Novel Scoring Functions Comprising QXP, SASA, and Protein Side-Chain Entropy Terms

Fabrizio Giordanetto,[‡] Simona Cotesta,[§] Cornel Catana,[||] Jean-Yves Trosset, Anna Vulpetti, Pieter F. W. Stouten, and Romano T. Kroemer*

Computational Sciences, Pharmacia Italia, Pfizer Group, Viale Pasteur 10, 20014 Nerviano, MI, Italy

Received January 20, 2004

Novel scoring functions that predict the affinity of a ligand for its receptor have been developed. They were built with several statistical tools (partial least squares, genetic algorithms, neural networks) and trained on a data set of 100 crystal structures of receptor−ligand complexes, with affinities spanning 10 log units. The new scoring functions contain both descriptors generated by the QXP docking program and new descriptors that were developed in-house. These new descriptors are based on solvent accessible surface areas and account for conformational entropy changes and desolvation effects of both ligand and receptor upon binding. The predictive r^2 values for a test set of 24 complexes are in the 0.712−0.741 range and RMS prediction errors in the 1.09−1.16 log K_d range. Inclusion of the new descriptors led to significant improvements in affinity prediction, compared to scoring functions based on QXP descriptors alone. However, the QXP descriptors by themselves perform better in binding mode prediction. The performance of the linear models is comparable to that of the neural networks. The new functions perform very well, but they still need to be validated as universal tools for the prediction of binding affinity.

INTRODUCTION

The correct prediction of the binding affinity of a small-molecule ligand for a specific macromolecular target is of critical importance in computer-aided drug design,[1,2] both in the "hit identification" and subsequent "lead optimization" stages of the drug discovery process. During hit identification, binding affinity predictions serve the purpose of discriminating between binders and nonbinders. During lead optimization they help select the most potent ligands from within a congeneric chemical class. Different computational approaches have been devised to predict binding affinity. These include free energy perturbation and thermodynamic integration, knowledge-based and energy component methods.

A statistical mechanical prediction of relative free energies of binding between two molecules can be derived using either free energy perturbation (FEP) calculations[3] or thermodynamic integration (TI).[4] In principle, they are the most appropriate methods for the prediction of binding affinity as they consider solvent molecules explicitly and account for the molecular flexibility of both ligand and target. However, their applicability is limited because of the potential problem of inadequate conformational space sampling, limitations in the chemical differences between ligands that can be efficiently studied, and because of the long simulation runs required.

Knowledge-based approaches are heuristic in nature. They assume that a given problem can be solved through the application of rules that were learned from a sufficiently large number of examples. When applied to the prediction of ligand−protein affinity, the idea is that the necessary information can be found in the experimentally derived three-dimensional structures of ligand−target complexes. Statistical analysis of large structural databases defines the frequency distributions of specific interactions between certain ligand and protein atoms.[5] In a new system, not present in the original database, it can be assumed that only those molecular interactions that are close to the frequency maxima of the interactions in the database favor the binding event and therefore contribute to the binding affinity, whereas interactions that have been found to occur with low frequency in the database are likely to destabilize binding and decrease the overall affinity. The observed frequency distributions are converted to what is usually referred to as "potentials of mean force" or "knowledge-based potentials". Several knowledge-based potentials have been proposed to predict binding affinity (e.g., SmoG,[5] Bleep,[6] PMF,[7] DrugScore[8]). All these approaches differ mainly in the size of the training database employed and in the type and description of the molecular interaction considered.

Energy component methods are based on the assumption that the change in free energy upon binding of a ligand to its target can be decomposed into a sum of individual contributions (eq 1).

$$\Delta G_{bind} = \Delta G_{int} + \Delta G_{solv} + \Delta G_{conf} + \Delta G_{motion} \quad (1)$$

The individual terms in this equation account for the main energetic contributions to the binding event, as follows: specific ligand−receptor interactions (ΔG_{int}), the interactions

* Corresponding author phone: +39-02-4838 5221; fax: +39-02-4838 3965; e-mail: romano.kroemer@pharmacia.com.

[‡] Current affiliation: Computational Chemistry, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden.

[§] Current affiliation: Molecular Structure and Design, Pharmaceuticals Division, F. Hoffmann - La Roche Ltd., CH-4070 Basel, Switzerland.

[||] Current affiliation: Pfizer Global Research & Development, 2800 Plymouth Road, Ann Arbor, MI 48105.

10.1021/ci0499626 CCC: $27.50
Published on Web 04/14/2004

of ligand and receptor with solvent (ΔG_{solv}), the conformational changes in the ligand and the receptor (ΔG_{conf}), and the "motions" in the protein and the ligand during the complex formation (ΔG_{motion}). In principle, a separation into individual terms is possible only if the system of interest is divided into mutually independent variables.[9] However, many of the individual terms are highly correlated with each other, and they can affect the binding affinity in more than one way (i.e., positive or negative contribution).[10] Moreover, the free energy contributions are not calculated as ensemble mean values, but they are usually computed from a single structure. Despite the very approximate character of these assumptions, energy component methods are very appealing, as the simplifications result in functions that can be evaluated very rapidly. These scoring schemes are therefore applicable to large database computational ("virtual") screening and de novo design procedures. Despite the approximations in the underlying theory, these methods have been successfully applied to the prediction of protein−ligand affinity.[11−13]

A popular class of energy-based methods follows the "regression" approach. These methods assume an, often linear, statistical relationship between the total free energy change upon binding and a number of descriptors of the binding event. The final relationship takes the form of a regression equation where each independent variable (X_i) contributes to the observed dependent variable (Y_j) through a set of functional forms f_i (eq 2).

$$Y_j = \sum_{i=1}^{N} f_{i,j}(X_i) \qquad (2)$$

Y_j is usually represented by the experimentally measured binding affinity of a ligand for a specific target, whereas the X_i are a set of explanatory variables that are supposed to contribute to the overall change in free energy. A large number of explanatory variables have been proposed for regression-based scoring functions.[14−18] Most frequently, they include descriptors for hydrogen bonds[14−16] and ion pairs,[14,18] the amount of buried and contact surface,[14,18] and molecular flexibility of the ligands.[14,17,18] $f_{i,j}$ denotes a mathematical relationship between X_i and Y_j and can include multiplication with a constant or more complex functional forms. These functional forms can be determined by different data modeling techniques such as multiple linear regression, genetic algorithms (GA), partial least-squares regression, and neural networks. Such analysis is based on experimentally derived three-dimensional structures of ligand−protein complexes and available binding affinity data. Many popular scoring functions have been derived using the regression-based approach (e.g., LUDI,[14] ChemScore,[17] Validate[18]).

Since the relationship between descriptors of the binding event for a certain ligand−target complex and their actual binding affinity is statistical in nature, one needs to be concerned about the parameter transferability to different targets and the predictivity for compounds outside the training set. The best predictions are generally obtained when a model is applied to the training set on which it was developed. If a certain chemical class or a specific type of molecular interactions was not represented in the experimental data, these will be poorly modeled by the final regression equation, and predictions on compounds belonging to that class or making those interactions will be unsatisfactory.

The present work describes new scoring functions that were developed using the regression-based approach. A series of new descriptors characterizing various aspects of receptor−ligand interactions are presented. Their importance in the corresponding scoring functions is discussed in light of their relationships with the biological activity. The main purpose was to develop scoring functions that could be used as a filter after docking. They are not meant to guide the evolution of the docking run (i.e., the generation of the poses) but depend on a previously calculated pose of the ligand in the binding site of the receptor. However, considering that the identification of the "correct" pose of a ligand is considered a prerequisite for a reliable prediction of biological affinity,[19] it was of interest to evaluate as well whether the scoring functions developed here are able to identify the correct pose among a pool of different binding modes.

METHODS

Data Set. The Protein Data Bank (PDB)[20] was searched for structures of protein−ligand complexes. Complexes between nucleic acids and proteins or small molecules were not taken into account as the energetics of this type of associations have been shown to be significantly different from the ones involving proteins.[21] The goal was to develop scoring functions in order to rank small-molecule drugs. Therefore, protein−protein complexes and protein−large peptide complexes were excluded as well. Some of the descriptors used in the scoring functions are force field-like (see below). Since current force fields have difficulties to adequately capture polarizability and induced electrostatic effects, molecular complexes in which ligands coordinate to metal ions in the binding site of the protein were also excluded. The resulting three-dimensional structures were subjected to a resolution filter and rejected if their resolution was >3.0 Å. For the complexes surviving these filters, all available information with respect to biological activity was retrieved from the literature. This process resulted in a total of 124 protein−ligand complexes with known binding data. The corresponding equilibrium constants were converted into pK_d values (the negative logarithm of the dissociation constant). This implies that larger (more positive) pK_d values correspond to higher biological affinities. The final data set contains activity values spanning nearly 10 log units and covers 11 different protein families (Table 1).

The data set thus obtained was divided into training/validation and test sets in order to build the scoring functions and to evaluate their predictivity. To derive training/validation and test sets that are comparable in terms of activity values, the activity values were binned. 12 bins were generated using a bin width of 0.8275 pK_d units. For each bin 80% of the ligand/protein complex pairs were put into the training/validation set and the remainder into the test set. In the case of the least occupied bin (containing four complexes only) the training/validation and test percentage were 75 and 25%, respectively. Complexes were distributed over the training/validation and test sets such that these sets had similar affinity distributions but contained different families of proteins. This yielded a total of 100 ligands in

Table 1. Molecular Complexes Employed To Train and Test Scoring Functions with PDB Accession Numbers, pK_d Values,[a] and a Description of the Proteins and Ligands

PDB entry	pK_d	protein/ligand	PDB entry	pK_d	protein/ligand
1aaq	8.4	HIV-1 protease/hydroxyethylene	1tng	2.93	trypsin/aminomethylcyclohexane
1abe	6.52	ABP/L-arabinose	1tnh	3.37	trypsin/4-fluorobenzylamine
1abf	5.42	ABP/D-fucose	1tni	1.7	trypsin/4-phenylbutylamine
1apb	5.82	ABP(P254G)/D-fucose	1tnj	1.96	trypsin/2-phenylethylamine
1apt	9.4	penicillopepsin/pepstatin analogue	1tnk	1.49	trypsin/3-phenylpropylamine
1apu	7.49	penicillopepsin/pepstatin analogue	1tnl	1.88	trypsin/t-2-phenylcyclopropylamine
1apv	9	penicillopepsin/IvaVV(H)Dfo-*N*-methylamide	1ulb	4.4	PNP/guanine
1apw	8	penicillopepsin/IvaVVDfo-*N*-methylamide	2ak3	3.86	adenylate kinase isoenzyme-3/AMP
1bap	6.85	ABP(P254G)/L-arabinose	2cgr	7.27	immunoglobulin/GAS
1bra	1.82	trypsin mutant/benzamidine	2ctc	3.89	carboxypeptidase A/L-phenyl lactate
1cla	5.28	chloramphenicol acetyltransferase/chloramphenicol	2dbl	8.7	DB3/pregnane analogue
1csc	7.1	citrate synthase/carboxymethyl coenzyme A	2gbp	7.4	GBP/D-glucose
1dbj	7.68	DB3/aetiocholanolone	2mcp	4.7	immunoglobulin/phosphocholine
1dbk	8.09	DB3/5-beta-androstanedione	2phh	3.36	PHBH/ADP ribose
1dbm	9.44	DB3/progesterone analogue	2phh	4.6	PHBH/p-hydroxybenzoate
1dhf	7.4	DHFR/folate	2sns	6.7	staphylococcal nuclease/2′-deoxy-3′,5′-diphosphothymidine
1dr1	5.57	DHFR/biopterin			
1drf	7.44	DHFR/folate	2ypi	4.82	TP isomerase/2-phosphoglycolate
1dwb	2.9	thrombin/benzamidine	3cla	4.94	chloramphenicol acetyltransferase/chloramphenicol
1dwc	7.41	thrombin/argatroban	3cpa	4	CPA/GY
1dwd	8.18	thrombin/NAPAP	3csc	5.15	citrate synthase/acetyl coenzyme A
1etr	7.41	thrombin/MQPA	3ptb	4.5	trypsin/benzamidine
1ets	8.22	thrombin/NAPAP	4cla	5.47	chloramphenicol acetyltransferase/chloramphenicol
1ett	6.19	thrombin/4-TAPAP	4dfr	8.62	DHFR/methotrexate
1fkb	9.7	FK506 binding protein/rapamycin	4fab	8.05	IgG kappa Fab 4-4-20/fluorescein dianion
1fkf	8.77	FK506 binding protein/FK506	4gr1	2.2	glutathione reductase/retro-GSSG
1hbv	6.37	HIV-1 protease/SB203238	4hvp	6.11	HIV-1 protease/MVT-101
1hpv	9.22	HIV-1 protease/VX-478	4phv	9.17	HIV-1 protease/L-700417
1hvi	10.07	HIV-1 protease/A-77003	4sga	3.27	proteinase A/Ace-Pro-Ala-Pro-Phe
1hvj	10.45	HIV-1 protease/A-78791	4tim	2.16	triosephosphate isomerase/2-phosphoglycerate
1hvl	9	HIV-1 protease/A-76889	5abp	6.64	ABP/D-galactose
1hvs	10.08	HIV-1 protease (V82A)/A-77003	5enl	3.8	enolase/2-phospho-D-glycerate
1ldm	5.44	lactate dehydrogenase/NAD	5sga	2.85	proteinase A/Ace-Pro-Ala-Pro-Tyr
1lgr	3.07	glutamine synthetase/AMP	5tim	2.3	triosephosphate isomerase/DTT
1lyb	11.42	cathepsin D/pepstatin	6abp	6.36	ABP(M108L)/L-arabinose
1mcb	4.84	immunoglobulin/peptide	6rnt	2.37	ribonuclease T1/2′-AMP
1mcf	5.15	immunoglobulin/peptide	6tim	3.21	triosephosphate isomerase/glycerol-3-phosphate
1mch	5.15	immunoglobulin/peptide	7abp	6.46	ABP(M108L)/D-fucose
1mcj	3.78	immunoglobulin/peptide	7cat	8	catalase/NADPH
1mcs	4.84	immunoglobulin/peptide	7dfr	4.96	DHFR/NADP+
1mdq	5.1	maltose binding protein A301GS/maltose	7dfr	6.1	DHFR/folate
1mfe	5.31	immunoglobulin/D-gal-D-abe-D-man	7est	7.6	elastase/TFAP
1pgp	5.7	6-PGDH/6-phosphogluconic acid	7hvp	9.62	HIV-1 protease/JG-365
1phh	7.35	PHBH/FAD	7tim	5.4	triosephosphate isomerase/phosphoglycolohydroxamate
1ppc	6.16	trypsin/NAPAP			
1pph	6.22	trypsin/3-TAPAP	8abp	6.6	ABP(M108L)/D-galactose
1ppk	7.66	penicillopepsin/phosphopeptide analogue	8hvp	9	HIV-1 protease/U-85548E
1 ppm	5.8	penicillopepsin/CBZ-Ala-Ala-Leu-P−(O)Phe-Ome	9abp	8	ABP(P254G)/D-galactose
1rnt	5.18	ribonuclease T1/2′-GMP	9hvp	8.35	HIV-1 protease/A-74704
1rus	3.08	rubisco/3-phosphoglycerate	9ldt	4.74	lactate dehydrogenase/oxamate
1tmt	6.24	thrombin/D-Phe-Pro-Arg	9ldt	5.43	lactate dehydrogenase/NADH
			Test Set		
1adf	4.58	alcohol dehydrogenase/beta-TAD	2ifb	5.44	FABP/palmitic acid
1dbb	9	DB3/progesterone	2pk4	4.32	plasminogen kringle 4/aminocaproic acid
1dih	5.74	dihydrodipicolinate R/NADPH	2r04	6.22	virus coat protein/compound IV
1fbp	4.82	fructose-1,6-bisphosphatase/AMP	3fx2	9.3	flavodoxin/FMN
1hsl	7.3	histidine-binding protein/H	3pgm	3.19	phosphoglycerate mutase/phosphoglycerate
1hvk	10.11	HIV-1 protease/A-76928	3tmn	5.9	thermolysin/VW
1l83	3.4	lysozyme/benzene	4ts1	5.61	tyrosyl-transfer RNA synthetase/tyrosine
1nnb	5.3	neuraminidase/DANA	5acn	2.8	aconitase/tricarballylic acid
1rbp	6.72	retinol-binding protein/retinol	5cna	2	concanavilin A/O1-methyl-mannose
1rne	8.7	renin/CGP-38560	6apr	7.77	rhizopuspepsin/pepstatin
firsta	5.35	transthyretin/3,3′-diiodo-L-thyronine	8atc	7.57	aspartate carbamoyltransferase/PALA
2dri	6.52	D-ribose binding protein/beta-D-ribose	9aat	8.22	aspartate aminotransferase/pyridoxal-5′-phosphate

[a] Taken from refs 14−18 or from the PDB structure's reference.

the training/validation set and 24 compounds in the test set. Descriptive statistics based on biological affinity values, ligand-based molecular characteristics, and structural information for the training/validation and test sets are shown in Figure 1. Table 1 presents the details for the corresponding molecular complexes.

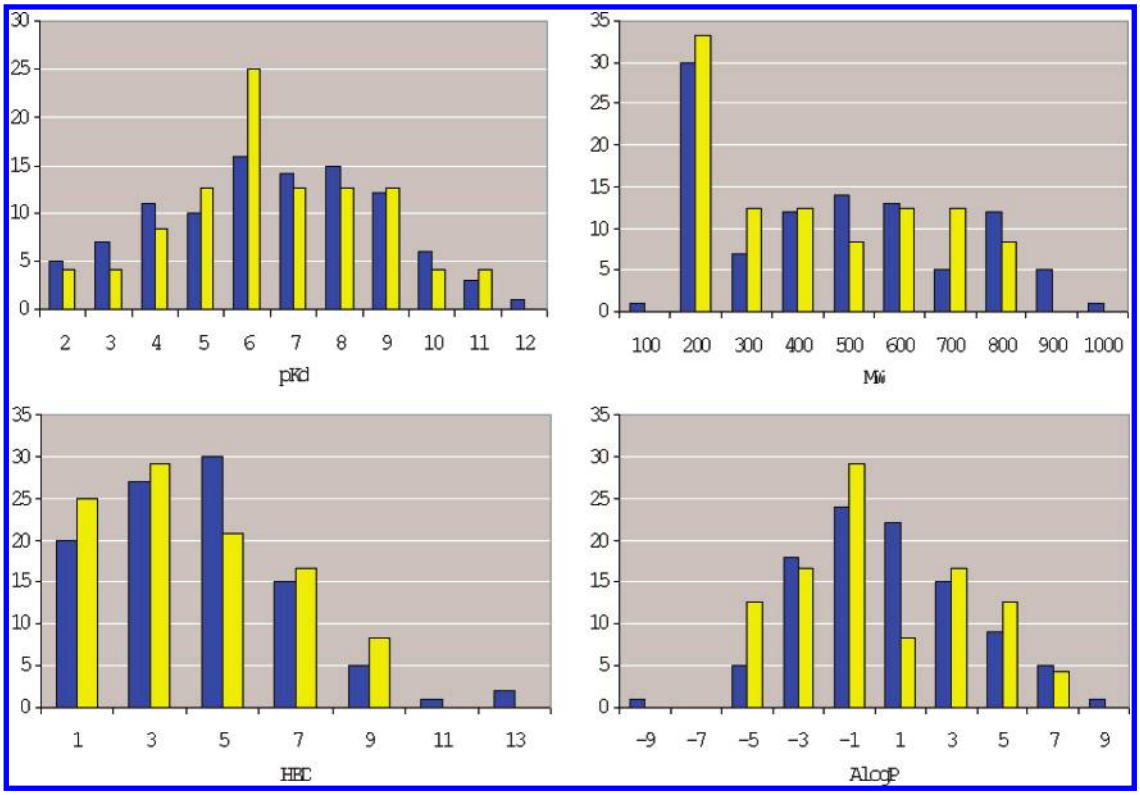

Figure 1. Descriptive statistics for the molecular complexes included in the training/validation (blue columns) and test (yellow columns) sets. HBC - hydrogen bond count.

The data set features a wide activity distribution with most of the compounds exhibiting affinity values in the nanomolar−micromolar range. The majority of the compounds are small organic molecules with characteristics that satisfy Lipinski's "Rules of 5" for optimal absorption profiles.[22] However, almost 30% of the ligands have molecular weight greater than 500. Many cofactors (e.g., FAD, NADPH) and pseudopeptides, exhibiting very different biological affinity values, are present in this subset of the data set. Since we aim to develop models with broad applicability and since these heavier ligands contribute diverse structural and biological activity information, it was deemed important to include them.

Each molecular complex was processed as follows. The ligand was extracted from the complex and checked for the correct atom and bond types. Ligand bond types were assigned manually as the PDB files lack bond order information for the ligands. Acidic and basic groups on the ligands were assigned likely protonation states for a biological environment (pH=7.4), and hydrogen atoms were added or removed accordingly. Water molecules and other cofactors were retained as part of the protein. The correct ionization state for charged amino acids and the correct rotamer for asparagine and glutamine residues were selected on the basis of the molecular surroundings and the ability to maintain hydrogen-bonded networks. Subsequently, the protein was cropped around the ligand, selecting only those residues that were within 10 Å from the ligand atoms. Finally, the energy of the molecular complexes was minimized using QXP.[23] Here, the ligand as well as all the side chains and water molecules present in the binding site was treated as flexible in order to eliminate steric overlap between atoms.

Force Field Descriptors. The receptor−ligand complexes were energy minimized using QXP.[23] It was therefore possible to use the force field terms calculated during the minimization in the development of new scoring functions. These descriptors include a contact score (*Cntc*), a hydrogen bond energy (*Hbnd*), a desolvation penalty term (*Pslv*), the internal energy of the ligand (*Intl*), and a bump monitor term (*Bump*).

Cntc is a distance-dependent ligand−receptor contact score. For each ligand/receptor atom pair (*i*, *j*), a distance dependent contact value $C_{i,j}$ is calculated. The contact function $F_{i,j}$ employed is a fourth-order polynomial of the van der Waals distance between the atoms, which is then weighted using the atomic volumes (V_i, V_j) of the two atoms involved (eq 3). The weighting by the atomic volume is done to account for desolvation effects. The final *Cntc* is the sum of the pair values computed for the ligand−receptor complex (eq 4).

$$C_{i,j} = F_{i,j} \cdot V_i \cdot V_j \tag{3}$$

$$Cntc = \sum_{i,j} C_{i,j} \tag{4}$$

Hbnd represents the hydrogen bond energy of the receptor−ligand complex. For all the receptor−ligand atom pairs in which one atom is a hydrogen bond donor (*i*) and the other is a hydrogen bond acceptor (*j*), the hydrogen bond energy $Hb_{i,j}$ is computed. This is defined based on a fourth-order polynomial function of the van der Waals distance between the two atoms $F_{i,j}$, which is multiplied by the corresponding hydrogen bond strengths (H_i, H_j), based on the electronegative character of the atoms, as shown in eq 5. The overall *Hbnd* is the sum of the hydrogen bond energies ($Hb_{i,j}$) for all the hydrogen bond pairs (eq 6).

$$Hb_{i,j} = F_{i,j} \cdot H_i \cdot H_j \quad (5)$$

$$Hbnd = \sum_{i,j} Hb_{i,j} \quad (6)$$

The main contact score *Cntc* does not discriminate between polar and nonpolar atoms. *Pslv* is used as a correction term and represents the sum of the individual polar desolvation scores $P_{i,j}$ for the atom pairs in which only one atom is able to form hydrogen bonds:

$$Pslv = \sum_{i,j} P_{i,j} \quad (7)$$

The polar desolvation $P_{i,j}$ for an atom pair formed by hydrogen bonded atoms *i* and *j* is defined as follows:

$$P_{i,j} = F_{i,j} \cdot H_i \cdot V_i \cdot V_j \quad (8)$$

Here, $F_{i,j}$ is the same functional form used in the *Cntc* term, H_i is the hydrogen bond strength of atom *i*, and V_i and V_j are the atomic volumes for atom *i* and *j*, respectively.

The *Intl* term describes the internal free energy of the ligand as a function of the torsional free energy, E_{tors}, and the internal contact score, *ICS*:

$$Intl = f(E_{tors} - ICS) \quad (9)$$

E_{tors} is the free energy that corresponds to the torsion potential which is derived from a number of experimental structures. As it is derived from experimental structures, it contains intrinsically both enthalpic and entropic contributions. In addition, *Intl* is modulated by the formation of internal contacts between atoms that are separated by more than three bonds. In principle, these internal contacts are energetically favorable, and therefore the ligand is assumed to be partially folded in solution. However, the formation of the internal contacts is concomitant with an entropic penalty, as the torsions between the contact atoms are frozen. Therefore, the *ICS* energy is divided by the number of torsions between the two atoms making an internal contact.

The *Bump* term accounts for the presence of steric clashes between the ligand and the receptor. The energy associated with a close contact between atom *i* and *j*, $E_{i,j}$, is zero if $D_{i,j} > C_{i,j}$ and proportional to the square of the difference between the actual distance $D_{i,j}$ and the contact distance $C_{i,j}$ (i.e., the sum of their van der Waals radii) otherwise:

$$E_{i,j} \propto (D_{i,j} - C_{i,j})^2 \quad (10)$$

Protein Solvent Accessible Surface Area Descriptors Including Entropy Contributions. This class of descriptors involves the calculation of the solvent accessible surface area (ASA), as defined by the algorithm of Lee and Richards,[24] using a probe radius of 1.4 Å. This is done for the ligand and the protein in both the free and bound forms, and the results are divided into contributions from polar (nitrogen and oxygen) and nonpolar atoms. Murphy and Freire showed that for a limited set of protein−protein complexes these quantities predicted unfolding energetics in good agreement with experimental data.[25] They proposed an amino acid-based conformational entropy contribution of the protein to the binding affinity[25] that we also use in our study. This term assumes that the conformational entropy of a side-chain *P_dS* is zero when fully buried and scales linearly with the solvent accessible surface until it reaches a maximum value when fully exposed

$$P\_dS = \sum_{i=1}^{N} \frac{\Delta ASA_{SC,i}}{ASA_{AXA,i}} \cdot \Delta S_{bu \rightarrow ex} \quad (11)$$

where $\Delta ASA_{SC,i}$ is the change in ASA of side-chain *i* on binding, $ASA_{AXA,i}$ is the ASA of the side-chain reference state, an extended Ala-X-Ala tripeptide, and the summation extends over all *N* side chains in the protein. $\Delta S_{bu \rightarrow ex}$ is the buried-to-exposed entropy gain for a side-chain of type *i* as defined by Lee et al.[26]

Ligand Entropy. The conformational entropy term discussed above cannot be applied to small organic molecules, as the original parametrization has been developed for the 20 natural amino acids.[26] Therefore, the number of rotatable bonds in the ligands, ignoring any terminal symmetric groups (e.g., methyl, trifluoromethyl), is computed to capture the change in the conformational entropy for the ligand upon complexation with the receptor. This descriptor was used alongside the related *Intl* term implemented in QXP, as they are conceptually different. The rotatable bonds term accounts for freezing of torsions and is purely entropic in nature, whereas *Intl* also includes enthalpic contributions (i.e., contacts).

Ligand Charge- and Atom Type-Based Solvent Accessible Surface Area Descriptors. This group of descriptors is an evolution of the previous set and aims to describe the different energetic costs associated with the burial of atoms with different physicochemical characteristics during the binding event. Here the simple two-class polar/nonpolar distinction based on atom types is replaced by multiple classes based on calculated partial charges**.** First, atomic charges are computed for all the atoms in the ligand−receptor complex using the Gasteiger−Marsili method.[27] Then, the atoms are divided in two classes, positive and negative, according to the value of the atomic charge. Moreover, the two classes comprising atoms that are either positively or negatively charged are further subdivided into three different atom type-based categories, containing hydrophilic, hydrophobic sp^2, and hydrophobic sp^3-hybridized atoms, respectively. Protein atom types are assigned on the basis of the PDB atom definitions and ligand atom types by the Tripos force field.[28] This procedure resulted in a total of six different atomic classes (i.e., hydrophilic positive, hydrophilic negative, hydrophobic positive, hydrophobic negative, hydrophobic sp^3-hybridized positive, hydrophobic sp^3-hybridized negative).

Model Building − General. The molecular complexes were energy minimized using QXP.[23] This served to remove any steric clashes present in the crystal structures and to calculate the corresponding force field descriptors. The remaining descriptors, i.e., solvent-accessible surface descriptors and the side chain-based conformational entropy term were computed from the energy-minimized structure of the complexes. The correlation coefficient was calculated for all the possible descriptor pairs and between each descriptor and the biological activity using Cerius2.[29] If the correlation coefficient was higher than 0.7 for a given descriptor pair, then the descriptor that was more correlated

with the biological activity (i.e., that had the larger value of $|r|$) was retained. This process resulted in a total of 25 descriptors. Although in theory different descriptors may correlate better with affinity when nonlinear techniques are used, it was decided to use a single set of descriptors for all model development efforts.

For all statistical techniques employed, the internal consistency of the models was evaluated by using randomization tests on the biological activity values and different types of cross-validation using Cerius2.[29] These included leaving one, five, and 10 observations out during model building. The process was repeated 100 times in the case of leave-five and leave-ten-out schemes. External predictivity was subsequently evaluated by predicting the biological activity values for the 24 complexes in the test set.

Model Building − Linear Methods. These models were built using stepwise linear regression (backward and forward), principal component regression (PCR), and partial least squares (PLS), using the QSAR module of Cerius2.[29] The maximum number of steps in the stepwise linear procedure was set to 100 with an *F*-value for the introduction of a new variable set to 4. For PCR, principal components were retained until at least 90% of the variance in descriptor space was explained.

Model Building − Genetic Algorithms. Models were built using both normalized (mean centered and variance scaled) and unnormalized descriptors. The genetic function approximation,[30] as implemented in the Cerius2 program,[29] was employed to allow for the generation of nonlinear models. Specifically, an initial population of 1000 randomly built models was evolved for 50 000 generations. During the evolutionary process, different mutational operators such as adding a new term are applied to the population of models. Each mutational operator is applied according to an associated probability value. In this study, the probability value was set to 50%. Creation of linear, off-set linear, quadratic, and off-set quadratic terms was allowed.

Model Building − Neural Networks. Back-propagated feed forward neural networks were built using the NETLAB toolbox implemented in MATLAB.[31] The training/validation set (100 complexes, Table 1) was divided into a model building (80 complexes) and a model validation set (20 complexes). The architecture of the back-propagated networks involved three different levels: an input layer, a hidden, and an output layer. The same 25 molecular descriptors, as previously described, were used after mean-centering and variance-scaling. PCA was performed on the descriptors, and the corresponding principal components were retained until at least 90% of the variance in descriptor space was explained, as previously described**.** Input variables for the networks consisted of the remaining 4 principal components, and the output (target) variables were represented by the normalized biological activities.

The number of neurons in the hidden layer was iteratively changed monitoring the predictive power of the resulting network. An optimal number of 8 hidden neurons resulted from this procedure. The hidden neurons were assigned a scalar bias and a 'tanh' activation function. The single output neuron was assigned a linear transfer function. Training was performed using the Levenberg−Marquardt algorithm[32] during which the network performance for both model building and model validation sets was monitored. Whenever the validation error increased for a specified number of iterations, the training was stopped (early stopping), and the resulting neural network at the minimum of the validation error was returned.

N-Way PLS Model. The previous model building techniques employed data from energy-minimized crystal structures. In doing so, one may encounter the following two problems: the resulting scoring functions do not separate correct from supposedly incorrect binding modes, and they work well when redocking cognate ligands but not when docking other well-binding ligands (for a recent example, refer to ref 33). We addressed these problems recently with a novel in-house approach (C. Catana et al., in preparation). In this approach ligands are docked to a cognate or noncognate crystal structure, and for each ligand multiple poses are saved. A small number of diverse poses (both in terms of calculated affinity, binding mode and interaction terms) are selected from this saved set and used to parametrize the scoring function, using Accelrys[29] and QXP[23] scoring function terms as variables, and n-way PLS as statistical technique. Although the approach has been specifically developed to work in real life situations where one has few or no crystal complexes, it works also very well when applying regular PLS to data exclusively derived from crystal structures.

Evaluation of Pose Prediction Quality. For this purpose, the compounds in the data set were docked to the corresponding crystal structures using 300 MC steps in QXP. For each ligand, a total of 20 conformations were saved and analyzed. To make sure that the correct poses were included, the energy-minimized ligand structures from the original PDB complexes were added to the set of 20 poses. The descriptors described above were computed on all those poses, which were then scored using the different scoring functions developed here.

The different poses of each ligand were evaluated by calculating the RMSD to the crystallographically observed structure. RMSD values smaller than 1.0 Å usually indicate that a ligand is correctly docked. However, there could be cases in which the ligand is correctly oriented overall, but one or a few moieties are not properly placed. These situations result in higher RMSD values, but the pose should not be considered incorrect. Therefore, it was decided to present the results with two different RMSD cutoffs (i.e., RMSD < 1.0 Å and RMSD < 2.0 Å).

RESULTS AND DISCUSSION

More than 1700 scoring functions were built employing the model building techniques and descriptors outlined above. In this study, the models were evaluated with respect to their ability to predict the biological activity of the test set. Their performance was assessed on the basis of two metrics that reflect differences between predicted and experimentally determined test set activities: the predictive r^2 (r^2_{pred}) and the root-mean-square prediction error (RMSE).

The scoring functions built using linear and genetic algorithms contain explicit coefficients and descriptors, while the ones developed with neural networks procedures hide the contribution of each descriptor to the final prediction and their behavior cannot be easily interpreted. For these reasons, these two types of functions are discussed separately.

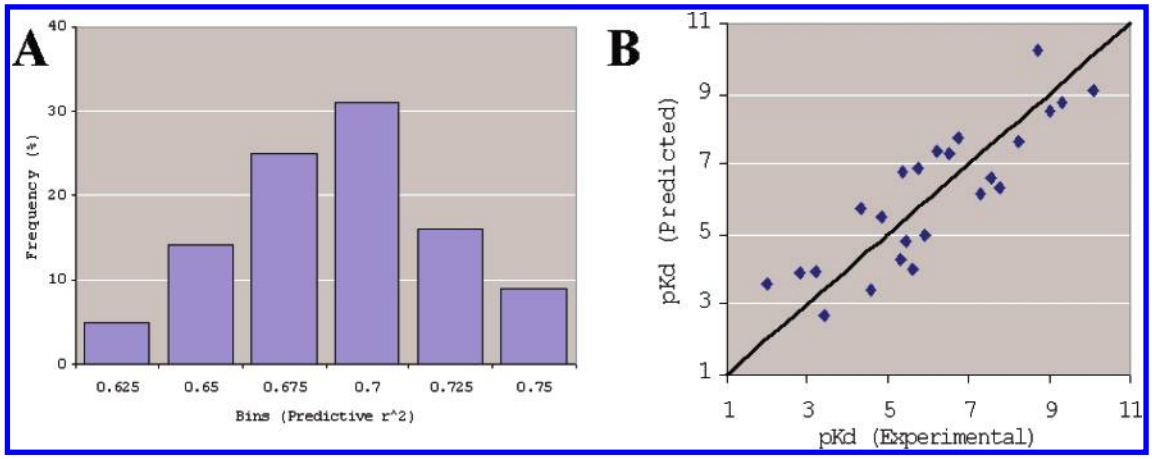

Figure 2. A: Frequency distribution of the r^2_{pred} values, as computed on the test set consisting of 24 complexes, for the best 200 scoring functions developed using linear regression and GA techniques. B: Experimental versus prediction values for the best scoring function obtained with stepwise multiple linear regression.

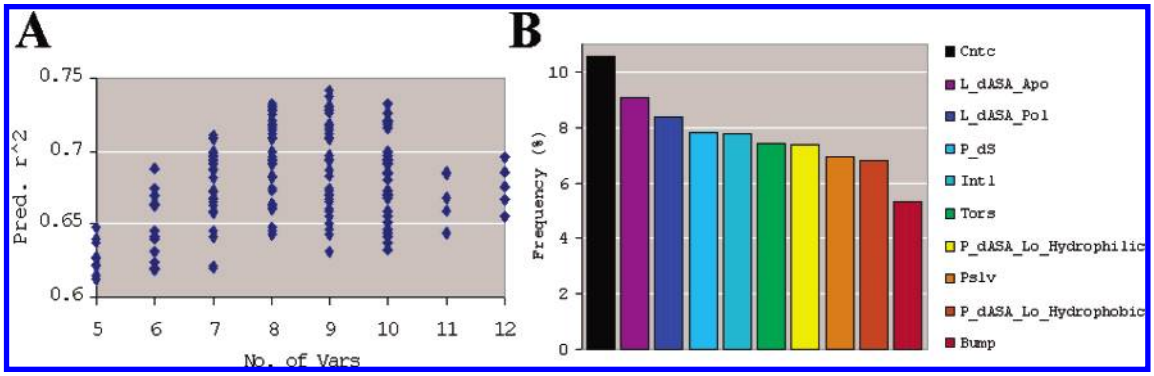

Figure 3. A: Dependence of the predictive power of linear and GA scoring functions on the number of independent variables employed. B: Descriptors most frequently employed by the best 200 linear and GA scoring functions.

Linear and Genetic Algorithms. The best 200 models (highest r^2_{pred} values) resulting from the application of linear regression and genetic algorithms were retained and examined further. These models originated from the various internal cross-validation procedures described before, and they differ in the number and type of independent variables employed and in their predictive power for the test set. They contain unnormalized descriptors only. Models were also built with normalized descriptors, but this did not improve the external predictivity and they are, therefore, not shown.

The frequency distribution of the r^2_{pred} values for the best 200 models built using linear and GA techniques is presented in Figure 2A. r^2_{pred} values range from 0.612 to 0.741 with a frequency peak in the 0.675−0.700 bin. The RMSE values for the 200 best models are in the 1.09−1.33 log unit range. The correlation plot of predicted and experimental activities for the best model (r^2_{pred} = 0.741) is shown in Figure 2B.

The resulting scoring functions show significant differences in the number of terms, ranging from 5 to 12, excluding the intercept. The relationship between the number of independent variables included in the models and the corresponding predictivity is depicted in Figure 3A. The most predictive model contains 9 different descriptors. However, several scoring functions with high r^2_{pred} values (e.g., 0.7−0.725) contain a lower number of variables (e.g., 7), as shown in Figure 3A. This indicates that a higher number of descriptors does not necessarily result in a significant gain in predictivity.

An interesting result was the absence of quadratic and offset quadratic descriptor dependencies in the top 200 scoring functions. The creation of such terms was allowed during the GA optimization to reflect possible nonlinear relationships between the biological activity and the various descriptors. As the GA algorithm evolves automatically toward better models, the absence of nonlinear terms in the final equations suggests that, at least for the data set employed in this study, linear relationships are sufficient to predict the observed binding affinities.

Figure 3B depicts the 10 descriptors that occur most frequently in the best 200 scoring functions and which, consequently, are expected to contribute most significantly to the observed affinity. Four descriptors stem from the latest (2003) version of QXP (*Cntc*, *Intl*, *Pslv*, and *Bump*) and 6 were specifically developed in this study (*Tors*, the number of rotatable bonds in the ligand; *L_dASA_Pol*, the change in polar solvent-accessible surface area for the ligand; *L_dASA_Apo*, the change in nonpolar solvent-accessible surface area for the ligand; *P_dS*, the side chain-based conformational entropy term for the protein *P_dASA_Neg_Hydrophilic*, the change in negatively charged hydrophilic solvent-accessible surface area for the protein and *P_dASA_Neg_Hydrophobic*, the change in negatively charged hydrophobic solvent-accessible surface area for the protein). All these terms contribute to the biological activity with optimizable weights (coefficients). Interestingly, each descriptor shown in Figure 3B is associated with comparable

Table 2. Descriptors Most Frequently Employed by the Best Linear and GA Scoring Functions[a]

term	description	contribution to pK_d
Cntc	favorable interactions	+
Bump	unfavorable interactions	-
Tors	ligand conformational entropy	-
Intl	ligand internal free energy	-
P_dS	protein conformational entropy	-
Pslυ	penalty for unpaired hydrogen bond atoms	-
P_dASA_Neg_Hydrophilic	ASA change for negatively charged hydrophilic protein atoms	-
P_dASA_Neg_Hydrophobic	ASA change for negatively charged hydrophobic protein atoms	+
L_dASA_Pol	ASA change for polar ligand atoms	-
L_dASA_Apo	ASA change for nonpolar ligand atoms	+

[a] Their overall contribution (sign of descriptor * sign of coefficient) to the predicted pK_d value (positive numbers correspond to higher biological activity values) is reported.

coefficients (i.e., sign and numerical value) across the best 200 scoring functions developed in this study, therefore suggesting that its contribution to the overall binding affinity can be generalized. More specifically, these descriptors can be grouped with respect to their physical meaning and to their influence on the final predictions as outlined in Table 2. Here, four main classes can be identified: the first two account for the favorable (*Cntc*) and unfavorable (*Bump*) interactions between the ligand and the protein, the third one describes the conformational entropy contributions of the ligand (*Intl* and *Tors*) and the protein (*P_dS*), and the last group (*L_dASA_Pol*, *L_dASA_Apo*, *P_dASA_Neg_Hydrophilic*, *P_dASA_Neg_Hydrophobic*, and *Pslυ*) captures the desolvation aspects of the binding event (Table 2).

The *Cntc* term is the sum of all the atomic contact pairs (distance-dependent and atomic volume-weighted) between the ligand and the protein and contributes favorably to the predicted binding affinity. Accordingly, a ligand that establishes extensive molecular contacts with the receptor receives better scores (Table 2).

The *Bump* term accounts for steric clashes (i.e., ligand and protein atoms that are closer together than the sum of their van der Waals radii) and this obviously decreases the biological activity (Table 2).

As described earlier, *Intl* accounts for the internal free energy of a ligand by calculating its torsional free energy (E_{tors}) in the docked conformation and considering the favorable interactions between ligand atoms that are separated by more than three bonds (*ICS*). Larger values of *Intl* correspond to a strained ligand (E_{tors} is high) and/or to a small *ICS* value, which in turn is due to few internal contacts (corresponding to a ligand bound in an extended conformation) or to internal contacts separated by many bonds (corresponding to a partially folded but flexible ligand). In any case a high value of *Intl* is detrimental to binding and, consequently, its coefficient is negative.

Tors is another term related to ligand flexibility. By simply counting the number of rotatable bonds in the ligand, it accounts for the negative entropy change that results from the freezing of a flexible ligand upon binding. Consistent with previous studies,[14−18] the *Tors* term contributes adversely to the predicted biological activity in our top 200 scoring functions (Table 2).

The top 200 scoring functions that we analyzed contained either the *Intl* or the *Tors* descriptor but not both terms together. This indicates that *Intl* and *Tors* are equivalent as descriptors of the conformational entropy contribution of the ligand to the binding affinity.

P_dS describes the conformational entropy contribution of the protein upon ligand binding.[25] It is derived from the change of solvent accessible surface area for the side chains in the protein binding site, as described in the Methods section. Side chains in an empty binding site are fully exposed and are able to sample a large number of conformational states. When the protein is bound to a ligand, side chains are buried and frozen in a single rotameric state. Entropy loss disfavors binding and *P_dS* affects the predicted affinity adversely (Table 2).

In the 200 best scoring functions five different descriptors (*Pslυ*, *L_dASA_Pol*, *L_dASA_Apo*, *P_dASA_Neg_Hydrophilic*, and *P_dASA_Neg_Hydrophobic*) account for the desolvation contributions. *L_dASA_Pol* describes the change of polar (negative or positive) solvent accessible surface area for the ligand during the binding. *P_dASA_Neg_Hydrophilic* represents the amount of hydrophilic negatively charged surface area that is buried during the binding for the protein. *L_dASA_Pol* and *P_dASA_Neg_Hydrophilic* account for the unfavorable process of transferring a polar solute (i.e., protein, ligand) from a fully exposed (solvated) to a buried (desolvated) state. Both terms affect the predicted biological activity negatively (Table 2). *Pslυ* is a force field descriptor that accounts for the contacts between polar hydrogens (i.e., hydrogen atoms bound to oxygen, nitrogen, or sulfur atoms) and nonpolar atoms in the ligand−protein complex. Because the energy penalty due to the desolvation of polar hydrogen atoms upon binding is not compensated by stabilizing polar−polar interactions, *Pslυ* has a negative coefficient in the 200 scoring functions (Table 2). *P_dASA_Neg_Hydrophobic* describes the area of negatively charged hydrophobic atoms (not sp^3-hybridized) on the protein that is buried during the binding. *L_dASA_Apo* encodes similar information, as it accounts for the nonpolar area of the ligand that is buried upon binding. The water molecules that surround the hydrophobic parts of the ligand and the protein are highly ordered, according to the so-called "iceberg" model.[34] During binding, the nonpolar parts of protein and ligand come in contact and free up the ordered layers of water molecules, resulting in favorable entropy changes. Again, the fitting procedure yields the correct sign of the coefficient for these two descriptors (*P_dASA_Neg_Hydrophobic*, and *L_dASA_Apo*), and they contribute positively to the overall binding affinity (Table 2).

Table 3 shows the details for the most predictive scoring function derived in this study (model 1) and another one that represents a tradeoff between predictive power and number of descriptors (model 2). Apart from the additional

Table 3. Details for the Best Scoring Functions Developed Using Stepwise Linear Regression and GA

	model 1	model 2
number of variables	9	7
r^2_{pred}	0.741	0.712
RMSE (pK_d)	1.09	1.16
pK_d =	1.796	2.059
	- 0.0069 * *L_dASA_Pol*	- 0.0059 * *L_dASA_Pol*
	+ 0.0008 * *L_dASA_Apo*	+ 0.0013 * *L_dASA_Apo*
	- 0.0502 * *P_dS*	- 0.0272 * *P_dS*
	- 0.117 * *Intl*	- 0.0079 * *Tors*
	+ 0.097 * *Cntc*	+ 0.109 * *Cntc*
	- 0.069 * *Pslv*	- 0.114 * *Pslv*
	- 0.00128 * *P_dASA_Neg_Hydrophilic*	- 0.0123 * *P_dASA_Neg_Hydrophilic*
	+ 0.0246 * *P_dASA_Neg_Hydrophobic*	
	+ 0.0036* *P_dASA_Neg_HydrophobicSp3*	

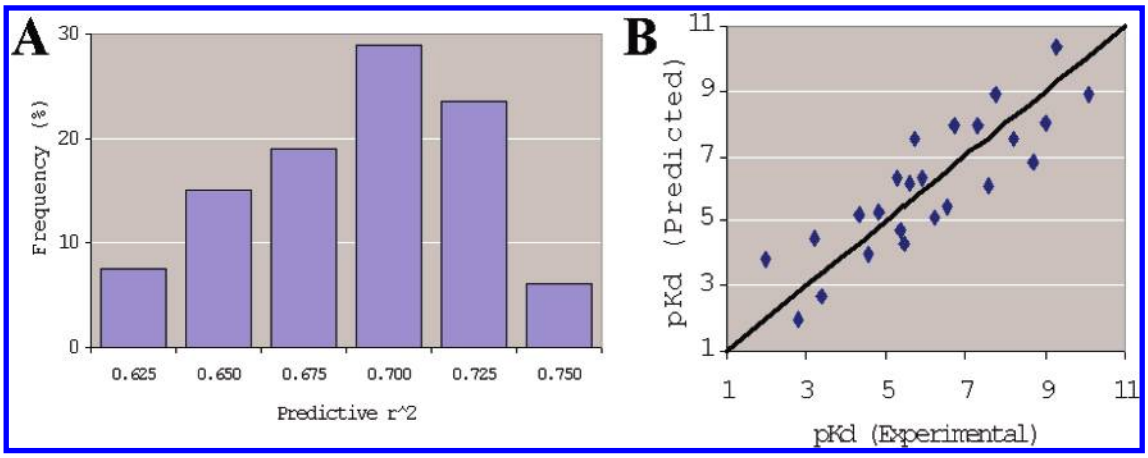

Figure 4. A: Frequency distribution of test set r^2_{pred} values for the best 200 neural networks developed in this study. B: Experimental versus averaged prediction values for the ensemble of the 40 best neural networks.

terms present in model 1, the two models are very similar, and the coefficients for the descriptors common to both functions are comparable. Interestingly, model 1 uses the *Intl* term as a measure of the conformational entropy of the ligand, whereas model 2 employs the number of rotatable bonds (*Tors*). In model 1 there are two more terms, compared to model 2: *P_dASA_Neg_Hydrophobic*, *P_dASA_Neg_HydrophobicSp*^{3}. These terms account for the burial of protein hydrophobic atoms upon binding. It is interesting to note that the burial of negatively charged and sp^3-hybridized hydrophobic atoms (*P_dASA_Neg_HydrophobicSp*^{3}) is less favorable than the one involving aromatic or conjugated hydrophobic atoms (*P_dASA_Neg_Hydrophobic*, Table 3). The additional two terms in model 1 are very similar to the ones already present in model 2, and they add very little information. A smaller number of variables reduces the risk of overfitting and is expected to lead to better general predictivity,[35] which is why we prefer model 2.

Neural Networks. The scoring functions developed using linear modeling techniques performed well, and the inclusion of quadratic dependencies between descriptors and biological activity in the GA optimization procedure did not increase their predictivity. Indeed, the best models were obtained using simple stepwise linear regression. Therefore, it was interesting to see whether better scoring functions could be obtained through the application of machine learning techniques, such as neural networks.

Neural networks perform optimally with large numbers of data points. Especially with a relatively small data set and large numbers of descriptors the risk of overtraining exists. For this reason we kept the number of descriptors as low as possible, retaining the most significant ones only, and we carefully monitored the networks for signs of overtraining.

The best 200 networks, according to their predictions on the external test set, are analyzed. Their r^2_{pred} values are in the 0.603−0.759 range (RMSE: 1.07−1.35 log units), and its distribution is depicted in Figure 4A. Overall, these results are slightly better than the ones obtained using linear modeling techniques with a total of 40 models displaying r^2_{pred} values greater than 0.725 and RMSE below 1.12 pK_d units. The use of an ensemble of neural networks for prediction purposes is a popular technique to avoid the occurrence of chance correlation.[36] The average predictions for the test set, as computed by the ensemble of the best 40 networks, are plotted against the corresponding observed biological activities in Figure 4B.

The contribution of each independent variable in a neural network cannot be assessed directly, as opposed to the previous scoring functions that were developed using linear regression and genetic algorithms. This is a clear disadvantage because a rational interpretability of the results is not possible, and therefore neural networks are often classified as "black boxes".[37] A possible way to assess the importance and the influence of each input descriptor on the overall network performance (i.e., the ability to learn the hidden relationships between input (descriptors) and output (affinities)) is to iteratively alter the number and type of descriptors.[37] This helps to understand the contribution of the independent variables to the predictive power of the resulting

Table 4. r^2_{pred} Statistics for Different Distributions of 200 Neural Networks Built Using Different Descriptor Sets

descriptors employed during training	average	median	SD	max	min
all	0.711	0.689	0.037	0.759	0.603
all minus most frequent set (Table 2)	0.349	0.346	0.025	0.489	0.287
most frequent set	0.699	0.691	0.042	0.745	0.619
most frequent set minus *Cntc*	0.596	0.589	0.033	0.629	0.507
most frequent set minus *Pslv*	0.673	0.659	0.029	0.702	0.602
most frequent set minus *Intl*	0.678	0.671	0.015	0.710	0.640
most frequent set minus *Tors*	0.670	0.662	0.036	0.699	0.619
most frequent set minus *L_dASA_Pol*	0.665	0.553	0.023	0.694	0.622
most frequent set minus *P_dASA_Neg_Hydrophilic*	0.688	0.680	0.034	0.717	0.639
most frequent set minus *L_dASA_Apo*	0.669	0.656	0.031	0.701	0.627
most frequent set minus *P_dS*	0.607	0.599	0.017	0.632	0.586
most frequent set minus *P_dASA_Neg_Hydrophobic*	0.672	0.668	0.024	0.708	0.633
most frequent set minus *Bump*	0.681	0.673	0.026	0.720	0.649
most frequent set minus *Intl, Tors*	0.587	0.581	0.032	0.622	0.536

network, and a series of tests was carried out accordingly. In the first test, all 10 descriptors that were most frequently observed in the linear scoring functions (Table 2) were removed from the input set. The 4 principal components, calculated from the remaining 15 descriptors, were used to train the networks as explained before in the Methods section. The resulting neural networks have r^2_{pred} values in the 0.287−0.489 range (Table 4), a clear deterioration of the predictive power. This indicates that the descriptors most frequently occurring in the linear models (Table 2) are indeed important for correct prediction of the binding affinity.

To determine the influence on the biological activity of each of these 10 descriptors by itself, one after the other was omitted from the input set, and the resulting retrained networks were used to predict the activities of the test set. Table 4 shows the results obtained from this procedure. For comparison, the results obtained with the most frequent set only, with all 25 descriptors, and with the 15 least important descriptors are provided as well.

Neural networks that are based only on the set of most frequent descriptors yield predictivity comparable to the ones that use the entire set of descriptors (Table 4). This is a further indication that those variables are most relevant in describing the receptor−ligand interactions. Neural networks built with more than the 25 uncorrelated descriptors (i.e., those with pairwise correlation coefficients < 0.7) exhibited lower predictive power than those built with the 10 most frequent descriptors listed in Table 2. Interestingly, it appears that the networks are not severely affected by the absence of any single descriptor, as the predictions for the test set are not much lower than those obtained with the full descriptor set (Table 4). The largest drop in the predictive power of the networks is observed when either *Cntc* or *P_dS* is removed (Table 2). In these cases, the best networks in the ensembles display r^2_{pred} values of 0.629 and 0.632, depending on whether *Cntc* or *P_dS* are omitted, respectively. *Cntc* accounts for the occurrence of favorable interactions, while *P_dS* describes the conformational entropy change of the protein upon binding. These descriptors are obviously important in describing binding. Two terms reflect the conformational entropy change of the ligand during the binding: *Tors* and *Intl*. When both are removed simultaneously, the predictive power of the networks is negatively affected. However, either of them seems to be sufficient to characterize the entropic contribution, as the average r^2_{pred} is comparable when either of them is not included (Table 4).

Comparison with Other Scoring Functions. The main aims of this paper are to illustrate how an existing scoring function containing interaction terms can be improved by inclusion of additional descriptors and to evaluate the consequences of such a procedure. Although the scoring functions described in this work are parametrized on crystal structure complexes, they are intended for use in docking experiments. QXP is the docking program that we use most heavily, predominantly in conjunction with its built-in pI scoring function. As a consequence, we are primarily interested in improving on the existing QXP-pI scoring function by combining its interaction terms with additional descriptors. Although it is of interest to compare the resulting functions with other scoring functions as well, this is beyond the scope of this paper.

To assess the performance of our scoring functions, they were benchmarked against related functions. Our scoring functions contain descriptors generated with both QXP-based and homegrown descriptors. Therefore, QXP's built-in pI scoring function and the N-way PLS model described in the methods section (which is built with Cerius2 and QXP descriptors) were used as general benchmarks, while a linear QXP descriptor model built only with the four terms (*Cntc*, *Pslv*, *Bump*, and *Intl*, see Figure 3B) that are most frequently used by the best 200 linear and GA models QXP was employed to determine the added value of the homegrown descriptors. Prior to calculating the QXP pI prediction errors and the correlation with experimental pK_d test set data, the pI function was calibrated by a linear fit of pI scores to experimentally observed pK_d values for the training/validation set using Cerius2.[29]

Table 5 compares the test set prediction quality of the three scoring functions developed in the present study (the 9 and 7 descriptor linear models and the Neural Network ensemble) with that of the benchmark functions (QXP pI, QXP descriptor, and N-way PLS models). The three new scoring functions perform much better than the QXP pI and descriptor models and somewhat better than the N-way PLS model. The introduction of the new homegrown descriptors significantly enhances the predictivity of the resulting models, as evidenced by the comparison with the QXP descriptor model. This indicates that this new class of descriptors, based on solvent accessible surface area changes, captures important aspects of the binding event.

Pose Prediction Accuracy. The scoring functions presented here were derived from the energy-minimized structures of complexes available in the PDB.[20] They show good

Table 5. Comparison of the Predictive Power of the Three Scoring Functions Developed in the Present Study and the Benchmark Functions

scoring function	r^2_{pred}	RMSE (pK_d units)
new models in this work		
model 1 (9 descriptors)	0.741	1.09
model 2 (7 descriptors)	0.712	1.16
neural network ensemble	0.737	1.10
benchmark models		
QXP pI model: 3.61 + 0.54 * pI	0.470	1.55
QXP descriptor model: 3.28 + 0.074 * *Cntc* −0.098 * *Pslv* −0.26 * *Bump* −0.093 * *Intl*	0.560	1.42
N-way PLS model	0.696	1.21

Table 6. Pose Prediction Results When Considering the Top-Ranked Conformation Only, Expressed in Terms of the Percentage of Compounds Whose Highest-Scoring Conformation Had an RMSD < 1.0 and < 2.0 Å, Respectively

scoring function	RMSD < 1.0 Å	RMSD < 2.0 Å
model 1 (9 descriptors)	32.5	47.9
model 2 (7 descriptors)	33.6	49.3
neural network ensemble	33.3	46.1
QXP pI model	55.7	79.8
QXP descriptor model	46.8	67.7

predictive power when applied to an external set, and they can therefore be employed to predict the biological affinity of any receptor−ligand complex. However, it was also of interest to evaluate the scoring functions in terms of their ability to score different poses of the same compound and to successfully rank the experimentally observed binding mode at the top. Table 6 displays the percentage of the compounds that were correctly docked, when considering the conformation ranked at the top by the different scoring functions. According to the stricter definition of docking accuracy (i.e., RMSD < 1.0 Å), a lower percentage of correct poses is found by all the scoring functions analyzed (Table 6). The QXP pI affinity score yielded the best results with 80% of its top ranked conformations having an RMSD < 2 Å. The QXP descriptor model performed second best with 68% of the top-scoring poses having an RMSD < 2 Å. All 3 new scoring functions developed here showed low pose prediction ability when compared to QXP, with linear 7-descriptor model 2 yielding slightly better results than its 9-descriptor counterpart (model 1) and than the ensemble of the best 40 neural networks.

The scoring functions developed in this study are clearly better in predicting the biological affinity than QXP (Table 5). However, the QXP pI scoring function is superior in the prediction of the correct pose (Table 6). The QXP descriptor model represents an improvement in terms of affinity prediction when compared to the QXP pI model (Table 5). However, its ability to identify the correct pose is reduced (Table 6). It is commonly accepted that the generation of the "correct" pose is a prerequisite for an accurate prediction of the biological affinity.[19] Nevertheless, the present study indicates that the pose and affinity prediction abilities of the scoring functions analyzed are anticorrelated. The scoring functions developed here were derived from "correct" poses (i.e., experimentally observed ones). No "incorrect" poses were used during the training of the scoring functions. Therefore one could argue that the molecular features associated with an incorrect pose were underrepresented during the training, and the resulting scoring functions were unable to learn how to discriminate between correct and incorrect poses. This could provide an explanation for the poor results obtained with respect to pose prediction. The scoring function employed by QXP had been derived by cominimizing the error in both pose and affinity predictions on a data set containing both crystal structures and docked ligands [Colin McMartin, personal communication]. The resulting prediction quality represents a tradeoff between the two main objectives (i.e., pose and affinity prediction). Therefore, it is not surprising that, at least for the data set employed here, the pose prediction results obtained with QXP are significantly better than the other scoring functions, whereas the affinity prediction is worse. It is conceivable that the N-way PLS model (which also uses non-native poses in the fitting procedure) yields better pose-prediction power on top of its high-quality affinity prediction, but this has yet to be established.

In a prospective docking experiment one may not know upfront which scoring function will work best, so the question is how to choose the "right" scoring function from a pool of functions as developed in this study. For linear models we would suggest a tradeoff between high predictivity and low number of terms, as was done here for the model containing 7 descriptors. Another—not yet explored—option would be to choose a number of scoring functions of high predictivity with different characteristics, such as number of terms and weights or types of terms, and to use a consensus scoring approach with these selected functions. For neural networks the safest option is to choose an ensemble, as was done in the present study.

CONCLUSIONS

The current study presents the development of novel descriptors and novel scoring functions that combine these descriptors with force field-like terms such as those implemented in the program QXP. The new descriptors account for the changes in solvation and entropy upon binding, both for the proteins and the ligands. The scoring functions were derived using regression-based approaches. The linear scoring functions exhibit good predictive power when evaluated on the external test set. Moreover, their terms have physical meaning, and they can help understand the energetic requirements for binding. Inclusion of the new descriptors led to significant improvements in affinity prediction compared to scoring functions that are based exclusively on QXP force field-like terms. With the current data set, the neural networks offer little advantage over the linear models. However, when larger, more diverse data sets become available, neural network approaches may become increasingly useful.

Although we obtained good results in terms of affinity prediction, the same scoring functions performed less satisfactorily in the prediction of crystallographic binding modes. This observation indicates that our understanding and mathematical description of the energetics of binding leave room for improvement. One may, therefore, consider using one scoring function specifically designed for pose prediction and another one developed for affinity prediction.

The new scoring functions described here are promising, but they need to be tested with additional experimental data

to further elucidate their predictive ability and to validate them as generally applicable tools for the prediction of binding affinity.

REFERENCES AND NOTES

(1) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand−receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.

(2) Martin, Y. C. Challenges and prospects for computational aids to molecular diversity. *Perspect. Drug Discov. Design* **1997**, *7*−*8*, 159−172.

(3) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.

(4) Zwanzig, R. W. High-temperature equation of state by a perturbation method. *J. Chem. Phys.* **1954**, *22*, 1420−1426.

(5) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de Novo design method based on simple, fast, and accurate free energy estimates .1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.

(6) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. BLEEP − Potential of mean force describing protein−ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.* **1999**, *20*, 1177−1185.

(7) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(8) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(9) Mark, A. E.; van Gunsteren, W. F. Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J. Mol. Biol.* **1994**, *240*, 167−176.

(10) Williams, D. H.; Maguire, A. J.; Tsuzuki, W.; Westwell, M. S. An analysis of the origins of a cooperative binding energy of dimerization. *Science* **1998**, *280*, 711−714.

(11) Reyes, C. M.; Kollman, P. A. Structure and thermodynamics of RNA-protein binding: using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *J. Mol. Biol.* **2000**, *297*, 1145−1158.

(12) Kuhn, B.; Kollman, P. A. A ligand that is predicted to bind better to avidin than biotin: Insights from computational fluorine scanning. *J. Am. Chem. Soc.* **2000**, *122*, 3909−3916.

(13) Wang, J. M.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.* **2001**, *123*, 5221−5230.

(14) 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.

(15) 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.

(16) Wang, R.; Liu, 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) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(18) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(19) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided. Mol. Des.* **2000**, *14*, 731−751.

(20) 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.

(21) Breslauer, K. J.; Remeta, D. P.; Chou, W. Y.; Ferrante, R.; Curry, J.; Zaunczkowski, D.; Snyder, J. G.; Marky, L. A. Enthalpy−entropy compensations in drug−DNA binding studies. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 8922−8926.

(22) 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 Deliv. Rev.* **2001**, *46*, 3−26.

(23) McMartin, C.; Bohacek, R. S. QXP: powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.

(24) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.

(25) Murphy, K. P.; Freire, E. Thermodynamics of structural stability and cooperative folding behavior in proteins. *Adv. Protein Chem.* **1992**, *43*, 313−361.

(26) Lee, K. H.; Xie, D.; Freire, E.; Amzel, L. M. Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. *Proteins* **1994**, *20*, 68−84.

(27) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(28) Sybyl, version 6.7.1, Tripos Inc., St. Louis, MO, 2000.

(29) Cerius2, version 4.7 ccO, Accelrys Inc., San Diego, CA, 2001.

(30) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantiative structure−activity-relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(31) MATLAB, version 12.0, The MathWorks, Natick, MA, 2003.

(32) Hagan, M. T.; Menhaj, M. B. Training Feedforward Networks with the Marquard Algorithm. *IEEE Trans. Neural Netw.* **1994**, *5*, 989−993.

(33) Filikov, A. Enrichment factors in molecular docking: common misconceptions. *Abstr. Pap. Am. Chem. Soc.226* **2003**.

(34) Frank, H. S.; Evans, M. W. Free volume and entropy in condensed systems. *J. Chem. Phys.* **1945**, *13*, 507−532.

(35) Good, I. J. What are degrees of freedom? *Am. Stat.* **1973**, *27*, 227−228.

(36) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural-network studies 0.1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.

(37) Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, 1995.

CI0499626