# Assessing Scoring Functions for Protein−Ligand Interactions

Philippe Ferrara,[†,§] Holger Gohlke,[†,§] Daniel J. Price,[†] Gerhard Klebe,[‡] and Charles L. Brooks III*[,†]

*Department of Molecular Biology (TPC6), The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, and Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany*

*Received October 1, 2003*

An assessment of nine scoring functions commonly applied in docking using a set of 189 protein−ligand complexes is presented. The scoring functions include the CHARMm potential, the scoring function DrugScore, the scoring function used in AutoDock, the three scoring functions implemented in DOCK, as well as three scoring functions implemented in the CScore module in SYBYL (PMF, Gold, ChemScore). We evaluated the abilities of these scoring functions to recognize near-native configurations among a set of decoys and to rank binding affinities. Binding site decoys were generated by molecular dynamics with restraints. To investigate whether the scoring functions can also be applied for binding site detection, decoys on the protein surface were generated. The influence of the assignment of protonation states was probed by either assigning "standard" protonation states to binding site residues or adjusting protonation states according to experimental evidence. The role of solvation models in conjunction with CHARMm was explored in detail. These include a distance-dependent dielectric function, a generalized Born model, and the Poisson equation. We evaluated the effect of using a rigid receptor on the outcome of docking by generating all-pairs decoys ("cross-decoys") for six trypsin and seven HIV-1 protease complexes. The scoring functions perform well to discriminate near-native from misdocked conformations, with CHARMm, DOCK-energy, DrugScore, ChemScore, and AutoDock yielding recognition rates of around 80%. Significant degradation in performance is observed in going from decoy to cross-decoy recognition for CHARMm in the case of HIV-1 protease, whereas DrugScore and ChemScore, as well as CHARMm in the case of trypsin, show only small deterioration. In contrast, the prediction of binding affinities remains problematic for all of the scoring functions. ChemScore gives the highest correlation value with $R^2 = 0.51$ for the set of 189 complexes and $R^2 = 0.43$ for the set of 116 complexes that does not contain any of the complexes used to calibrate this scoring function. Neither a more accurate treatment of solvation nor a more sophisticated charge model for zinc improves the quality of the results. Improved modeling of the protonation states, however, leads to a better prediction of binding affinities in the case of the generalized Born and the Poisson continuum models used in conjunction with the CHARMm force field.

## 1. Introduction

Molecular recognition is a problem of fundamental importance in biology. An understanding of its principles would result in a more efficient application of medicinal chemistry. In particular, it is important for drug discovery to capture the physical principles responsible for the recognition of a drug by its protein target. The ever growing amount of structural information makes computer-aided structure-based ligand design methods a useful alternative strategy to experimental high-throughput screening to find novel leads in a drug development program.[1]

Computer-aided structure-based methods are aimed at predicting the binding mode of a ligand in the binding site of a protein or any molecular target and at obtaining an estimate of the binding affinity. These methods involve two computational steps: docking and scoring. In the docking step, multiple protein−ligand configurations, called *poses*, are generated. Several current docking programs have the ability to generate poses close to the native structure (usually an rmsd of ≤2 Å is accepted as close) in many cases.[2−7] Then, a scoring function is used to calculate the affinity between the receptor and the ligand for each pose. There are several requirements a useful scoring function should satisfy. First, the poses must be ranked correctly; i.e., those that resemble most closely the experimental structures should score best. In addition, if multiple ligands are docked, their binding free energies need to be ranked accurately. In a virtual screening simulation, weak binders should be distinguished from nonbinders. Finally, a scoring function must be sufficiently fast to be applied in a docking algorithm. This makes it almost impossible to use methods that require the generation of a correctly weighted ensemble of conformations to obtain the free energy of binding, although calculations of binding affinities based on algorithms that identify the most stable conformations of the free and bound species have been reported.[8,9] These studies highlight

* To whom correspondence should be addressed. Phone: (1 858) 784 8035. Fax: (1 858) 784 8688. E-mail: brooks@scripps.edu.
[†] The Scripps Research Institute.
[§] These authors contributed equally.
[‡] Philipps-Universität Marburg.

10.1021/jm030489h CCC: $27.50 
Published on Web 05/04/2004

the important role of the configurational entropy, which is not explicitly taken into account in calculations using the single-structure approach.

Scoring functions can be grouped into three classes: force-field-based, knowledge-based, and empirical scoring functions (see refs 10 and 11 for a review). Force-field-based scoring functions apply classical molecular mechanics energy functions. They approximate the binding free energy of protein−ligand complexes by a sum of van der Waals and electrostatic interactions. Solvation is usually taken into account using a distance-dependent dielectric function, although solvent models based on continuum electrostatics have been developed.[12,13] Nonpolar contributions are usually assumed to be proportional to the solvent-accessible surface area. A drawback is that the energy landscapes associated with force-field potentials are usually rugged, and therefore, minimization is required prior to any energy evaluation.

Empirical scoring functions estimate the binding free energy by summing interaction terms derived from weighted structural parameters. The weights are obtained by fitting the scoring function to experimental binding constants of a training set of protein−ligand complexes. The archetypical scoring function pioneered by Böhm consists of five contributions, which represent hydrogen bonds, ionic and lipophilic interactions, and the loss of external and configurational entropy upon binding.[14,15] The main drawback of empirical scoring functions is that it is unclear whether they are able to predict the binding affinity of ligands structurally different from those used in the training set.

Knowledge-based scoring functions represent the binding affinity as a sum of protein−ligand atom pair interactions. These potentials are derived from the protein−ligand complexes with known structures, where probability distributions of interatomic distances are converted into distance-dependent interaction free energies of protein−ligand atom pairs using the "inverse" Boltzmann law.[16] However, the structures deposited in the Protein Data Bank (PDB) do not provide a thermodynamic ensemble at equilibrium, and therefore, a knowledge-based potential should be considered as a statistical preference rather than a potential of mean force. A key ingredient of a knowledge-based potential is the reference state, which determines the weights between the various probability distributions. Recently, several approaches to derive these potentials have been proposed.[17−20] They differ in their definition of the reference state, the protein and ligand atom types, and the list of protein−ligand complexes from which they were extracted.

No scoring function performs in a satisfactory way, which led to a pragmatic compromise, the so-called consensus scoring approach. Here, several scoring functions are combined and only those poses that receive high scores by two or more scoring functions are considered favorable.[21] It was shown that this method yields a large reduction of false positives when applied either to choosing the ligands with the lowest binding free energies among a set of ligands or to selecting the best poses between different docked configurations of a particular ligand.

Here, we present an assessment of nine scoring functions, most of which are implemented in widely used docking programs. The scoring functions cover the three classes described above: CHARMm[22] and DOCK-chemical[5] represent force-field-based methods; ChemScore[23] and the potentials implemented in GOLD[3,24] and AutoDock[6,25,26] are empirical scoring functions; DrugScore[20] and PMF[19] are knowledge-based potentials. Finally, DOCK-contact counts the number of contacts between the ligand and the receptor. The study was performed on data from the Ligand−Protein Database (LPDB), which is World Wide Web accessible (http://lpdb.scripps.edu) and comprises 189 protein−ligand complexes.[27] This data set corresponds to 49 different receptors with both high-resolution structure (2.1 Å on average) and known experimental binding affinity. In this respect, the current study is the most comprehensive comparison of scoring functions reported so far.

Several studies analyzing the performance of docking programs in combination with various scoring functions have been reported in the case of virtual screening applications.[21,28,29] In such a study, the foremost goal is to identify true hits in a database of mainly nonbinders. This point is not addressed in this study, but we were interested in correctly ranking ligands already known to bind, which is of primary interest in lead optimization. In addition, to separate the docking problem from the scoring problem, nearly 100 decoys have been constructed for each protein−ligand pair whose deviations from the crystal structure represent a continuous spectrum in the neighborhood of the binding site. This set was then rescored by all functions. Finally, misdocked structures far from the binding site were generated to test whether the scoring functions can successfully detect binding pockets.[27] The present study was motivated by earlier work by Vieth et al. to identify key features of binding energy landscapes necessary and sufficient for the development of successful docking and scoring algorithms.[30,31] We note that a study similar in spirit has been reported very recently by Wang et al. on a data set of 100 complexes, which assesses 11 scoring functions in their ability to recognize native poses among a set of decoys and to predict binding affinities.[32] The next paragraphs explain to which extent our study goes beyond theirs.

The experimental conditions, such as pH or salt concentration, under which crystallization and the binding assay are performed and the packing in the crystal form can have a profound impact on the binding mode and the affinity of the ligand. This has been recently revealed in a study of trypsin crystals,[33] which shows the occurrence of protonation and crystal form dependent binding modes. For a series of aliphatic cyclic ureas bound to HIV-1 protease, the binding energies computed by the Poisson equation significantly depend on the protonation state of the two active site aspartic acids.[9] Therefore, it may be anticipated that it is important to correctly model the protonation state of the ligand and the receptor, at least for the scoring functions that make use of partial charges. It is worth noting that the pH of crystallization of the complexes deposited in the LPDB is as low as 3.0 and as high as 8.5. We investigated this issue by adjusting, for some

**Table 1.** Description of the Data Sets

| no. | data set | no. of complexes | p$K_i$ range | $R^2$ [a] | protonation states | zinc charge model |
|---|---|---|---|---|---|---|
| 1a | all | 189 | 12 | 0.36 | modified | ab initio |
| 1b | all | 189 | 12 | 0.36 | standard | ab initio |
| 2 | all/ChemScore + AutoDock | 116 | 11 | 0.35 | modified | ab initio |
| 3 | aspartic protease | 52 | 7 | 0.05 | modified | NA |
| 4 | oxidoreductase | 37 | 8 | 0.23 | modified | ab initio |
| 5 | serine protease | 25 | 7 | 0.81 | NA | NA |
| 6 | metalloprotease | 13 | 10 | 0.58 | modified | ab initio |
| 7 | immunoglobulin | 10 | 6 | *0.50* | modified | NA |
| 8 | lyase | 10 (8[b]) | 8 (3[b]) | 0.18 | NA | ab initio |
| 9 | L-arabinose binding protein | 9 | 2 | 0.16 | NA | NA |
| 10 | mhc | 7 | 2 | 0.01 | NA | NA |
| 11 | others | 26 | 11 | 0.09 | modified | ab initio |

[a] Square of the correlation coefficient ($R^2$) between the experimental binding affinities and the logarithm of the ligand molecular weights. $R^2$ values in italic denote an anticorrelation. [b] Without 1avn and 1ebg.

of the complexes, the protonation state according to experimental evidence when available.

Solvation plays an important role in molecular recognition, and accurately incorporating solvent effects in docking approaches represents a major challenge. Solvation models based on continuum electrostatics have been implemented in docking[13,34] and de novo ligand design programs.[12] In a virtual screening simulation against three receptors, it was shown that including ligand solvation improves the ranking of known ligands and leads to low-energy compounds with net formal charges consistent with those of known ligands.[34] Hence, in this study, we analyzed in depth the role of electrostatics in scoring by using various solvent models in conjunction with the CHARMm force field.

Zinc in the binding sites of metalloproteins performs essential biological functions and often contributes considerably to the binding affinity of small-molecule ligands. However, it is notoriously difficult to evaluate metal−ligand interactions. More specifically, different charge sets may significantly influence binding energies computed by force-field scoring functions. Recently, it has been shown that the use of partial atomic charges determined by semiempirical calculations leads to a better recognition of true ligands in database docking.[35] Thus, we investigated the influence of the zinc charge model by comparing scoring results obtained for two sets of charges. First, zinc is modeled as a +2 ion; second, the charge transfer between the zinc and its coordinating protein groups is taken into account, leading to a Zn charge below +2 (see section 2).

It is obvious that docking algorithms that treat the receptor as rigid will encounter problems in docking a ligand into a binding site if the latter undergoes a significant conformational change upon binding. It is unclear to what extent small changes in the receptor compromise the accuracy of the scoring functions. From the point of view of virtual screening applications, it is common practice to preferentially use coordinates of a receptor obtained with a ligand analogue. Studies that assess the performance of scoring functions for cross-docking are rare. Murray et al. carried out docking and cross-docking experiments for three fairly rigid enzymes (thrombin, thermolysin, and neuraminidase) using the program PRO_LEADS, which makes use of the scoring function ChemScore.[36] Docking and cross-docking simulations were also performed for a set of 34 protein−ligand complexes, which represents 17 pairs of complexes of the same protein bound to two different ligands.[37] Binding energies were computed by the DOCK energy potential and PMF. In both studies, the decrease in performance was significant. We investigated the influence of the receptor structure by generating cross-decoys for six trypsin and seven HIV-1 protease complexes. We selected these two enzymes because their receptors display a significantly different plasticity. In the former case, the receptor is fairly rigid, whereas it is more flexible in the latter.

## 2. Methods

In this section, we provide details of data preparation and briefly outline the scoring functions used in this study.

**2.1. Data Preparation.** A detailed description of the selection and preparation of complexes in the LPDB is given in ref 27. We focus on the modifications since the first release of the database. To date, the LPDB comprises 189 complexes, which correspond to 49 different receptors and cover a range of binding affinities of 12 orders of magnitude (Table 1).

Recently, the importance of using error-free experimentally determined structural data for the development and evaluation of scoring functions has been noted.[38,39] In our case, using minimized crystal geometries alleviates the problem of short-range interactions sometimes found in the crystal. Furthermore, a survey of our data reveals that in around 15% of the cases, crystal contacts are found to the ligand. To investigate the influence of crystal contacts on the outcome of the prediction of binding affinities, we also used decoys for affinity prediction in addition to the minimized crystal structure. However, we did not find any significant change in the results. In the case of identification of near-native poses, some of the crystal geometries may not be relevant as references due to the presence of crystallographically related contacts. This point is not addressed in this study.

**2.1.1. Protonation State Assignment.** A first set of structures was generated by using default values to determine the protonation states of the titratable groups. As such, amines were protonated, carboxylate groups were negatively charged, and hydroxyl groups are considered to be neutral. Imidazole rings were considered neutral, with the hydrogen on the $\delta$ nitrogen, except when a hydrogen bond involving the N$\epsilon$ as a donor could be formed. Then, the hydrogen was placed on the $\epsilon$ nitrogen.

As mentioned, the ligand binding mode and the binding affinity can be strongly pH-dependent, which led us to build a second set of conformations by adjusting the protonation states of ligands and protein residues according to either experimental evidence or chemical intuition. This information (when available) was extracted from the literature describing the crystallization of the complex structure. For example, it was suggested that most of the ligands in the LPDB that are bound to cytochrome *c* peroxidase should be protonated.[40] It is worth noting that the pH of crystallization for this series of ligands is close to 4.5. In some of the thermolysin complexes, a contact between the ligand and the receptor involves two carboxylate oxygens separated by a distance of around 2.8 Å. In this case, a hydrogen atom was placed on the oxygen belonging to the receptor. In the case of the aspartic protease family, experimental and theoretical studies show that there is not a consensus choice for the protonation state of the two catalytic aspartic acids in the binding site.[41−49] Even when it is assumed that the catalytic aspartates are in a monoprotonated form, the position of the proton remains unclear. An influence of the chemical nature of the inhibitor and the pH of the experiment on the ionization state of the active site has been described.[50] Since the pH of crystallization of most of the aspartic protease complexes in the LPDB ranges between 4.5 and 5.5, the two catalytic aspartates are probably not in the form prevalent at physiological pH. In two cases (1hpx and 1hbv), the proton was placed on the outer oxygen of Asp25A, according to experimental evidence.[51,52] For the remaining aspartic protease complexes, we made the somewhat arbitrary choice to place a proton on one of the two inner oxygens of the carboxylate groups of the two aspartic acids.

**2.1.2. Metal−Ligand Interactions.** Modeling interactions between zinc and surrounding residues as solely ionic leads to assigning a +2 charge for Zn. However, charge transfer between zinc and its coordinating protein residues reduces the zinc charge, which may influence the outcome of binding affinity calculations. To investigate this influence, the charge on the zinc and neighboring protein atoms was computed by Hartree−Fock calculations at the HF/6-31G*//HF/6-31G* level using Gaussian.[53] The following coordinated groups were included in the calculations: for a histidine, all atoms until the C$\gamma$; for a cysteine until the C$\beta$; for an aspartate until the C$\beta$; for a glutamate until the C$\gamma$. Ligand atoms were not taken into account in the calculations. The bond between the zinc and its coordinated atoms was considered to be covalent. Therefore, for the complexes having a zinc bound to the ligand, only those decoys were used for which the distance between the zinc and its closest ligand atom was less than the corresponding average distance extracted from the Cambridge Structural Database augmented by 0.25 Å.[54]

**2.1.3. Decoy Generation.** After these changes, the crystal structures were reminimized, and for each minimized protein−ligand pair, two sets of decoys have been generated.[27] First, starting from the minimized crystal structure, we used the replica method in CHARMM to move simultaneously 25 copies of the ligand within the rigid binding site. One-thousand steps of Langevin molecular dynamics simulations at 300 K with NOE-like restraints pulled the ligand away from the minimized complex. These restraints are similar to the distance restraints built from NOE data and used in protein determination by NMR. The center of mass of the ligand was not restricted. At the end of each move, the 25 ligand positions were minimized using the steepest descent method until the energy gradient was less than 0.05 kcal mol$^{-1}$ Å$^{-1}$ (with a maximum number of steps set to 1000), followed by conjugate gradient minimization with the same termination criterion. We will refer to these decoys as "binding site decoys", since they represent a quasi-continuum of decoys within the binding site. Nearly 100 binding site decoys were generated for each protein−ligand complex. It should be noted that the binding modes of the decoys might not be as diverse as in a docking simulation with different random initial conformations. The distribution of the root-mean-square deviation from the crystal structure ($rmsd_N$) of the binding site decoys is shown in Figure 1 with the corresponding cumulative probability distribution. It can be seen that nearly half of the binding site decoys have an $rmsd_N$ below 3.5 Å. In the generation of the second decoy set, a spherical grid comprising 1820 points was built around the protein. All points overlapping with the protein were removed. The ligand was translated to each point of the grid. For each translation, 3−15 conformations were generated from a random rotation of the ligand. The ligands were minimized in the same way as the binding site decoys. These decoys will be referred to hereafter as "surface decoys".

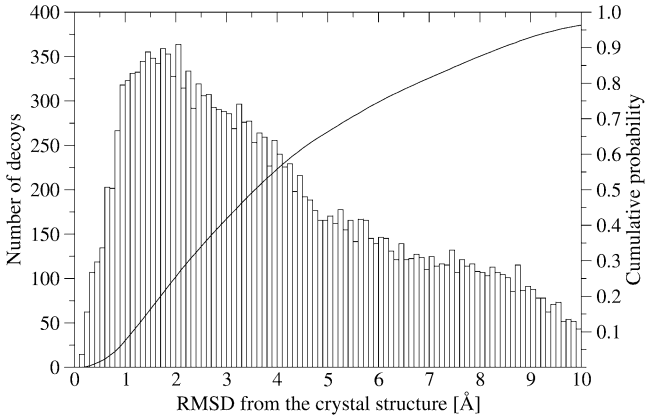


**Figure 1.** Histogram of the root-mean-square deviation from the crystal structure of the binding site decoys (left *y*-axis) with the corresponding cumulative probability distribution (right *y*-axis).

Minimization of the decoys with the CHARMm force field may introduce a bias, which could favor the CHARMm scoring function. To investigate this effect, we reminimized all the decoy configurations with the Tripos force field of the CScore module in SYBYL and rescored them by all scoring functions. We did not find any significant change in the recognition rates of near-native configurations.

**2.1.4. Cross-Decoy Generation.** From the LPDB, we selected six trypsin (1tni, 1tng, 1tpp, 1ppc, 1pph, and 3ptb) and seven HIV-1 protease (1ajv, 1gno, 1hih, 1hps, 1htf, 1hvi, and 2upj) complexes and generated cross-decoys for all the protein−ligand pairs within a particular enzyme. To generate a "native" conformation for a particular protein−ligand pair, the two binding sites

were first superimposed and the ligand was then minimized in the rigid binding site of the related protein using the conjugate gradient method until the energy gradient was less than 0.5 kcal $mol^{-1}$ $Å^{-1}$ (with a maximum number of steps set to 1000). Binding site decoys were generated as described above for the natural ligands, and no surface decoys were constructed. In total, cross-decoys corresponding to 72 protein−ligand combinations were generated.

**2.2. Scoring Functions.** An all-atom model was used for CHARMm, DOCK, and AutoDock. Partial charges for the protein were set up using the CHARMm force field.[22] A modified version of INSIGHT II[55] was used to assign the ligand partial charges, which makes use of a bond increment scheme. Some of the bond increments were modified to better reproduce the 6-31G* ESP charges (the modified bond increments are available upon request). CHARMm was evaluated either in vacuo or in conjunction with various solvation models: namely, the Poisson equation,[56] the generalized Born (GB) model,[57] and either a constant (CDIE) or distance-dependent dielectric function (RDIE). Implicit electrostatic solvation models approximate the energy associated with solvating a charged solute, represented by a low-dielectric medium, in a solvent modeled by a high-dielectric medium. Calculations based on the Poisson equation are usually considered to be the benchmark for continuum electrostatics. They were carried out using the CHARMM[58] PBEQ module, which makes use of grid-based finite difference techniques (we will subsequently refer to them as FDP). The grid spacing was set to either 0.25 Å ($FDP_{0.25}$) or 0.4 Å ($FDP_{0.4}$), the ionic strength was set to zero, and the solute dielectric boundary was defined as the Lee−Richards molecular surface.[59] The solvent and solute dielectric constants were set to 80 and either 1 or 4, respectively. A simpler and faster continuum dielectric approximation is the so-called generalized Born (GB) model, which approximates the reaction field by a Coulomb potential.[57] Numerical and analytical solutions to the GB model have been proposed.[60−62] Here, we used a recently developed analytical GB model in which a correction term to the Coulomb field approximation has been introduced in conjunction with a more accurate description of the molecular surface.[63,64] The model contains five parameters and utilizes a molecular volume built from a superposition of atomic functions. The calculations were performed for $\epsilon = 1$, and the solvation energies were scaled by $(\epsilon - 1/80)/(1 - 1/80)$ to obtain the energies for higher values of the dielectric constant. An even more simplified solvent model is based on a linear distance-dependent dielectric function (called hereafter RDIE, $\epsilon(r) = cr$, where $c$ is a constant that can range from 1 to 80 and $r$ is an atom−atom distance) to approximate the screening effects of the electrostatic interactions. For the CDIE/RDIE calculations, a sigmoidal switching function[58] between 11 and 14 Å was employed for both the van der Waals and electrostatic terms. For the GB calculations, the van der Waals and the electrostatic interactions were truncated between 24 and 26 Å with a switching function.

Three scoring functions, called contact, energy, and chemical scoring, are available in the DOCK 4.0.1 program suite.[5] The contact scoring function counts the number of heavy atom contacts between the ligand and the receptor. The energy score is based on the nonbonded interaction energies of the Weiner et al. force field.[65] AMBER charges were used for the protein, while Gasteiger−Marsili[66] atomic charges as implemented in SYBYL[67] were assigned to the ligands. A linear distance-dependent dielectric function ($\epsilon(r) = 4r$) was used for the Coulomb potential, and the cutoff for steric and electrostatic interactions was set to 10 Å. The chemical score is based on the energy score except that the attractive part of the van der Waals interaction is scaled depending on the interacting atoms. The gridded_score flag was turned off in DOCK to enable scoring as a continuous function.

DrugScore[20] and AutoDock[6,25,26] were evaluated by using their original implementation. The CScore (Consensus Score) module[68] implemented in SYBYL[67] was used to assess ChemScore,[23] Gold,[3,24] and PMF.[19] It should be noted that the scores calculated by the original scoring function and CScore can differ.[28] DrugScore and PMF are knowledge-based potentials, which were derived by using 1376 and 697 protein−ligand complexes, respectively, taken from the PDB. In DrugScore, 17 atom types were defined, whereas PMF is based on 16 protein atom types and 34 ligand atom types. A knowledge-based, solvent accessible surface area dependent solvation term is included in DrugScore. The GOLD scoring function is a sum of a hydrogen-bonding energy, a steric interaction energy (a 4-8 potential) between the ligand and the protein, and an internal energy for the ligand, which consists of a van der Waals energy and a torsional potential.[3,24] ChemScore is a regression-based scoring function, which uses contact terms for lipophilic and metal-binding contributions and a hydrogen-bonding term. ChemScore also includes a term that penalizes restriction of conformational degrees of freedom upon binding.[23] ChemScore was calibrated on a data set of 82 protein−ligand complexes. AutoDock (version 3.0) is also a regression-based scoring function, which consists of a van der Waals, an electrostatic, a hydrogen bonding, and a desolvation energy term, augmented by an entropic term that measures the loss of torsional degrees of freedom upon binding. The desolvation free energy is taken to be proportional to the volume around the atoms that are exposed to the solvent.[69] The scoring function was parametrized on a set of 30 protein−ligand complexes.[6] To evaluate the influence of the charge set, AMBER/Gasteiger−Marsili and CHARMm/INSIGHT II charges were assigned to the complexes, and only the results obtained by the AMBER/Gasteiger−Marsili charges will be discussed, since both alternatives yield very similar results.

Finally, we note that the binding energies $\Delta E$ were calculated according to $\Delta E = E_{complex} - E_{protein} - E_{ligand}$ and, therefore, include only interaction energies between the ligand and the protein. For the CHARMm force field, taking into account the intraligand energy leads to a small improvement in the recognition of near-native configurations and to similar correlation coefficients with the experimentally determined binding affinities (data not shown).

**2.3. Data Analysis.** The performance of the scoring functions is evaluated in terms of their ability (a) to

identify near-native ligand poses (≤2 Å) among a set of decoy structures and (b) to correctly rank different ligands with respect to their binding affinities. The latter ability depends on the former because it cannot be expected that correct binding affinities are obtained using misdocked protein−ligand configurations as the structural basis.[30] We did not investigate how often the minimized crystallographic structure scores best because it does not always represent the global energy minimum and is seldom generated by docking tools. Instead, we used the percentage of complexes with a root-mean-square deviation from the crystal structure of less than 2 Å of the best ranked structure as the criterion to evaluate the success of a potential to recognize near-native binding modes.

Progress in understanding ligand−protein interactions has been made by using ideas of the statistical energy landscape theory.[30,70,71] It has been suggested that the ruggedness of the binding energy surface can be associated with structural flexibility and different types of binding mechanisms.[7,70] This finding led us to assess the discriminative power of a particular scoring function, which measures its ability to discriminate between well-docked structures and misdocked structures. To do so, we used a criterion that has been defined previously in a study of the CHARMm force field as a scoring function for flexible docking.[30] It is based on the *Z* score, which is defined as

$$Z(E) = \frac{(E - \bar{E})}{\sigma} \quad (1)$$

where $E$ represents the binding energy of the ligand−receptor complex, $\bar{E}$ is the mean energy, and $\sigma$ is the standard deviation of the energy distribution. In this work, this energy distribution corresponds either to the binding energies of the well-docked conformations or to those of the misdocked conformations (see below). The discriminative power, DP, of a given scoring function is then defined as

$$\mathrm{DP} = \frac{1}{N}\sum_{i=1}^{N}(Z^{i,D}_{\min} - Z^{i,\mathrm{M}}_{\min})f_i \quad (2)$$

where $i$ refers to the different complexes and $N$ is the number of complexes. $Z^{i,\mathrm{D}}_{\min}$ and $Z^{i,\mathrm{M}}_{\min}$ represent the $Z$ scores for the lowest energy structure among the well-docked and misdocked conformations, respectively. $f_i$ is the fraction of the well-docked structures with $Z$ scores lower than those of the misdocked structures. The definition of well-docked and misdocked was taken as before,[30] i.e., the structures with a $\mathrm{rmsd_N}$ smaller than 2 Å and larger than 4 Å, respectively. A DP value of zero means no discriminative power, and the lower the value of DP, the more reliable is the energy function in finding relevant solutions.

Finally, we used the square of the Pearson correlation coefficient ($R^2$) to evaluate the ability of a particular scoring function to predict experimental binding affinities. For instance, an $R^2$ of 0.36 between the experimental binding affinities and the molecular weight of the ligands means that 36% of the variation in the logarithm of the binding potency can be explained by a variation in the molecular weight. To relate the scores calculated by a knowledge-based or a force-field-based scoring function to an absolute binding affinity, it would be necessary to scale all computed energies. This was not done in this study, and therefore, we do not provide the standard deviations from the observed affinities.

**2.4. Description of the Data Sets.** Table 1 lists the data sets that were used to analyze the scoring functions. The first two sets (1a, 1b) each comprise the whole database (189 complexes). However, they differ in the protonation state assigned to the ionizable groups. Most of the complexes used to calibrate ChemScore and AutoDock belong to the LPDB. Therefore, to better evaluate the predictability of these two regression-based scoring functions, we built set 2, where 69 complexes that belong to the training set of ChemScore and AutoDock were removed. In a drug design project, one is interested in either the relative binding free energy between two different ligands for the same receptor (affinity) or the relative binding free energy between two different receptors for the same ligand (specificity). Therefore, we also considered sets 3−11, where all 189 complexes were classified according to their receptor.

## 3. Results and Discussion

**3.1. Comparison of the Generalized Born Model with a Model Based on the Poisson Equation.** Recently, a new solution for the GB model has been described.[63,64] Yet the model has only been validated for different protein conformations and not for protein−ligand complexes. Since the solvation contribution to the binding free energy (which is on the order of 10−100 kcal $\mathrm{mol^{-1}}$) is determined as the difference between solvation free energies of the complex, the receptor, and the ligand (the first two usually being on the order of $10^3$−$10^4$ kcal $\mathrm{mol^{-1}}$), already small absolute errors in these solvation free energies lead to a large (relative) error in the predicted binding free energy. Furthermore, it has been questioned whether continuum solvent models parametrized to reproduce vacuum-to-water free energy data can be used for binding reactions without reparametrization.[72] Here, we present an extensive validation of this model based on a comparison with FDP results. The benchmark FDP solvation energies were obtained using a grid spacing of 0.25 Å. These calculations are computationally intensive, and therefore, we tested the accuracy of the GB method on a set of only 84 complexes. Table 2 shows that the GB energies for the complex, receptor, and ligand agree very well with the $\mathrm{FDP_{0.25}}$ energies with an average absolute error below 1% for the complex and receptor. The agreement is almost as good as between the $\mathrm{FDP_{0.25}}$ and $\mathrm{FDP_{0.4}}$ results. Figure 2 displays the corresponding binding free energies obtained by the GB, $\mathrm{FDP_{0.4}}$, and $\mathrm{FDP_{0.25}}$ models. The cancellation of errors is much better in the case of $\mathrm{FDP_{0.4}}$ than for the GB energies, leading to an rms error of 2.52 kcal $\mathrm{mol^{-1}}$ of $\mathrm{FDP_{0.4}}$ compared to $\mathrm{FDP_{0.25}}$, whereas in the case of GB an rms error of 31.83 kcal $\mathrm{mol^{-1}}$ is obtained. The square of the correlation coefficient between the GB and $\mathrm{FDP_{0.25}}$ binding energies is $R^2 = 0.58$. Removing the 12 endothiapepsin and 18 cytochrome *c* peroxidase complexes results in an improved regression with $R^2 = 0.76$, which, however, is still worse than that obtained in the $\mathrm{FDP_{0.4}}$/$\mathrm{FDP_{0.25}}$ case ($R^2 = 0.99$).

**Table 2.** Comparison of Electrostatic Energies Calculated by the GB and FDP Methods on a Data Set of 84 Complexes[a]

| | GB | | | $FDP_{0.4}$[b] | | |
|---|---|---|---|---|---|---|
| | complex | receptor | ligand | complex | receptor | ligand |
| absolute average[c] (%) | 0.51 | 0.60 | 2.10 | 0.48 | 0.49 | 1.88 |
| slope[d] (kcal mol$^{-1}$) | 0.997 | 0.997 | 1.011 | 1.004 | 1.004 | 1.009 |
| intercept[d] (kcal mol$^{-1}$) | 22.80 | 35.17 | 1.71 | −4.77 | −7.55 | −0.68 |
| $R^2$ [d] | 0.9999 | 0.9999 | 0.9993 | 0.9999 | 0.9999 | 0.9999 |

[a] Benchmark energies were obtained from FDP with a grid spacing of 0.25 Å ($FDP_{0.25}$). [b] FDP energies with a grid spacing of 0.4 Å. [c] Error (%) = $\langle|GB/FDP_{0.4} - FDP_{0.25}|/FDP_{0.25}\rangle$. [d] Slope, intercept, and square of the correlation coefficient $R$ of the least-squares fit line of the GB and $FDP_{0.4}$ to the $FDP_{0.25}$ energies.

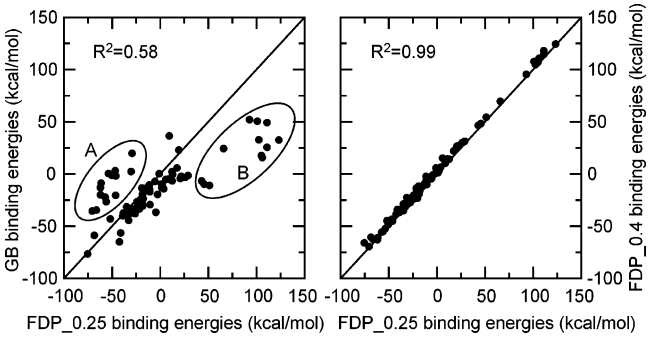


**Figure 2.** Correlation between the $FDP_{0.25}$ and GB binding energies (left) and between the $FDP_{0.25}$ and $FDP_{0.4}$ binding energies (right). $FDP_{0.25}$ and $FDP_{0.4}$ denote finite difference Poisson calculations with a grid spacing of 0.25 and 0.4 Å, respectively. The circles A and B comprise the cytochrome *c* peroxidase and endothiapepsin complexes, respectively.

### 3.2. Recognition of Near-Native Configurations.

In this section, we consider only binding site decoys to analyze the ability of the scoring functions to recognize near-native poses. In the forthcoming analyses, the results for the CHARMm force field were obtained by the RDIE method with a dielectric constant of 4 ($\epsilon(r) = 4r$). We will refer to them as CHARMm-4r. This choice allows a direct comparison with the DOCK-energy calculations, which were performed with the same treatment for the electrostatic interactions. The influence of the solvation models is discussed in the next section.

Figure 3 shows for different data sets the percentage of complexes for which the lowest energy decoy has an $rmsd_N$ of less than 2 Å. CHARMm-4r, DOCK-energy, ChemScore, DrugScore, and AutoDock reach a high rate of success with percentages ranging from 80% to 90%, whereas all the other functions yield recognition rates of 53% to 65%. A decomposition of the CHARMm potential energy values yields a recognition rate of 80% for the sets 1a and 1b if only the van der Waals potential is used. A receptor-based analysis shows that the metalloprotease and lyase sets yield the lowest recognition rates for most of the scoring functions. This might be due to the presence of a zinc in the binding site of these complexes.

All the scoring functions fail to recognize near-native configurations for the complexes 1cny, 1cnw, and 1avn. In all these cases, the ligand is significantly solvent-exposed and many decoys have a more favorable van der Waals interaction energy than in the crystal structure. It can be expected that these decoys should be disfavored by entropic and solvation effects, which are not correctly captured by the analyzed scoring functions, at least in these cases.

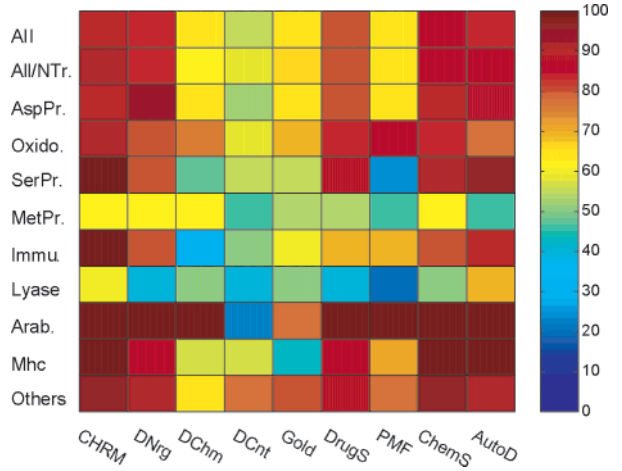


**Figure 3.** Percentage of complexes for which the lowest energy decoy is within 2 Å from the crystal structure. The scoring functions are represented on the *x*-axis (CHRM, CHARMm-RDIE ($\epsilon(r) = 4r$); DNrg, DOCK-Energy; DChm, DOCK-chemical; DCnt, DOCK-contact; DrugS, DrugScore; ChemS, ChemScore; AutoD, AutoDock), and the various data sets are represented on the *y*-axis (All, whole set (189 complexes); All/NoTr., All without the complexes used to calibrate ChemScore and AutoDock; AspPr., aspartic protease; Oxido., oxidoreductase; SerPr., serine protease; MetPr., metalloprotease; Immu., immunoglobulin; Arab., L-arabinose binding protein; Mhc, major histocompatibility protein).

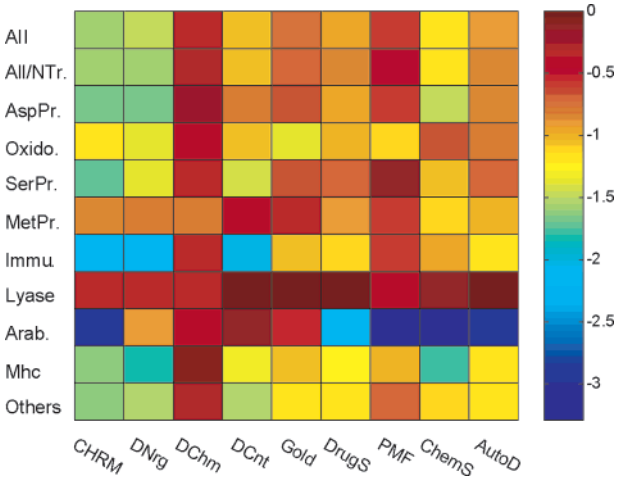


**Figure 4.** Discriminative power. A discriminative power value of zero means no discriminative power, and the lower the value, the more discriminative is the scoring function. See Figure 3 for the definition of the scoring functions and the data sets.

Figure 4 displays the results for the discriminative power. For the data set 1a, the force-field-based potentials CHARMm-4r and DOCK-energy perform best, with discriminative power values of −1.58 and −1.47, respectively. They are followed by DOCK-contact, Drug-

Score, and ChemScore, with DP values of −1.20, −1.06, and −0.93, respectively.

A similar analysis was performed using the docked and misdocked structures together. For the data set 1a, the percentage of complexes for which the lowest energy decoy has an $rmsd_N$ of less than 2 Å was within 2% of that obtained using only the docked structures for all the scoring functions except PMF and DOCK-chemical (data not shown). For these potentials, the deterioration in the performance is 7% and 11%, respectively. This result shows that the analyzed scoring functions discriminate well the docked structures from the decoys that are far from the binding site. This result is of practical importance in docking when the location of the binding site is not known. One should note, however, that most of the binding sites that we considered are deep. It is obviously more difficult to locate shallow binding pockets, which may lead to a deterioration of our results. A successful docking simulation using AutoDock without prior knowledge of the binding site has been reported.[73] Recently, a new method for determining binding sites on proteins has been successfully validated on seven protein−ligand complexes.[74]

We have found that with respect to our data set the steric complementarity between the receptor and its natural ligand is crucial for recognizing near-native structures among a set of decoys. Our finding about the role of steric interaction is consistent with a study that evaluated the two docking functions, DOCK-energy and PMF.[37] This comparison was based on a set of 34 protein−ligand complexes. It was found that omitting the electrostatic term in CHARMm has a small impact on the ability of the molecular mechanics potential to recognize near-native poses. The success rate, defined as the percentage of complexes for which the lowest energy decoy has an $rmsd_N$ smaller than 1.5 Å, was 79% for the DOCK-energy potential and 59% for the PMF potential. These results agree well with ours. Although the significance of steric interactions may imply that the electrostatic contribution is negligible, it should be noted that this effect may be overemphasized because of the preparation of the data set. Namely, minimizing receptor and ligand in the crystal conformation may lead to a perfect "induced fit" (which may carry over for only slightly deviating decoys), whereas for all decoys, only the ligand geometries were minimized while keeping the receptor rigid. Thus, taking into account the rather steep potential curves in the case of van der Waals potentials compared to the electrostatic ones, a less-than-ideal "induced fit" of more strongly deviating decoys would be more easily recognized by the van der Waals potential. In fact, the role of electrostatics has been demonstrated in that optimization of electrostatic interactions can be used to increase affinity and specificity.[75]

Very recently, an evaluation of 11 scoring functions was reported.[32] It was based on a data set of 100 protein−ligand complexes, 44 of which are deposited in the LPDB, and the decoys were generated by the docking program AutoDock. The recognition rates were on average between 10% and 20% inferior to ours except for ChemScore and D-Score, where D-Score stands for the CScore implementation of the DOCK-energy potential. In these two cases, we found much higher recognition rates. For the latter scoring function, the discrepancy comes partly from the fact that we used the academic version of DOCK instead of D-Score. We also assessed D-Score, and the results were significantly worse with respect to DOCK-energy (data not shown). For ChemScore, the origins of the discrepancy are not clear. We note that in the study of Wang et al.[32] a high recognition rate was obtained for the FlexX scoring function, which is also a regression-based scoring function with contributions similar to those of ChemScore. It cannot be excluded that the difference comes from the fact that the decoys were not generated in the same way in the two studies. DrugScore was validated on two sets of protein−ligand structures comprising 91 and 100 complexes, respectively.[20] The decoys were generated by the docking tools FlexX and DOCK, respectively. The percentages of complexes found by DrugScore for which the top-ranking pose has an $rmsd_N$ below 2 Å were 73% and 70%, respectively. In addition, it was shown that DrugScore and DOCK-energy yield similar results and perform better than DOCK-chemical. These findings are consistent with our study. Furthermore, while it was shown that DrugScore recognizes well-docked structures slightly better than AutoDock,[76] we found that AutoDock yields slightly higher recognition rates than DrugScore.

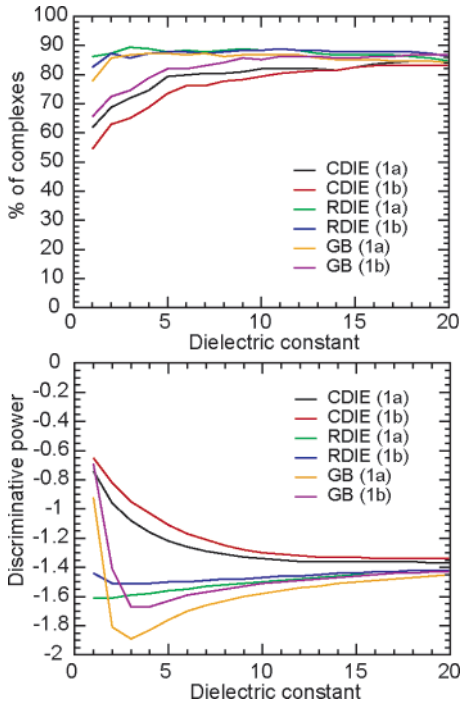


**Figure 5.** Percentage of complexes for which the lowest energy decoy is within 2 Å from the crystal structure (top) and discriminative power (bottom) for the CDIE, RDIE, and GB models as a function of the solute dielectric constant. Sets 1a and 1b comprise 189 complexes with the modified and standard protonation states, respectively.

**3.2.1. Effect of Solvation Models.** Here, we analyze the role of a variety of solvation models in conjunction with the CHARMm force field with respect to the recognition of near-native poses. In addition to the RDIE model (see previous section), we also consider the CDIE and GB models (see section 2 for a description). Figure 5 (top) depicts the percentage of complexes for which the lowest energy decoy lies within 2 Å from the crystal structure as a function of the solute dielectric constant.

These curves are almost flat for a dielectric constant larger than 20 and are, therefore, not shown. With respect to a van der Waals potential, the RDIE and GB models give slightly better results (around 10%), although in the latter case this is only achieved by using the modified protonation states. For $\epsilon = 4$, the success rate for the RDIE and GB models for set 1a is 89% and 87%, respectively.

Figure 5 (bottom) shows the behavior of the discriminative power with respect to the dielectric constant. The CHARMm van der Waals potential yields a DP value of −1.34 and −1.33 for sets 1a and 1b, respectively. The best results are obtained by the GB model with the modified protonation states. For $\epsilon = 4$, GB gives a DP value of −1.83 compared to −1.58 and −1.16 for RDIE and CDIE, respectively. The GB model yields the best discriminative power among all the scoring functions analyzed in this study. At low $\epsilon$ ($\epsilon < 5$), RDIE discriminates much better correctly docked structures from misdocked ones than CDIE. These results suggest the use of the RDIE model (with a dielectric constant of 3−4, where the discriminative power is highest) as a computationally efficient alternative to the GB model.

**3.2.2. Effect of the Protonation States.** Binding affinities determined by an empirical scoring function are relatively insensitive to the atomic details of the receptor and the ligand. To a lesser extent, this fact also holds for knowledge-based potentials. Therefore, adding or removing a proton is not expected to change significantly the binding energies calculated by such methods. For instance, ChemScore does not distinguish hydrogen bonds from ionic interactions. In general, empirical and knowledge-based scoring functions do not require a protonation model. This fact is also true for the GOLD scoring function, which does not make use of partial charges. Indeed, we consistently found very similar results for these scoring functions with either the standard or the modified protonation states. Larger effects may be expected for force-field-based potentials. For the CHARMm force field, this effect is investigated in Figure 5. For low values of the dielectric constant, adjusting the protonation states slightly improves the ability of the CHARMm potential to recognize well-docked structures. If we consider only the 84 complexes having a modified protonation state, the increase in the percentage of complexes for which the lowest energy decoy has an $rmsd_N$ smaller than 2 Å is 17%, 11%, and 29% for the CDIE, RDIE, and GB models with $\epsilon = 1$, respectively. For higher values of $\epsilon$, the magnitude of the electrostatic interactions is reduced, and therefore, modifying the protonation state has a smaller effect. Lower values of discriminative power are also obtained by modifying the protonation states. The results are most similar for the RDIE model, which yields a high success rate already using the standard protonation states for any value of the dielectric constant. The DOCK-energy and chemical scoring functions, which make use of a distance-dependent dielectric function, yield results similar to those from the CHARMm−RDIE model; i.e., a slight increase of performance is observed when adjusting the protonation states. These results suggest that different modeling of the protonation state of the ligand and the receptor does not influence the recognition of well-docked configurations. On the other hand, a virtual screening experiment against dihydrofolate reductase has shown that the protonation state of methotrexate and pteridine affects significantly their binding energies computed by the DOCK force field.[13] Furthermore, these two ligands obtained a high rank only in their protonated form, which is in agreement with experiment.

**3.2.3. Effect of the Zinc Charge Model.** Ligand−metal interactions are notoriously difficult to model, and using different interaction models may influence significantly binding energies computed by force-field potentials. This effect is investigated here for zinc using a charge of +2, thus assuming ionic interactions, or determined by ab initio methods, taking into account charge transfer between the metal and liganded protein residues (see section 2). For the CHARMm force field and $\epsilon = 4$, the two charge models yield very similar percentages of recognition (around 70%). Considering this result, one needs to keep in mind that we only considered decoys for which the distance between the zinc and its coordinated ligand atoms is close to the value in the crystal structure (see section 2).

**3.2.4. Analysis of Cross-Decoys.** We selected CHARMm-4r, DrugScore, and ChemScore to analyze the set of cross-decoys because they have yielded the highest decoy recognition rates on the "native" complexes and represent the three classes of scoring function. Figure 6 shows the $rmsd_N$ of the lowest energy configuration for the all-pairs decoys on trypsin (top panels) and HIV-1 protease (bottom panels). On average, there is a slight decrease in performance with respect to the case of the decoys (about 12%), except for CHARMm-4r in the case of HIV-1 protease, in which the recognition rates decrease from 86% for the decoys to 52% for the cross-decoys. This result may be attributed to the increased steepness of the CHARMm force field compared to the other two functions. However, if we consider an $rmsd_N$ of 2.5 Å as a threshold below which the recognition is considered to be successful, the recognition rate decreases from 86% for the decoys to 74% for the cross-decoys for CHARMm-4r in the case of HIV-1 protease. This result shows that the larger decrease of performance for CHARMm-4r with respect to DrugScore and ChemScore is somewhat misleading. Furthermore, the rigidity of the trypsin receptors, as well as the smaller size of the ligands, makes the recognition of well-docked configurations in the case of the trypsin complexes much easier than in the case of the HIV-1 complexes. A decomposition of the CHARMm potential energy values yields a recognition rate of 67% and 71% for the trypsin and HIV-1 decoys, respectively, if only the van der Waals potential is used. These percentages are 67% and 55% for the trypsin and HIV-1 cross-decoys, respectively. This result shows that the crucial role of steric complementarity between the ligand and the receptor that was highligthed previously in decoy recognition is still valid in cross-decoy recognition. The average rmsd between the trypsin receptors is 0.51 Å (with contributions of 0.15 and 0.96 Å for the backbone and side chain atoms, respectively), whereas this value is 0.87 Å for the HIV-1 receptors (0.58 and 1.12 Å for the backbone and side chain atoms, respectively). DrugScore yields success rates for the cross-decoys that are even slightly higher than for the decoys.

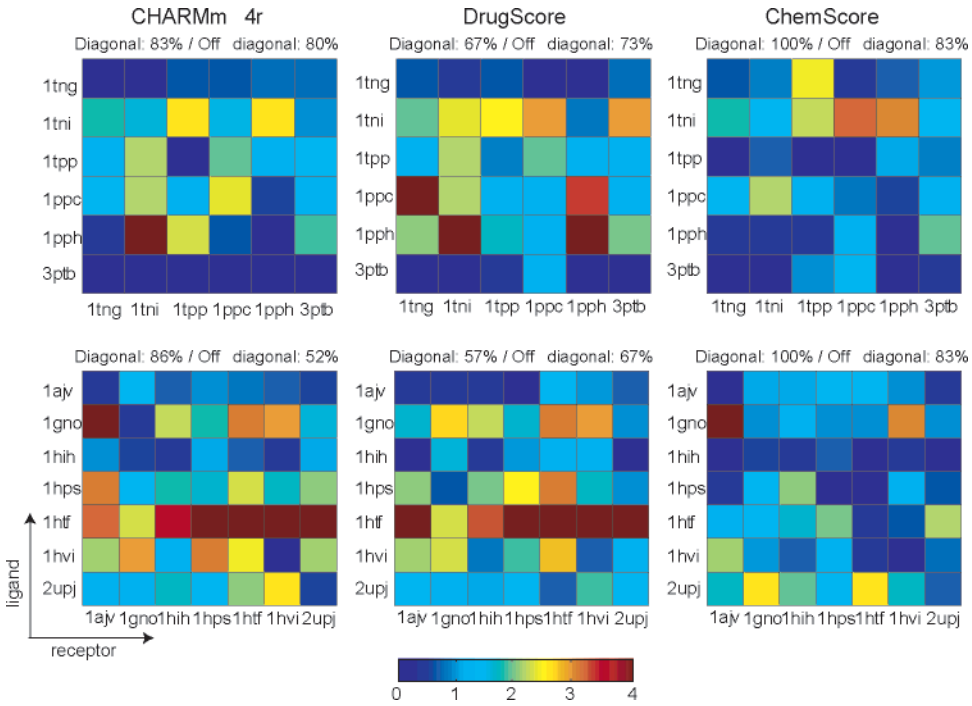


**Figure 6.** The rmsd (in Å) of the lowest energy configuration from the native structure for the all-pairs decoys on trypsin (top panels) and HIV-1 protease (bottom panels). The results for CHARMm-4r, DrugScore, and ChemScore are shown in the left, middle, and right panels, respectively. The *x*-axis and *y*-axis represent the different receptors and ligands, respectively. The average recognition rates for the decoys (diagonal elements) and cross-decoys (off-diagonal elements), i.e., the percentages of complexes for which the lowest energy decoy is within 2 Å from the crystal structure, are indicated above the plots.

This may highlight an advantage of this knowledge-based scoring function because it is more robust to small changes in the receptor conformation than the force-field-based or empirical scoring functions. On a per complex basis, large variations in performance can be observed. For instance, for DrugScore, the ligand 1ppc is well recognized against the 1ppc, 1tpp, and 3ptb receptors and poorly recognized against 1tng and 1pph. The rmsd between 1ppc and the other trypsin receptors ranges between 0.51 and 0.74 Å. This result indicates that even small changes in the receptor conformation can have a large impact on the results of a docking simulation, and therefore, it may be important to incorporate receptor flexibility into docking algorithms.

Cross-docking simulations can help to identify the "ideal" receptor, i.e., the one that yields low $rmsd_N$ values for the majority of the ligands docked into this receptor. This receptor would be 3ptb and 2upj in the case of trypsin and HIV-1 complexes, respectively (Figure 6). This information might be useful in a virtual screening experiment, when there is no other obvious way to choose a receptor structure for a given target. Such an approach is not applicable, however, when the receptors display considerable flexibility because in this case the binding site cannot be considered the same across the various receptors. In a separate evaluation of docking and cross-docking simulations for three enzymes (thrombin, thermolysin, and neuraminidase) that made use of ChemScore, it was reported that the recognition rate decreases from 76% to 49% on going from native to non-native docking.[36] On the basis of a set of 34 protein−ligand complexes, Perez and Ortiz compared the performance of DOCK and PMF in docking and cross-docking.[37] In docking, the success rate was 79% and 59% for DOCK and PMF, respectively, while in cross-docking it decreases to 56% and 41%, respectively. We found similar deterioration in performance for CHARMm-4r and ChemScore but not for DrugScore, which yields recognition rates for the cross-decoys that are similar to those obtained for the decoys.

**3.2.5. Influence of the Decoy and Cross-Decoy Preparation.** Decoys and cross-decoys were minimized to an energy gradient of 0.05 kcal $mol^{-1}$ $Å^{-1}$ in the rigid binding site of the receptor. Since the poses generated in a docking simulation are usually not minimized to the same extent, we investigated the outcome of using minimized and nonminimized decoys and cross-decoys on the recognition of well-docked structures. This comparison was performed for the same sets of trypsin and HIV-1 complexes shown in Figure 6, and the recognition rates for the decoys and cross-decoys are listed in Table 3. The main difference with respect to the minimized decoys and cross-decoys lies in the large decrease in performance observed for ChemScore. On the other hand, CHARMm-4r and DrugScore perform nearly equally well for the minimized and nonminimized (cross-)decoys. It is also instructive to analyze the behavior of the discriminative power in going from minimized to nonminimized (cross-)decoys (Table 4). A significant drop in discriminative power is observed for CHARMm-4r with values close to zero for the nonminimized (cross-)decoys. This finding suggests that it is important to generate minimized poses (albeit with more computational expense) to dock successfully a ligand with a force-field-based scoring function. This result is also true for ChemScore but to a lesser extent. As noticed previously for the recognition rates, DrugScore gives discriminative power values that differ the

**Table 3.** Recognition of Well-Docked Structures among a Set of Decoys and Cross-Decoys[a]

| | trypsin | | | HIV-1 protease | | |
|---|---|---|---|---|---|---|
| | CHARMm-4r | DrugScore | ChemScore | CHARMm-4r | DrugScore | ChemScore |
| decoys, min | 83 | 67 | 100 | 86 | 57 | 100 |
| decoys, non-min | 100 | 50 | 50 | 86 | 71 | 86 |
| cross-decoys, min | 80 | 73 | 83 | 52 | 67 | 80 |
| cross-decoys, non-min | 67 | 57 | 53 | 55 | 55 | 43 |

[a] The percentage of complexes for which the lowest energy decoy is within 2 Å from the crystal structure is reported for minimized and nonminimized sets of decoys and cross-decoys.

**Table 4.** Recognition of Well-Docked Structures among a Set of Decoys and Cross-Decoys[a]

| | trypsin | | | HIV-1 protease | | |
|---|---|---|---|---|---|---|
| | CHARMm-4r | DrugScore | ChemScore | CHARMm-4r | DrugScore | ChemScore |
| decoys, min | −0.92 | −0.64 | −0.87 | −1.06 | −0.84 | −1.36 |
| decoys, non-min | −0.18 | −0.77 | −0.56 | −0.11 | −0.36 | −0.48 |
| cross-decoys, min | −1.28 | −0.79 | −1.24 | −0.80 | −0.74 | −1.13 |
| cross-decoys, non-min | −0.13 | −0.80 | −0.23 | −0.09 | −0.62 | −0.33 |

[a] The discriminative power values are reported for minimized and nonminimized sets of decoys and cross-decoys.

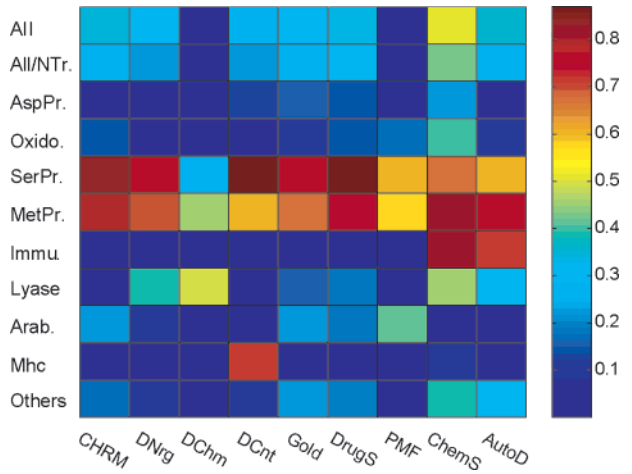


**Figure 7.** Square of the correlation coefficients ($R^2$) between experimental and calculated binding energies. Values were set to zero in the case of an anticorrelation. See Figure 3 for the definition of scoring functions and data sets.

least between the two sets, indicating that the (free) energy surface of this scoring function is the least rugged.

**3.3. Prediction of Binding Affinities.** In this section, we analyze the ability of the scoring functions to rank binding energies. Figure 7 shows the square of the correlation coefficients ($R^2$ value) for the different data sets using the minimized crystal structure. ChemScore achieves the highest correlation with an $R^2$ value of 0.51 for set 1a. ChemScore also yields the highest $R^2$ value ($R^2 = 0.43$) for set 2, where the complexes used to calibrate this scoring function and AutoDock were removed. In particular, ChemScore outperforms all other scoring functions except AutoDock for the immunoglobulin set, although no immunoglobin was included in the training set. It also gives higher $R^2$ values than most of the other scoring functions for the data sets for oxidoreductase, lyase, and others. The training set of ChemScore contains 6 oxidoreductase, 1 lyase, and 11 other complexes that belong also to the LPDB. Furthermore, ChemScore ranks binding affinities significantly better than the other regression-based scoring function AutoDock. Although the results always depend on the set of complexes, this suggests that ChemScore was more broadly parametrized; for instance, a larger diversity of interactions was present in the training set. In general, the other scoring functions perform well for the serine protease and metalloprotease families and unsatisfactorily for the remaining data sets. The aspartic protease and oxidoreductase sets yield a low correlation despite the wide range of binding affinities of 7 and 8 orders of magnitude, respectively. The CHARMm-4r and DOCK-energy obtain similar correlation coefficients for most of the sets. The $R^2$ value between the binding affinities computed by these two potentials is 0.79. DOCK-chemical is the poorest scoring function to rank binding affinities, whereas DOCK-contact, despite its simplicity, yields results that are only slightly inferior to CHARMm-4r or DrugScore. The correlation between the binding affinities computed by all the scoring functions, including also the logarithm of the ligand molecular weight as a scoring function (see also below), is generally high except when the $R^2$ values are computed by either DOCK-chemical or CHARMm-CDIE and all the other scoring functions. For set 1a, the average $R^2$ value between the all-pairs scoring functions without DOCK-chemical and CHARMm-CDIE and between these two scoring functions and all the other ones is 0.66 and 0.24, respectively.

Correlation coefficients similar to our results were obtained for DrugScore, Gold, and PMF in a previous study.[32] In contrast, we found better results for DOCK-energy, ChemScore, and AutoDock than in ref 32. As for the recognition of near-native poses, the origin of the discrepancy for DOCK comes partly from the fact that we used the academic version instead of the CScore module. Different correlation values for ChemScore and AutoDock may indicate a set dependence and, therefore, a clear weakness of the regression-based scoring functions. The ability of DrugScore and PMF to rank binding affinities was previously investigated for various sets of protein−ligand complexes.[19,77] These two knowledge-based functions achieved comparable results, for instance, high correlation coefficients for the serine proteases and metalloproteases. In our study, DrugScore and PMF yield correlation values of 0.33 and 0.08 for set 1a, respectively. It cannot be ruled out that the poor performance of PMF originates partly from the fact that we did not use the original implementation.

The reason that all the scoring functions except DOCK-chemical perform well for the serine protease and metalloprotease families might relate to the fact that for these two sets the binding affinities correlate well with the ligand size. The $R^2$ value between the experimental binding affinities and the logarithm of the ligand molecular weights is 0.81 and 0.58, respectively. The logarithmic function was chosen because a survey of experimental data has shown that the binding free energy depends on the number of heavy atoms roughly in a logarithmic way; i.e., it increases initially but then levels off after 15 heavy atoms.[78] ChemScore and AutoDock yield high $R^2$ values for the immunoglobulin data set, whereas all the other scoring functions give either no correlation (DrugScore) or even an anticorrelation. On the other hand, we note that the binding energies calculated by the PB and GB models correlate well with the experimental binding affinities of the immunoglobulins for $\epsilon \leq 2$. In this case, the correlation coefficient depends strongly on the dielectric constant, since it drops to zero for $\epsilon \geq 3$ (see also below). For this set, the experimental binding affinities anticorrelate with the size of the ligand, and therefore, the success of ChemScore and AutoDock can be attributed to their entropic term, which takes into account the decrease in rotational degrees of freedom upon binding. When this contribution is removed, the square of the correlation coefficient for AutoDock decreases from 0.72 to 0.02, and the $R^2$ value between the entropic contribution alone and the experimental values is 0.60. Since the CScore module does not provide the different contributions of the scoring functions, this effect is unknown for ChemScore, but a similar result is expected. It can be anticipated that adding to the other scoring functions an entropic term like the one implemented in ChemScore or AutoDock should improve the results in this case. However, it should be kept in mind that these entropic contributions are based on simplified models. In a virtual screening experiment, it is likely that higher scores will be attributed to larger ligands if the scoring function contains only terms for favorable interactions. This problem can be overcome by considering the restriction of the degrees of freedom upon binding of (ideally) both binding partners. As a cheap alternative, normalizing the binding energies on the basis of the total number of heavy atoms has been proposed.[79]

Other reasons than the lack of an entropic term must be invoked to understand the failure of the scoring functions to rank binding affinities for the other classes. High correlation coefficients might not be expected for the L-arabinose binding proteins and the major histocompatibility proteins because of the small range of binding affinity values in connection with the uncertainties in the experimental values. This argument does not hold for the aspartic protease and oxidoreductase sets, however. In the former case, the enzyme undergoes a significant conformational change upon binding. Moreover, crystallographic studies often show the presence of a conserved water molecule bridging the inhibitor and the two flaps. In all scoring functions used, ligand−water−receptor interactions are not considered, however. In the oxidoreductase case, most of the ligands are small and interact with the heme, whereas others are bound to a zinc. As a result, the calculation of binding affinities is particularly challenging.

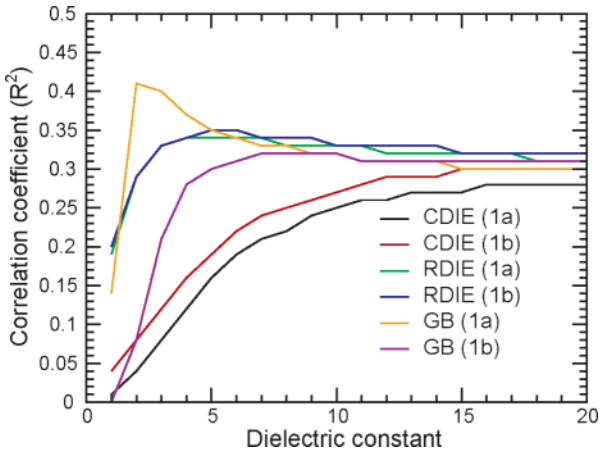


**Figure 8.** Square of the correlation coefficients ($R^2$) between experimental and calculated binding energies for the CDIE, RDIE, and GB models as a function of the solute dielectric constant. Sets 1a and 1b comprise 189 complexes with the modified and standard protonation states, respectively.

**3.3.1. Analysis of Solvation Models.** We next evaluate the performance of various solvation models used in conjunction with the CHARMm force field for ranking ligands with respect to binding energies. FDP calculations were carried out with a grid spacing of 0.4 Å and a solute dielectric constant of either 1 ($FDP_{\epsilon 1}$) or 4 ($FDP_{\epsilon 4}$). The $FDP_{\epsilon 1}$ model yields no correlation with experiment (data not shown). Better results are obtained by $FDP_{\epsilon 4}$ for data set 1a. Yet the correlation is moderate ($R^2 = 0.35$), only slightly larger than when using only a van der Waals potential ($R^2 = 0.29$) and comparable to the value obtained by CHARMm-4r ($R^2 = 0.34$).

We then analyzed the performance of simpler solvent models. These include the CDIE, RDIE, and GB models. Figure 8 shows the correlation between results obtained with these models and experimental binding affinities as a function of the solute dielectric constant for data sets 1a and 1b. The curves are almost flat for a dielectric constant larger than 20 and are therefore not displayed. It can be seen that none of these models perform significantly better than a van der Waals potential ($R^2 = 0.29$). The GB model yields the best result with $R^2 = 0.37$ for $\epsilon = 4$ using the modified protonation states. At low $\epsilon$ ($\epsilon < 5$), RDIE gives much better results than CDIE and almost as good results as GB with $\epsilon = 4$, albeit with considerably less computational cost. The choice of the optimal value of the dielectric constant depends on the receptor. Our results show that for the GB model the highest correlation coefficients for the data sets 3−11 are obtained using $\epsilon$ values that range between 1 and 6. Furthermore, solvation energies computed by an electrostatic continuum model depend strongly on the set of partial atomic charges and radii. Very recently, it has been shown that slightly scaling the atomic radii has a profound impact on binding energies computed by the Poisson equation for benzamidine bound to trypsin.[72] This parameter dependence clearly points to a limit of using such models for binding free energy calculations. Finally, a comparison of the GB and PB models with explicit solvent results has been reported for the calculation of the binding free energy of an

octapeptide ligand to the murine MHC class I protein.[80] Good agreement was found for the neutralized ligand, but large discrepancies were obtained for the ionized ligand.

**3.3.2. Effect of the Protonation States.** For all the scoring functions (without considering CHARMm; see below), we found very similar correlation values between the standard and modified protonation states. On the other hand, the results depend strongly on the solvation model used with the CHARMm force field (Figure 8). A higher correlation between experiment and FDP or GB is achieved when using the complexes with the modified protonation states. The improvement is significant for GB at low $\epsilon$. For $\epsilon = 4$, the $R^2$ value goes from 0.36 to 0.55 when considering the 84 complexes with a modified protonation state. Similar observations can be made for the $FDP_{\epsilon 4}$ results. In contrast, these correlation values are very similar for RDIE. For the CDIE model, the results are better using the standard protonation states. This finding suggests that using more realistic protonation states in conjunction with a reasonably accurate treatment of solvation effects may lead to better agreement with experimental binding affinities. Nevertheless, neither the GB model nor the Poisson equation ranks significantly better binding affinities than a distance-dependent dielectric function. The result for the RDIE model does not imply that modifying the protonation state has no impact on the RDIE binding scores. For $\epsilon = 4$, the average binding energies for the 84 complexes with a modified protonation state differ by 34, 8, and 20 kcal/mol between the standard and modified protonation states for the CDIE, RDIE, and GB model, respectively. This result shows that adding a proton has a large impact on binding energies calculated by a force-field-based scoring function.

**3.3.3. Effect of the Zinc Charge Model.** The correlation values for the metalloproteins are very low (smaller than 0.1) for all the scoring functions using a +2 charge for zinc (data not shown). It is therefore of interest to investigate the effect of using charges obtained by ab initio calculations. The correlation achieved by the $FDP_{\epsilon 1}$ ($R^2 = 0.36$) and $FDP_{\epsilon 4}$ ($R^2 = 0.23$) calculations in this case originates only from three complexes, 1avn, 6cpa, and 7cpa, which have the highest and the two lowest binding affinities among this set. When these three complexes are removed (which reduces the range of binding affinities from 11 to 6 orders of magnitude), the correlation drops to below 0.1. Similar conclusions are found for the other scoring potentials. Thus, it might be that our ab initio zinc charges are not accurate enough, since we did not take into account the charge transfer between zinc and the ligand atoms. Furthermore, the zinc charge is probably not the same in the bound and unbound states. However, if the interactions between zinc and surrounding atoms are similar for a series of ligand, a large cancellation of errors can be expected.[81]

**3.3.4. Hydrophobic Effect.** To model the effect of hydrophobic desolvation, we added to the CHARMm force field a contribution proportional to the solvent-accessible surface area (SASA). This term was investigated in connection with the CDIE, RDIE, GB, and PB models. The surface tension constant ($\gamma$) was varied between 5 and 30 cal $mol^{-1}$ $Å^{-2}$. Higher values of $\gamma$ have little physical meaning. For set 1a, the binding energies computed by the SASA term alone and the van der Waals potential correlate with an $R^2$ value of 0.94. This indicates that the SASA term is not expected to yield much improvement. Furthermore, the $R^2$ value between the experimental affinities and the energies calculated by the SASA model is 0.29. On a receptor basis, this correlation is the highest for the serine protease and metalloprotease families with $R^2$ values of 0.77 and 0.56, respectively. It is very low for the remaining classes with even a strong anticorrelation for the immunoglobulin family ($R = -0.78$). Again, these findings reflect the correlation (anticorrelation) of experimental binding affinities with respect to the size of the ligand for these data sets. In total, adding a SASA contribution to the CHARMm leads only to small improvements. Recently, deficiencies of a hydrophobic model based on the solvent-accessible surface area have been discussed. Hydration free energies of the cycloalkanes fall below the linear correlation for the *n*-alkane analogues.[82,83] Simulations of small alkanes have suggested that the SASA model can only describe the thermodynamics of cavity formation but does not correctly model the favorable van der Waals interactions between interior atoms of the solute and the solvent.[84−86] These contributions may play an important role in cases where the number of solvent-exposed and buried atoms changes considerably, such as in binding reactions.

## 4. Conclusion

We presented an assessment of nine scoring functions for protein−ligand interactions using a database of 189 protein−ligand complexes. Most of the potentials that we analyzed recognize well near-native configurations among a set of decoys. CHARMm, DOCK-energy, DrugScore, ChemScore, and AutoDock showed the best performance in discriminating near-native from misdocked structures. For these scoring functions, the recognition rate was between 80% and 90%, which is fairly remarkable and shows their usefulness in the docking problem. The analysis of cross-decoys versus decoys as well as minimized versus nonminimized poses has shown that the knowledge-based potential DrugScore is less sensitive to the atomic details of the receptor than the regression-based scoring function ChemScore or the CHARMm force field. The cross-decoy results also lead to the recommendation of evaluating receptor structures with respect to their "dockability" prior to predictive docking calculations.

We have found here that steric complementarity between the ligand and the receptor is more important than electrostatics to identify a near-native pose. As a result, the treatment of solvation effects should have a minor impact on the ability of a force-field potential to recognize near-native configurations. In this regard, we have shown that a computationally cheaper distance-dependent dielectric function works almost as well as a generalized Born model. However, it cannot be excluded that the steric effect may be overemphasized because of the preparation of the data set. In the case of absolute binding free energy prediction, errors do not cancel out and different solvation contributions are expected to play a larger role.

On the other hand, our work has also confirmed that the prediction of binding affinities still represents a major challenge. For the 189 complexes in the LPDB, only ChemScore achieves a fair correlation between the binding scores and experimentally determined binding energies. Most of the scoring functions perform well only for the serine protease and metalloprotease families and unsatisfactorily for the remaining data sets. For these two sets, the experimental binding affinities correlate well with the size of the ligands, which may explain their success. Including terms that account for changes in the degrees of freedom of the binding partners upon binding is expected to yield an improvement in these cases.

We have investigated in detail the effect of adjusting the protonation state of binding site titratable groups. Improvement in the case of binding affinity prediction was achieved for the generalized Born model and the Poisson equation used in conjunction with the CHARMm potential. Despite this, these two models do not rank significantly better binding affinities than a distance-dependent dielectric function. It might be necessary to reparametrize these models to obtain better results in binding free energy calculations. Since more accurate charges for zinc do not lead to a better agreement with experiment, the prediction of affinities for metalloproteins remains problematic. Finally, we note that all of the decoys and updated data have been integrated into the LPDB and are available at http://lpdb.scripps.edu.

**Acknowledgment.** We thank Dr.Olivier Roche for initial construction of the LPDB. The Swiss National Science Foundation is gratefully acknowledged for financial support to P.F., and H.G. gratefully acknowledges a Feodor-Lynen fellowship from the Alexander-von-Humboldt Foundation, Germany. Financial support from the NIH (Grants GM37554 and RR12255) is also appreciated.

**Supporting Information Available:** Three tables listing the percentage of complexes for which the lowest energy decoy is within 2 Å from the crystal structure, the discriminative power values of the scoring functions, and the correlation of experimental and calculated binding affinities using the minimized crystal structure. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Bleicher, K. H.; Böhm, H. J.; Müller, K.; Alanine, A. I. Hit and lead generation beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.
(2) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM−A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.
(3) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(4) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(5) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.
(6) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(7) Wu, G.; Robertson, D. H.; Brooks, C. L., III; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKERA CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24*, 1549−1562.
(8) Luo, R.; Gilson, M. K. Synthetic adenine receptors: Direct calculation of binding affinity and entropy. *J. Am. Chem. Soc.* **2000**, *122*, 2934−2937.
(9) Mardis, K. L.; Luo, R.; Gilson, M. K. Interpreting trends in the binding of cyclic ureas to HIV-1 protease. *J. Mol. Biol.* **2001**, *309*, 507−517.
(10) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.
(11) Bühm, H. J.; Stahl, M. The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2002; Vol. 18, pp 41−87.
(12) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caisch, A. Exhaustive docking of molecular fragments on protein binding sites with electrostatic solvation. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 88−105.
(13) Zou, X.; Sun, Y.; Kuntz, I. D. Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model. *J. Am. Chem. Soc.* **1999**, *121*, 8033−8043.
(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) Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859−883.
(17) DeWitte, R.; Shakhnovich, E. 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.
(18) Mitchell, J. B. O.; Laskowski, R. A.; Alexander, A.; Thornton, J. M. BLEEP−Potential of mean force describing protein−ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.
(19) 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.
(20) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(21) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(22) Momany, F. A.; Rone, R. Validation of the general-purpose QUANTA. 3.2/CHARMm force-field. *J. Comput. Chem.* **1992**, *13*, 888−900.
(23) 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.
(24) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(25) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 195−202.
(26) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of exible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.
(27) Roche, O.; Kiyama, R.; Brooks, C. L., III. Ligand−protein database: Linking protein−ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592−3598.
(28) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(29) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(30) Vieth, M.; Hirst, J.; Kolinski, A.; Brooks, C. L., III. Assessing energy functions for flexible docking. *J. Comput. Chem.* **1998**, *19*, 1612−1622.
(31) Vieth, M.; Hirst, J.; Dominy, B. N.; Daigler, H.; Brooks, C. L., III. Assessing search strategies for flexible docking. *J. Comput. Chem.* **1998**, *19*, 1623−1631.

(32) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(33) Stubbs, M. T.; Reyda, S.; Dullweber, F.; Möller, M.; Klebe, G.; Dorsch, D.; Mederski, W.; Wurziger, H. pH-dependent binding modes observed in trypsin crystals: Lessons for structure-based drug design. *ChemBioChem* **2002**, *3*, 246−249.
(34) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 4−16.
(35) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **2002**, *322*, 339−355.
(36) Murray, C. W.; Baxter, C. A.; Frenkel, D. The sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 547−562.
(37) Perez, C.; Ortiz, A. R. Evaluation of docking functions for protein−ligand docking. *J. Med. Chem.* **2001**, *44*, 3768−3785.
(38) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new set for validating predictions of protein−ligand interaction. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457−471.
(39) Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem., Int. Ed.* **2003**, *42*, 2718−2736.
(40) Fitzgerald, M. M.; Musah, R. A.; McRee, D. E.; Goodin, D. B. A ligand-gated, hinged loop rearrangement opens a channel to a buried artificial protein cavity. *Nat. Struct. Biol.* **1996**, *3*, 626−631.
(41) Ferguson, D. M.; Radmer, R. J.; Kollman, P. A. Determination of the relative binding free energies inhibitors to HIV-1 protease. *J. Med. Chem.* **1991**, *34*, 2654−2659.
(42) Harte, W. E., Jr.; Beveridge, D. L. Prediction of the protonation state of the active site aspartyl residues in HIV-1 protease−inhibitor complexes via molecular dynamics simulation. *J. Am. Chem. Soc.* **1993**, *115*, 3883−3886.
(43) Yamazaki, T.; Nicholson, L. K.; Torchia, D. A.; Wingfield, P.; Stahl, S. J.; Kaufman, J. D.; Eyermann, C. J.; Hodge, C. N.; Lam, P. Y. S.; Ru, Y.; Jadhav, P. K.; Chang, C.-H.; Weber, P. C. NMR and X-ray evidence that the HIV protease catalytic aspartyl groups are protonated in the complex formed by the protease and a non-peptide cyclic urea-based inhibitor. *J. Am. Chem. Soc.* **1994**, *116*, 10791−10792.
(44) Chen, X.; Tropsha, A. Relative binding free energies of peptide inhibitors of HIV-1 protease: The influence of the active site protonation state. *J. Med. Chem.* **1995**, *38*, 42−48.
(45) Gomez, J.; Freire, E. Thermodynamic mapping of the inhibitor site of the aspartic protease endothiapepsin. *J. Mol. Biol.* **1995**, *252*, 337−350.
(46) Smith, R.; Brereton, I. A.; Chai, R. Y.; Kent, S. B. H. Ionization states of the catalytic residues in HIV-1 protease. *Nat. Struct. Biol.* **1996**, *3*, 946−950.
(47) Tokarski, J. S.; Hopfinger, A. J. Constructing protein models for ligand−receptor binding thermodynamic simulations: An application to a set of peptidometic renin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 779−791.
(48) Kulkarni, S. S.; Kulkarni, V. M. Structure based prediction of binding affinity of human immunodeficiency virus-1 protease inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1128−1140.
(49) Piana, S.; Sebastiani, D.; Carloni, P.; Parrinello, M. Ab initio molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site. *J. Am. Chem. Soc.* **2001**, *123*, 8730−8737.
(50) Tawa, G. J.; Topol, I. A.; Burt, S. K.; Erickson, J. W. Calculation of relative binding free energies of peptidic inhibitors to HIV-1 protease and its I84V mutant. *J. Am. Chem. Soc.* **1998**, *120*, 8856−8863.
(51) Baldwin, E. T.; Bhat, N.; Gulnik, S.; Liu, B.; Topol, I. A.; Kiso, Y.; Mimoto, T.; Mitsuya, H.; Erickson, J. W. Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine. *Structure* **1995**, *3*, 581−590.
(52) Hoog, S. S.; Zhao, Z.; Winborne, E.; Fisher, S.; Green, D. W.; DesJarlais, R. L.; Newlander, K. A.; Callahan, J. F.; Moore, M. L.; Huffman, W. L.; Abdel-Meguid, S. S. A check on rational drug design: crystal structure of a complex of human immunodeficiency virus type 1 protease with a novel-turn mimetic inhibitor. *J. Med. Chem.* **1995**, *38*, 3246−3252.
(53) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*; Gaussian, Inc.: Pittsburgh, PA, 1998.
(54) Harding, M. M. The geometry of metal−ligand interactions relevant to proteins. *Acta Crystallogr. D* **1999**, *55*, 1432−1443.
(55) *INSIGHT II*; Molecular Simulations: San Diego, CA, 2002.
(56) Warwicker, J.; Watson, H. C. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **1982**, *157*, 671−679.
(57) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(58) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(59) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.
(60) Scarsi, M.; Apostolakis, J.; Caisch, A. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A* **1997**, *101*, 8098−8106.
(61) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.
(62) Dominy, B. N.; BrooksI, C. L., III. Development of a generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **1999**, *103*, 3765−3773.
(63) Lee, M. S.; Salsbury, F. R.; Brooks, C. L., III. Novel generalized Born models. *J. Chem. Phys.* **2002**, *116*, 10606−10614.
(64) Lee, M. S.; Feig, M.; Salsbury, F. R.; Brooks, C. L., III. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003**, *24*, 1348−1356.
(65) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1985**, *7*, 230−252.
(66) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity. A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.
(67) *SYBYL*;, Tripos Associates: St. Louis, MO, 2002.
(68) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.
(69) Stouten, P. F. W.; Froemmel, C.; Nakamura, H.; Sander, C. An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. Simul.* **1993**, *10*, 97−120.
(70) Tsai, C.-J.; Kumar, S.; Ma, B.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **1999**, *8*, 1181−1190.
(71) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Retjo, P. A.; Freer, S. T.; Rose, P. W. Complexity and simplicity of ligand−macromolecule interactions: the energy landscape perspective. *Curr. Opin. Struct. Biol.* **2002**, *12*, 197−203.
(72) Rankin, K. N.; Sulea, T.; Purisima, E. O. On the transferability of hydration-parametrized continuum electrostatics models to solvated binding calculations. *J. Comput. Chem.* **2003**, *24*, 954−962.
(73) Hetenyi, C.; van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **2002**, *11*, 1729−1737.
(74) Glick, M.; Robinson, D. D.; Grant, G. H.; Richards, W. G. Identification of ligand binding sites on proteins using a multi-scale approach. *J. Am. Chem. Soc.* **2002**, *124*, 2337−2344.
(75) Kangas, E.; Tidor, B. Optimizing electrostatic affinity in ligand−receptor binding: Theory, computation, and ligand properties. *J. Chem. Phys.* **1998**, *109*, 7522−7545.
(76) Sotriffer, C. A.; Gohlke, H.; Klebe, G. Docking into knowledge-based potential fields: A comparative evaluation of DrugScore. *J. Med. Chem.* **2002**, *45*, 1967−1970.
(77) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and "hot spots" for protein−ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.
(78) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−1002.

(79) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−272.
(80) Zhang, L. Y.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. Solvents models for protein−ligand binding: Comparison of implicit solvent Poisson and surface generalized Born models with explicit solvent simulations. *J. Comput. Chem.* **2001**, *22*, 591−607.
(81) Grzybowski, B. A.; Ishchenko, A. V.; Kim, C.-Y.; Topalov, G.; Chapman, R.; Christianson, D. W.; Whitesides, G. M.; Shakhnovich, E. I. Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1270−1273.
(82) Ben-Naim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81*, 2016−2027.
(83) Simonson, T.; Brünger, A. T. Solvation free energies estimated from macroscopic continuum theory: An accuracy assessment. *J. Phys. Chem.* **1994**, *98*, 4683−4694.
(84) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. A "universal" surface area correlation for molecular hydrophobic phenomena. *J. Am. Chem. Soc.* **1999**, *121*, 9243−9244.
(85) Gallicchio, E.; Kubo, M. M.; Levy, R. M. Enthalpy−entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B* **2000**, *104*, 6271−6285.
(86) Pitera, J. W.; van Gunsteren, W. F. The importance of solute−solvent van der Waals interactions with interior atoms of biopolymers. *J. Am. Chem. Soc.* **2001**, *123*, 3163−3164.

JM030489H