around the binding pockets of



(E) Local sequence alignment of all matches in the training set

Descriptor	Segment1		Segment2		Segment3		Lig	SCOP fold/EC family
1coma_#94	3-7	..IRGIH	42-48	VVQMLLS	91-96	VMMTVQ	PRE	Bacillus chorismate mutase-like/5.4.99.5
1dg5a_#153	111-115	..EVTEV	141-145	..EWRFS	150-155	RYRLYS	TOP	Dihydrofolate reductases/1.5.1.3
1dy4a_#417	88-94	GNSLSIG	130-134	DVDVS..	414-419	AKVTFS		Concanavalin A-like lectins/glucanases/3.2.1.91
1c0ia2#1233	1199-1205	GQTVLVK	1223-1229	YITPRPG	1231-1235	.EVICG	BE2	FAD-linked reductases, C-terminal domain/1.4.3.3; 1.4.3.4; 1.14.13.2; 1.5.3.1;
1c0la2#1233	1199-1205	GQTVLVK	1223-1229	YITPRPG	1231-1235	.EVICG	FLA	
1el5a2#267	223-227	..QVVGF	252-257	IYYGFP.	264-269	LKLGYH	DMG	
1el9a2#267	223-227	..QVVGF	252-257	IYYGFP.	264-269	LKLGYH	MTG	
1elia2#267	223-227	..QVVGF	252-257	IYYGFP.	264-269	LKLGYH	PYC	
1ojaa2#343	292-298	GSVIKCI	326-330	..YTLDD	340-345	IMGFIL	ISN	
1phh_2#222	183-189	FGWLGLL	208-214	FALCSQR	219-224	SHYYVQ	DHB	
1qwua2#932	524-528	..FTLDD	554-560	HVVMHNT	929-934	LDKFIF		Supersandwich/3.2.1.114
1w96a1#508	454-460	CTACRIT	487-491	WGYFS..	505-510	QFGHIF	S1A	Barrel-sandwich hybrid/6.4.1.2
1w96a1#510	452-458	GHCTACR	487-491	..WGYFS	507-512	GHIFAF	S1A	
2a06e#194	71-77	MSKIEIK	184-190	SYEFTSD	192-196	.MVIVG		-/1.10.2.2
2web__#318	177-183	GSLTYTG	307-311	YVVFD..	315-320	PQLGFA		Acid proteases/3.4.23.20
2wec__#318	177-183	GSLTYTG	307-311	YVVFD..	315-320	PQLGFA		
2wed__#318	177-183	GSLTYTG	307-311	YVVFD..	315-320	PQLGFA		

Figure 5. The local descriptor 1dgw.1#Y367 (the local descriptor centered on amino acid Y367 in SCOP domain 1dgw.1) has 16 matches in the training set. (A-D) Matches to four different enzyme-ligand complexes with enzymes from four different SCOP folds. The local descriptor is shown in blue, and residues that contact the ligand as sticks and ligands (ligand code in PDB: PRE, TOP, BE2, S1A) are shown in red. (E) The local sequence alignment resulting from the structural alignment of all matches to 1dgw.1#Y367 in the training set. Contacts are marked in red.

proteins. Figure 5 shows an example of a local descriptor that describes a part of the binding pocket of several very different protein structures in the training set. Overall, the 405 local descriptors in our library describe 75% of all amino acids that are in contact with the ligands in the training set (Supporting Information S4). The remaining uncovered contacts are mainly associated with descriptors that are so rare that they were filtered from the library (occurring in fewer than 5 different structures). Only 8% of the contacts did not match any local descriptor and were thus located in

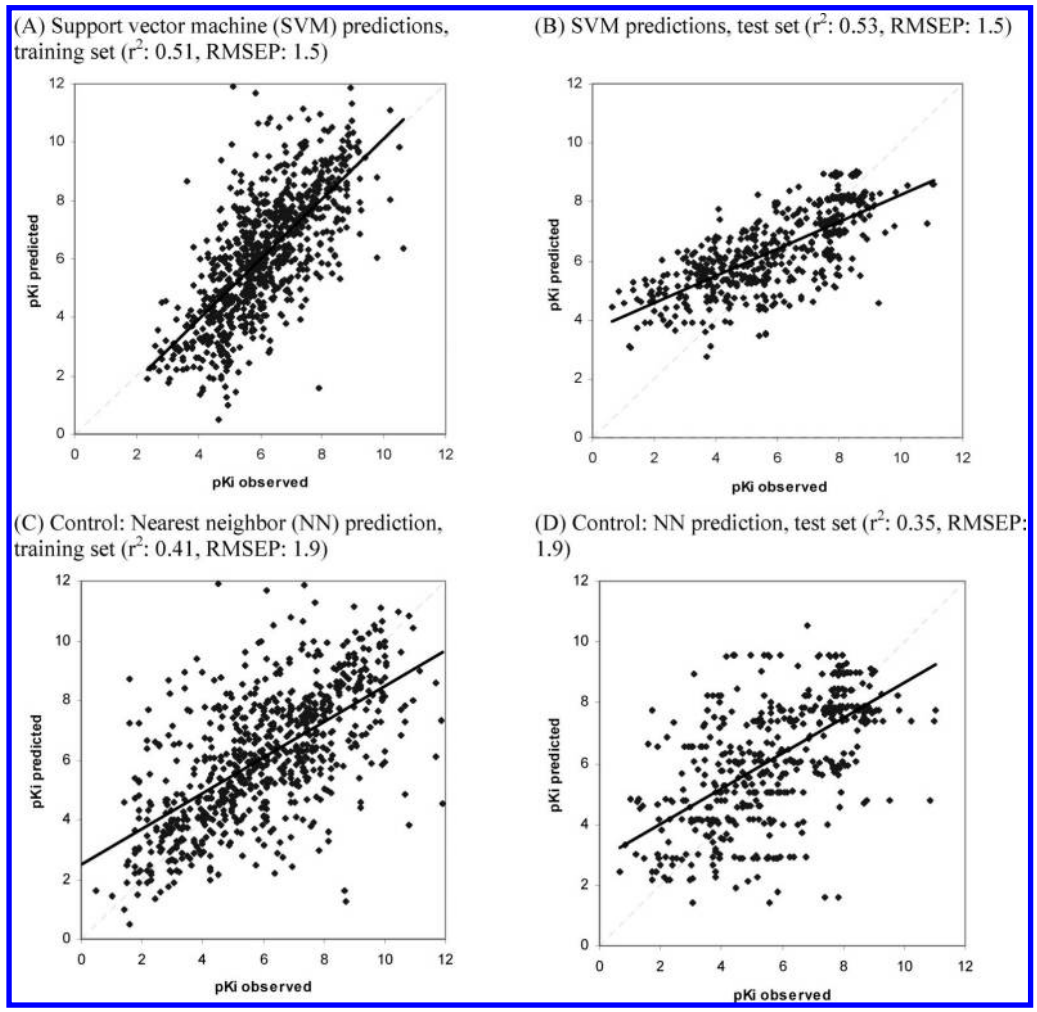


Figure 6. (A) Accumulated results from validating the modeling approach on the training set in a 10-fold cross-validation setting. (B) Results from validating the model on the external test set with unknown ligand binding mode. (C,D) Same as A and B, but a simple nearest neighbor predictor is used instead of SVM. The solid lines are linear trend lines fitted using the least-squares method.

parts of the protein that were not dense enough to contain three independent backbone segments close to each other. Local descriptors are large and specific enough to describe conserved structural motifs (e.g., parts of binding pockets) but at the same time small enough to describe common properties of otherwise dissimilar structure (see Figure 5). Another advantage is that they can be matched to new protein structures without requiring knowledge of the position of the binding pocket in those proteins. Thus, they constitute a very appealing framework for inducing generalized interaction models.

Generalized Model of Enzyme-Ligand Interactions. We represented each complex using local descriptors to describe proteins and QSAR descriptors to describe ligands (details in Materials and Methods). Each of the 405 local descriptors in our library is either present or absent in a protein structure. Thus, each protein can be represented by a vector of 405 bits (one for each local descriptor), and each local descriptor can be represented by a vector of 811 bits (one for each complex in the training set). However, this representation only takes into account the 3D structure of the local descriptor and does not account for its amino acid content. We decided to expand this representation to include every position in a local descriptor that has at least one contact with the ligand. For example, 13 positions in Figure 5E are in contact with the ligand in at least one complex, the third position in the second segment being the position most frequently recognized as a contact. Each such position was represented by five Z-scales reflecting various amino acid properties (these scales are principal components of a large number of different amino acid properties[17]). This resulted in 4555 contacting positions and thus five times as many Z-scales.

We employed the support vector regression method to induce the predictive model[47] using a nonlinear kernel function (details in Materials and Methods). Figure 6A,B shows the relationship between observed and predicted binding affinity for the training set (10-fold cross-validation) and the external test set (model built on all training data). The results for the external test set (r^2 0.53, p-value 1.4E-85 using the standard *t* test) are similar to those obtained for the training set (r^2 0.51, p-value 1.1E-125) and show that the model, and consequently also the local descriptor library, generalizes to unseen protein−ligand complexes. In fact, no correlation was found between the model's ability to predict binding affinity for complexes in the test set and the similarity of the corresponding test complex to its closest neighbor in the training set (Figure 7). This shows that the model is based on general properties related to interaction and that its ability to predict does not degrade when it faces completely new complexes. The generality of the model is further confirmed by its superior performance over the simple nearest neighbor

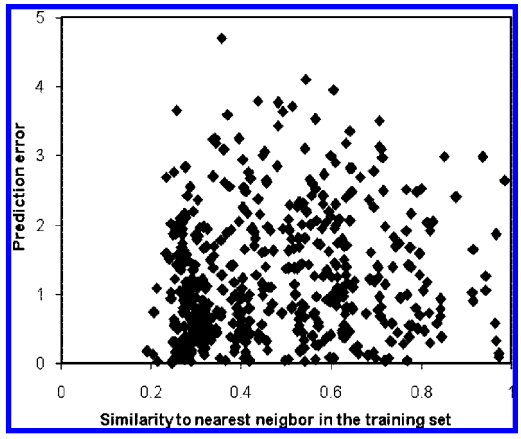


Figure 7. Prediction error (absolute difference between observed and predicted pK_i value) for complexes in the test set plotted against the similarity to the nearest complex in the training set (see text for details).

approach where the binding affinity value of the most similar complex in the training set is used as a prediction (Figure 6C,D). As before, similarity is calculated as the average between the fractional sequence identity of the proteins and the Tanimoto score of the ligands.

In virtual screening, one is generally faced with the problem of ranking a large number of ligands according to how strongly they bind to one specific protein (or *vice versa*). This is particularly challenging for the generalized model described in this paper, because it is trained on data that contain relatively few series, that is, a number of similar ligands that bind to the same protein or *vice versa*. On the other hand, our results suggest that the current model does not require a large number of ligands for each protein, which is in stark contrast to QSAR approaches. Figure 8 shows that the model is able to correctly rank affinity order for a large number of complexes in the test set. Approximately three out of four complexes containing the same protein (but different ligands), and about nine out of ten complexes containing the same ligand (but different proteins), are correctly ranked when the test set data are limited to the series in which the observed difference in pK_i values is greater than 2 (Figure 8A). This is not an unreasonable limitation considering that the data amassed from four different databases contained affinity measures from inhibition assays performed under variable conditions. Although pK_i values in general are comparable between assays, 16% of the enzyme-ligand complexes in the data had a standard deviation above 0.1. Moreover, the average standard deviation for these complexes was as high as 1.9. Obviously, it would be useful to know to what degree one can trust a ranking when the true binding affinity values are unknown. Figure 8B shows that the model displays similar ranking performance as above when one requires that the difference in *predicte*d pK_i values be greater than 1 (Figure 8B). In other words, given two complexes with the same protein, but different ligands, and a predicted difference in binding affinity of at least 1, the probability that their ranking is correct is 0.73. The corresponding probability for complexes with different proteins, but the same ligand, is 0.91. Figure 8A,B also shows that the nearest neighbor approach is unable to rank complexes; in fact it performs hardly better than random guessing.

Case Study: *Zea mays* Polyamine Oxidase. Polyamine oxidase (PAO) is involved in polyamine metabolism and production of hydrogen peroxide in animals and plants and plays a key role in development and programmed cell death. Until recently, maize PAO was the only PAO for which the tertiary structure had been determined, and therefore it is an important model for the study of the catalytic mechanism and for the design of new inhibitors for both plant and animal enzymes.[49] The test set contained seven ligands interacting with maize PAO (Figure 9A,B). The model generated predictions with a ranking of 0.90, an r^2 of 0.76, and an RMSEP of 1.7 (Figure 9E). Although the ranking is almost correct, Figure 9E clearly shows that only a small part of the observed variation in this series is predicted. This illustrates some of the above-mentioned difficulties of predicting series using a model trained on data that does not contain many series. PAO has only 36% sequence identity to its closest neighbor in the training set (human monoamine oxidase B; PDB code 2bk3). Figure 9A shows the parts of the protein structure that are covered by local descriptors from our library. It is interesting that the model still performs fairly well considering that the binding pocket is in a coil-rich region in the interface between two domains and is only partly described by the local descriptors.

Case Study: *Plasmodium falciparum* Dihydroorotate Dehydrogenase. Dehydroorotate dehydrogenase (DHODH) catalyzes the oxidation of dihydroorotate to orotate utilizing

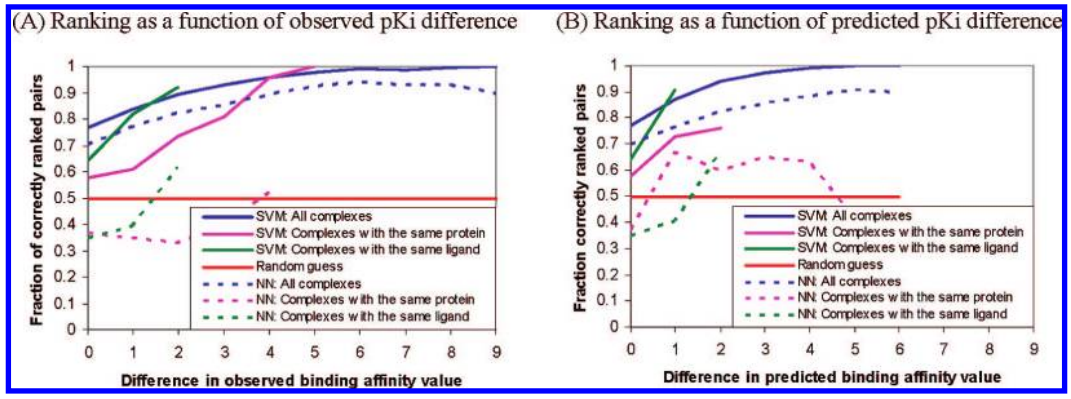


Figure 8. The model's ability to correctly rank pairs of protein−ligand complexes in the test set by binding affinity values. The *y*-axis measures the fraction of pairs that were correctly ranked, that is, where the complex with the highest observed binding affinity also had the highest predicted binding affinity. The *x*-axis divides the data into subsets of pairs where the difference in observed (A) or predicted (B) binding affinity values for two complexes in a pair is greater than 0, 1, 2, etc. Curves are drawn for all pairs of complexes as well as for pairs of complexes where the protein or the ligand is the same.

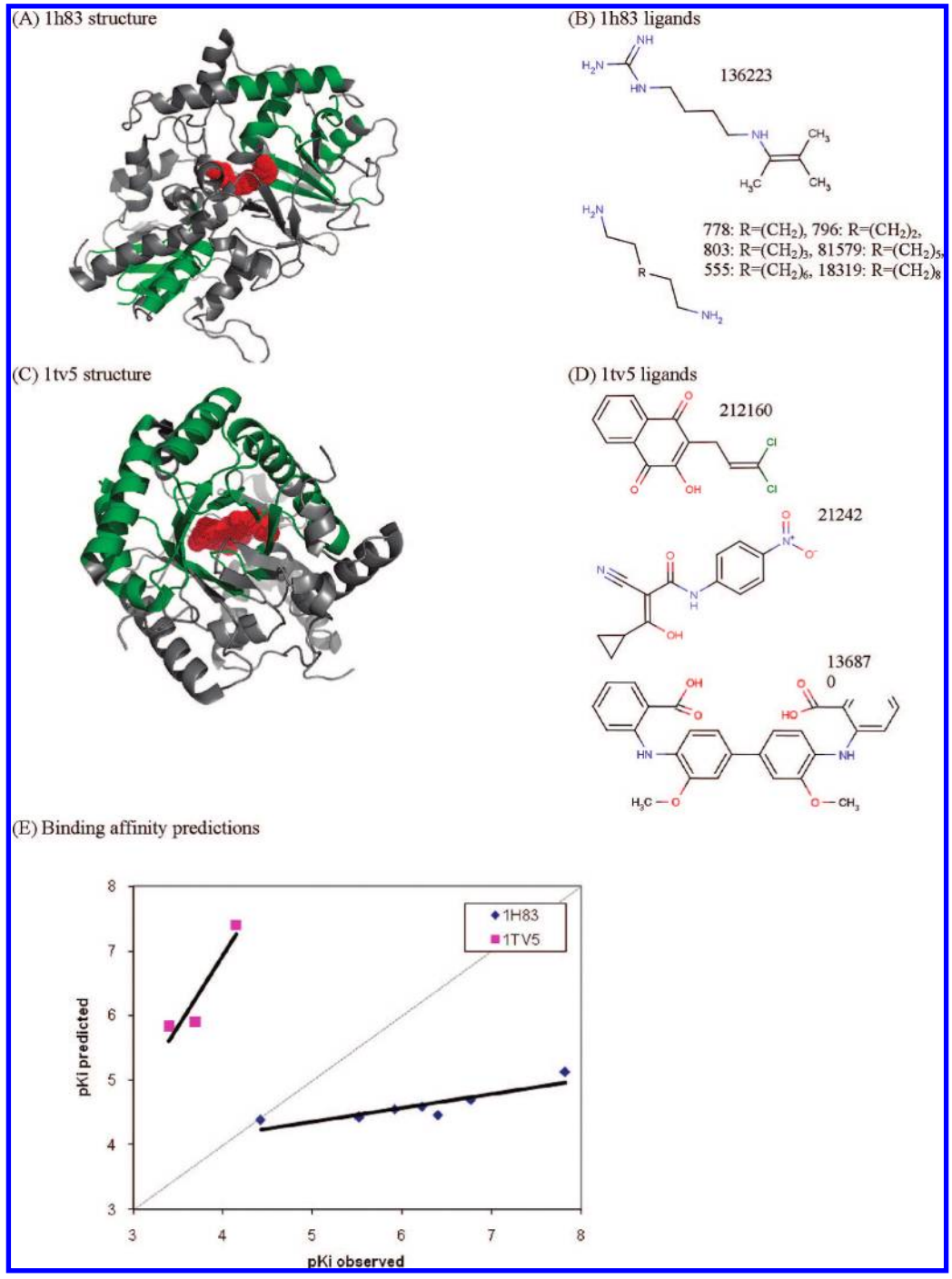


Figure 9. (A) The crystal structure of polyamine oxidase from *Zea mays* (PDB entry 1h83). The cocrystallized ligand octane 1,8-diamine is colored in red. The local descriptors matching the enzyme are highlighted in green. (B) Ligand structure 136223, 778, 769, 803, 81579, 555, and 18319 from the Brenda database mapped to interact with 1h83. (C) *Plasmodium falciparum* dihydroorotate dehydrogenase (PDB entry 1tv5) and the cofactor flavin nucleotide (FAD). (D) Ligand 212160, 21242, and 13687 mapped to interact with 1tv5. (E) Predicted versus observed binding affinity for the two series of ligands in the test set. Solid lines are linear trend lines.

the flavin cofactor FMN, which is a key step in pyrimide biosynthesis. Inhibitor studies of *Plasmodium falciparum* DHODH have shown that the enzyme is a promising target for the development of new antimalarial drugs.[50,51] The test set contained three ligands mapped to interact with DHODH (Figure 9C,D). The protein substructures in our library describe the part of the protein structure that surrounds the FMN binding site. The three ligands are ranked correctly with an r^2 of 0.87 and a RMSEP of 2.7 (Figure 9E). In this case, the correct ranking is accompanied by large overprediction which results in a high RMSEP. The protein displayed low similarity to the training set with PDB entry 1jqv (*Lactococcus lactis* dihydroorotate dehydrogenase A) as the most similar protein with a sequence identity of 32%.

Model Quality and Practical Application. The overall prediction accuracy presented in this study ($r^2 = 0.53$ and RMSEP = 1.5 for the test set) is not on par with what is typically reported for QSAR models. However, this is to be expected considering the diversity of the data both in terms of proteins/ligands and in terms of experimental techniques and laboratories producing the experimental binding affinity values. The practical application of the current model is

probably limited to providing rough predictions in order to, for instance, separate low binding affinity complexes from high binding affinity complexes as shown in Figure 8. The case studies in particular reveal limitations in predicting series. Since the training data mostly consists of unique complexes, series in the test set often result in predictions with limited variation. Still, this model clearly outperforms a naïve nearest neighbor approach both in terms of r^2 estimates and in terms of ranking complexes in series. Our results also compare favorably to other published studies where diverse data sets are used. Lindström et al.[19] reported an r^2 of 0.25 and a RMSEP of 1.92 on an external test set in their comprehensive protein−ligand study. Bock and Gough[21] reported a cross-validation r^2 of 0.43 and an RMSEP of 1.2 in their G-protein coupled receptor study. Hence, our results are an important step toward building general quantitative models with practical applications in areas such as *in silico* drug screening and prediction of adverse side effects.

CONCLUDING REMARKS

To obtain full coverage of the protein−ligand space, it would be necessary to investigate the binding of the chemical universe against entire proteomes. Such an investigation would yield a high resolution map of the space enabling, for instance, an accurate prediction of cross-interactions of putative drugs with other proteins in the proteome. However, this would be practically impossible, due to the almost infinite number of potential druglike compounds. Therefore, different strategies to travel the ligand and protein space are needed. One strategy is to focus on cross interactions with other proteins of the same protein family. This has been applied successfully by Martin et al.[52] and Guba et al.[53] where ligands with high specificity to their targets were discovered by comparing and screening GPCRs within the same family. Another approach is to coarsely sample the ligand and protein space by a diverse set of complexes representative of the universe and thus obtain a rough roadmap that could help navigate the protein−ligand space. This requires means to compare complexes as the proteins and ligands vary in size, structure, and sequence. This study shows that it is possible to induce a predictive model on a diverse data set that includes enzymes from all enzyme classes. Although local descriptors require that the protein structure is known, it is reasonable to expect that the applicability of these stable local conformations will survive the transition from experimentally determined protein structures to homology models. This would increase the scope of this approach enormously and will be pursued in future research. We have previously shown that local descriptors can be correctly assigned to sequences using position-specific scoring matrices, even in the case of nonsignificant sequence similarity to the training set of known structures.[22] Thus, we expect that our approach can be used and extended to obtain generalized models that allow predictions of protein−ligand cross interactions across proteomes.

ACKNOWLEDGMENT

We would like to thank Marcin Kierczak, Robin Andersson, Stefan Enroth, Adam Ameur, and Jakub Orzechowski Westholm at the Linnaeus Centre for Bioinformatics for many thoughtful and fruitful discussions on computational methodology and Mikael Nilsson at the Department of Cell and Molecular Biology for answering numerous questions on enzyme kinetics. We would also like to thank Emil Lundberg at the IT support department of the Uppsala Biomedical Centre for technical assistance with the Brenda database. Computations were partly carried out at the Polish CoE BioExploratorium Computing Centre financed by the following projects: WKP_1/1.4.3/1/2004/44/44/115/2005 and WKP_1/1.4.3/2/2005/74/193/385/2006. Financial support was obtained from the Swedish Research Council (JESW: grant number VR 04X-05957), the Knut and Alice Wallenberg Foundation, the Swedish Governmental Agency for Innovation Systems (VINNOVA) and the Polish Ministry of Science and Higher Education (PD: grant PBZ-MIN-014/P05/2004). Finally, we would like to thank the anonymous referee whose comments led to a substantial improvement in this manuscript.

Supporting Information Available: Training and test set data for model induction and validation (S1), test set ligand structures stored in sdf format (S2), training and test set ligand descriptors (S3), and local descriptors of protein structure for training and test sets (S4). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.
(2) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
(3) Dudek, A. Z.; Arodz, T.; Galvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb. Chem. High Throughput Screening* **2006**, *9*, 213–228.
(4) McInnes, C. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494–502.
(5) Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. *Protein Eng.* **2002**, *4*, 305–311.
(6) 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.
(7) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(8) Sauton, N.; Lagorce, D.; Villoutreix, B.; Miteva, M. MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics* **2008**, *9*, 184.
(9) Gasteiger, J. Introduction. In *Chemoinformatics: a textbook*; Gasteiger, J., Engel, T., Eds.; Wiley-WHC: Weinheim, 2003; pp 1−13.
(10) Terfloth, L. Calculation of structure descriptors. In *Chemoinformatics*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH: Weinheim, 2003; pp 401−431.
(11) Lill, M. A. Multi-dimensional QSAR in drug discovery. *Drug Discovery Today* **2007**, *12*, 1013–1017.
(12) Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. S. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1525*, 180–190.
(13) Prusis, P.; Mucaniece, R.; Andersson, P.; Post, C.; Lundstedt, T.; Wikberg, J. E. S. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1544*, 350–357.
(14) Strömbergsson, H.; Prusis, P.; Midelfart, H.; Lapinsh, M.; Wikberg, J. E. S.; Komorowski, J. Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. *Proteins* **2006**, *63*, 24–34.
(15) Lapinsh, M.; Prusis, P.; Mutule, I.; Mutulis, F.; Wikberg, J. E. S. QSAR and proteo-chemometric analysis of the interaction of a series of

organic compounds with melanocortin receptor subtypes. *J. Med. Chem.* **2003**, *46*, 2572–2579.
(16) Lapinsh, M.; Prusis, P.; Uhlen, S.; Wikberg, J. E. S. Improved approach for proteochemometrics modeling: application to organic compound-amine G protein-coupled receptor interactions. *Bioinformatics* **2005**, *21*4289–4296.
(17) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.
(18) Kontijevskis, A.; Prusis, P.; Petrovska, R.; Yahorava, S.; Mutulis, F.; Mutule, I.; Komorowski, J.; Wikberg, J. E. S. A look inside HIV resistance through retroviral protease interaction maps. *PLoS Comput. Biol.* **2007**, *3*, e48.
(19) Lindström, A.; Pettersson, F.; Almquist, F.; Berglund, A.; Kihlberg, J.; Linusson, A. Hierarchical PLS modeling for predicting the binding of a comprehensive set of structurally diverse protein-ligand complexes. *J. Chem. Inf. Model.* **2006**, *46*, 1154–1167.
(20) Jain, T.; Jayaram, B. Computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes. *Proteins* **2007**, *67*, 1167–1178.
(21) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402–1414.
(22) Hvidsten, T. R.; Kryshtafovych, A.; Komorowski, J.; Fidelis, K. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics* **2003**, *19*, II81−II91.
(23) Strömbergsson, H.; Kryshtafovych, A.; Prusis, P.; Fidelis, K.; Wikberg, J. E. S.; Komorowski, J.; Hvidsten, T. R. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins* **2006**, *65*, 568–579.
(24) Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **2000**, *28*, 304–305.
(25) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
(26) Boström, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion. *J. Med. Chem.* **2006**, *49*, 6716–6725.
(27) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522–526.
(28) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.
(29) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **2005**, *60*, 333–340.
(30) 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.
(31) 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*, 235–342.
(32) Golovin, A.; Oldfield, T. J.; Tate, J. G.; Velankar, S.; Barton, G. J.; Boutselakis, H.; Dimitropoulos, D.; Fillon, J.; Hussain, A.; Ionides, J. M.; John, M.; Keller, P. A.; Krissinel, E.; McNeil, P.; Naim, A.; Newman, R.; Pajon, A.; Pineda, J.; Rachedi, A.; Copeland, J.; Sitnov, A.; Sobhany, S.; Suarez-Uruena, A.; Swaminathan, G. J.; Tagari, M.; Tromm, S.; Vranken, W.; Henrick, K. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **2004**, *32*, D211–216.
(33) Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **2004**, *32*, D431–433.
(34) Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.* **2005**, *33*, D266–268.
(35) Liu, H.; Hu, Z. Z.; Zhang, J.; Wu, C. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* **2006**, *22*, 103–105.
(36) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
(37) ProCheck operating manual appendix E. http://www.biochem.ucl.ac.uk/~roman/procheck/manual/manappe.html (accessed March 3, 2008).
(38) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comp. Method.* **1990**, *3*, 537–547.
(39) Dragon-Software for the calculation of molecular descriptors, 2.1; Talete srl: Milan, Italy, 2002.
(40) Hvidsten, T. R.; Kryshtafovych, A.; Fidelis, K. Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. *Proteins* **2008**, in press.
(41) Brenner, S. E.; Koehl, P.; Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **2000**, *28*, 254–256.
(42) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453.
(43) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277.
(44) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
(45) The R suite of statistics software. http://www.r-project.org/ (accessed April 7, 2008).
(46) Johnson, R. A.; Wichern, D. W. Clustering, distance methods and ordination. In Applied multivariate statistical analysis; Prentice-Hall: Upper Saddle River, NJ, 1998; pp 726−799.
(47) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed Oct 1, 2007).
(48) Hastie, T.; Tibshirani, R.; Friedman, J. In *The elements of statistical learning.* Springer-Verlag: New York, 2001; pp 371−376.
(49) Cona, A.; Manetti, F.; Leone, R.; Corelli, F.; Tavladoraki, P.; Polticelli, F.; Botta, M. Molecular basis for the binding of competitive inhibitors of maize polyamine oxidase. *Biochemistry* **2004**, *43*, 3426–3435.
(50) Baldwin, J.; Michnoff, C. H.; Malmquist, N. A.; White, J.; Roth, M. G.; Rathod, P. K.; Phillips, M. A. High-throughput screening for potent and selective inhibitors of Plasmodium falciparum dihydroorotate dehydrogenase. *J. Biol. Chem.* **2005**, *280*, 21847–21853.
(51) Heikkila, T.; Ramsey, C.; Davies, M.; Galtier, C.; Stead, A. M.; Johnson, A. P.; Fishwick, C. W.; Boa, A. N.; McConkey, G. A. Design and synthesis of potent inhibitors of the malaria parasite dihydroorotate dehydrogenase. *J. Med. Chem.* **2007**, *50*, 186–191.
(52) Martin, R. E.; Green, L. G.; Guba, W.; Kratochwil, N.; Christ, A. Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach. *J. Med. Chem.* **2007**, *50*, 6291–6294.
(53) Guba, W.; Green, L. G.; Martin, R. E.; Roche, O.; Kratochwil, N.; Mauser, H.; Bissantz, C.; Christ, A.; Stahl, M. From astemizole to a novel hit series of small-molecule somatostatin 5 receptor antagonists via GPCR affinity profiling. *J. Med. Chem.* **2007**, *50*, 6295–6298.

CI800200E