Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes

Richard A. Friesner,* Robert B. Murphy,† Matthew P. Repasky,† Leah L. Frye,‡ Jeremy R. Greenwood,† Thomas A. Halgren,† Paul C. Sanschagrin,† and Daniel T. Mainz†

Department of Chemistry, Columbia University, New York, New York 10027, Schrödinger, Limited Liability Company, 120 West 45th Street, New York, New York 10036, Schrödinger, Limited Liability Company, 101 SW Main Street, Portland, Oregon 97204

Received December 16, 2005

A novel scoring function to estimate protein−ligand binding affinities has been developed and implemented as the Glide 4.0 XP scoring function and docking protocol. In addition to unique water desolvation energy terms, protein−ligand structural motifs leading to enhanced binding affinity are included: (1) hydrophobic enclosure where groups of lipophilic ligand atoms are enclosed on opposite faces by lipophilic protein atoms, (2) neutral−neutral single or correlated hydrogen bonds in a hydrophobically enclosed environment, and (3) five categories of charged−charged hydrogen bonds. The XP scoring function and docking protocol have been developed to reproduce experimental binding affinities for a set of 198 complexes (RMSDs of 2.26 and 1.73 kcal/mol over all and well-docked ligands, respectively) and to yield quality enrichments for a set of fifteen screens of pharmaceutical importance. Enrichment results demonstrate the importance of the novel XP molecular recognition and water scoring in separating active and inactive ligands and avoiding false positives.

1. Introduction

In two previous papers[1,2] we have described the Glide high throughput docking program and provided performance benchmarks for docking and scoring capabilities. These results have established Glide as a competitive methodology in both areas.[2−5] However, it is clear from enrichment results (ref 2) that there remains substantial room for improvement in separating "active" from "inactive" compounds. In this paper we outline and present results obtained from significantly enhanced sampling methods and scoring functions, hereafter collectively referred to as "extra-precision" (XP) Glide. The key novel features characterizing XP Glide scoring are (1) the application of large desolvation penalties to both ligand and protein polar and charged groups in appropriate cases and (2) the identification of specific structural motifs that provide exceptionally large contributions to enhanced binding affinity. Accurate assignment of these desolvation penalties and molecular recognition motifs requires an expanded sampling methodology for optimal performance. Thus, XP Glide represents a single, coherent approach in which the sampling algorithms and the scoring function have been optimized simultaneously.

The goal of the XP Glide methodology is to semiquantitatively rank the ability of candidate ligands to bind to *a specified conformation of the protein receptor*. Because of the rigid receptor approximation utilized in Glide and other high throughput docking programs, ligands that exhibit significant steric clashes with the specified receptor conformation cannot be expected to achieve good scores, even if they in reality bind effectively to an alternative conformation of the same receptor. Such ligands may be thought of as unable to "fit" into that specified conformation of the protein. For docking protocols to function effectively within the rigid-receptor approximation, some ability to deviate from the restrictions of the hard wall van der Waals potential of the receptor conformation used in docking must be built into the potential energy function employed to predict the ligand binding mode. In XP and SP Glide, this is accomplished by scaling the van der Waals radii of nonpolar protein and/or ligand atoms; scaling the vdW radii effectively introduces a modest "induced fit" effect. However, it is clear that there are many cases in which a reasonable degree of scaling will not enable the ligand to be docked correctly. For example, a side chain in a rotamer state that is very different from that of the native protein−ligand complex may block the ligand atoms from occupying their preferred location in the binding pocket. There will always be borderline situations, but in practice we have found it possible to classify the great majority of cases in cross-docking experiments as either "fitting" or "not fitting". The former are expected to be properly ranked by XP Glide (within the limitation of noise in the scoring function), while the latter require an induced-fit protocol[6,7] to correctly assess their binding affinity.[4] In the present paper, we focus on complexes where the ligand fits appropriately into the receptor, as judged by two factors: (1) the ability to make key hydrogen bonding and hydrophobic contacts and (2) the ability to achieve a reasonable root-mean-square deviation (RMSD), as compared to the native complex or as obtained by analogy with the native complex of a related ligand. Comparison by analogy is often necessary when dealing with a large dataset of active ligands, only a few of which may have available crystal structures.

Our discussion of XP Glide is divided into four different sections. First, in section 2, we describe the novel terms leading to enhanced binding affinity that have been introduced to account for our observations with regard to protein−ligand binding in a wide range of systems. The origin of these terms lies in the theoretical physical chemistry of protein−ligand interactions; however, developing heuristic mathematical representations that can be used effectively in an empirical scoring function, taking into account imperfections in structures due to the rigid receptor approximation and/or limitations of the docking algorithm, requires extensive analysis of, and fitting

* To whom correspondence should be addressed. Phone: 212-854-7606. Fax: 212-854-7454. E-mail: rich@chem.columbia.edu.
† Schrödinger, L.L.C., NY.
‡ Schrödinger, L.L.C., OR.

10.1021/jm051256o CCC: $33.50 © 2006 American Chemical Society
Published on Web 09/23/2006

Pharmacyclics LLC
Exhibit 2036
BeiGene v. Pharmacyclics
PGR2024-00009
EX2036.0001

to, experimental data. Key aspects of this analysis, along with illustrative examples, are provided in section 2 in an effort to provide physical insight as well as formal justification for the model. In developing XP Glide, we have attempted to identify the principal driving forces and structural motifs for achieving significant binding affinity contributions with specific protein−ligand interactions, above and beyond the generic terms that have appeared repeatedly in prior scoring functions. We have found that a relatively small number of such motifs are dominant over a wide range of test cases; the ability to automatically recognize these motifs, and assign binding affinity contributions, potentially represents an advance in the modeling of protein−ligand interactions based on an empirical scheme.

In section 3, we evaluate the performance of our methodology in self-docking, with regard to both the ability to generate the correct binding mode of the complex and the prediction of binding affinity, using docked XP structures for the complexes. In section 4, the performance of the scoring function in enrichment studies (ability to rank known active compounds ahead of random database ligands) for a substantial number of targets, containing qualitatively different types of active sites, is investigated. Our treatment of the data differs significantly from what has generally prevailed in previous papers in the literature; in evaluating scoring accuracy, we distinguish cases where there are significant errors in structural prediction, as opposed to systems where the structural prediction is reasonably good, but the scoring function fails to assign the appropriate binding affinity. By using only well-docked structures to parametrize and assess scoring functions, a way forward toward a globally accurate method, in which multiple structures are employed in docking and/or induced fit methods are utilized to directly incorporate protein flexibility, is facilitated.

The parameterization of XP Glide is carried out using a large and diverse training set comprising 15 different receptors and between 4 and 106 well-docked ligands per receptor. A separately developed test set incorporating four new receptors, and additional ligands for two receptors already in the training set, is also defined. All of the receptor and ligand data is publicly available (as is our decoy set, which has been posted on the Schrodinger Web site and is freely available for downloading) and we provide extensive references documenting the origin of each ligand. The results reported below have been obtained with the Glide 4.0 release.

The development of data sets suitable for the analysis described above is highly labor intensive; consequently, our current test set is too small to draw robust conclusions, and the results reported herein must be regarded as preliminary. While the test set results are encouraging with regard to demonstration of a respectable degree of transferability, a rigorous assessment of the performance to be expected on a novel receptor will have to be performed in future publications. Nevertheless, qualitative and consistent improvement in the results for both training and test set, at least as compared to the alternative scoring functions available in Glide, is demonstrated. Finally, in the conclusion, we summarize our results and discuss future directions.

2. Glide XP Scoring Function

The major potential contributors to protein−ligand binding affinity can readily be enumerated as follows:

(1) Displacement of Waters by the Ligand from "Hydrophobic Regions" of the Protein Active Site. Displacement of these waters into the bulk by a suitably designed ligand group will lower the overall free energy of the system. Waters in such regions may not be able to make the full complement of hydrogen bonds that would be available in solution. There are also entropic considerations; if a water molecule is restricted in mobility in the protein cavity, release into solvent via ligand-induced displacement will result in an entropy gain. As one ligand releases many water molecules, this term will contribute favorably to the free energy. Replacement of a water molecule by a hydrophobic group of the ligand retains favorable van der Waals interactions, while eliminating issues concerning the availability of hydrogen bonds. Transfer of a hydrophobic moiety on the ligand from solvent exposure to a hydrophobic pocket can also contribute favorably to binding by withdrawing said hydrophobic group from the bulk solution.

(2) Protein−Ligand Hydrogen-Bonding Interactions, as well as Other Strong Electrostatic Interactions Such as Salt Bridges. In making these interactions, the ligand displaces waters in the protein cavity, which can lead to favorable entropic terms of the type discussed above in (1). Contributions to binding affinity (favorable or unfavorable) will also depend on the quality and type of hydrogen bonds formed, net electrostatic interaction energies (possibly including long range effects, although these generally are considered small and typically are neglected in empirical scoring functions), and specialized features of the hydrogen-bonding geometry, such as bidendate salt bridge formation by groups such as carboxylates or guanidium ions. Finally, differences in the interactions of the displaced waters, as compared to the ligand groups replacing them, with the protein environment proximate to the hydrogen bond, can have a major effect on binding affinity, as is discussed in greater detail below.

(3) Desolvation Effects. Polar or charged groups of either the ligand or protein that formerly were exposed to solvent may become desolvated by being placed in contact with groups to which they cannot hydrogen bond effectively. In contrast to the two terms described above, such effects can only reduce binding affinity.

(4) Entropic Effects Due to the Restriction on Binding of the Motion of Flexible Protein or Ligand Groups. The largest contributions are due to restriction of ligand translational/orientational motion and protein and ligand torsions, but modification of vibrational entropies can also contribute. As in the case of desolvation terms, such effects will serve exclusively to reduce binding affinity.

(5) Metal−Ligand Interactions. Specialized terms are needed to describe the interaction of the ligand with metal ions. We shall defer the discussion of metal-specific parameterization to another publication, as this is a complex subject in its own right, requiring considerable effort to treat in a robust fashion.

A large number of empirical scoring functions for predicting protein−ligand binding affinities have been developed.[8−19] While differing somewhat in detail, these scoring functions are broadly similar. A representative example, the ChemScore[8] scoring function, is discussed in our comments below, though similar comments would apply to many of the other scoring functions cited in refs 8−19. We briefly summarize how ChemScore treats the first four potential contributors to the binding affinity presented above:

(1) ChemScore[8] contains a hydrophobic atom−atom pair energy term of the form

$$E_{\text{phobic_pair}} = \sum_{ij} f(r_{ij}) \tag{1}$$

Here, i and j refer to lipophilic atoms, generally carbon, and $f(r_{ij})$ is a linear function of the interatomic distance, r_{ij}. For r_{ij} less than the sum of the atomic vdW radii plus 0.5 Å, f is 1.0.

Between this value and the sum of atomic vdW radii plus 3.0 Å, *f* ramps linearly from 1.0 to zero. Beyond the sum of atomic vdW radii plus 3.0 Å, *f* is assigned a value of zero.

This term heuristically represents the displacement of waters from hydrophobic regions by lipophilic ligand atoms. Numerous close contacts between the lipophilic ligand and protein atoms indicate that poorly solvated waters have been displaced by lipophilic atoms of the ligand that themselves were previously exposed to water. The resulting segregation of lipophilic atoms, and concomitant release of waters from the active site, lowers the free energy via the hydrophobic effect, which is approximately captured by the pair scoring function above. Terms based on contact of the hydrophobic surface area of the protein and ligand, while differing in details, essentially measure the same free energy change and have a similar physical and mathematical basis.

Various parameterizations of the atom−atom pair term have been attempted, including efforts such as PLP,[9] in which every pair of atom types is assigned a different empirical pair potential. However, it is unclear whether this more detailed parameterization yields increased accuracy in predicting binding affinities. A key issue is whether a correct description of the hydrophobic effect can be achieved in all cases by using a linearly additive, pairwise decomposable functional form.

(2) ChemScore evaluates protein−ligand hydrogen-bond quality based on geometric criteria, but otherwise does not distinguish between different types of hydrogen bonds or among the differing protein environments in which those hydrogen bonds are embedded.

(3) ChemScore does not treat desolvation effects.

(4) ChemScore uses a simple rotatable-bond term to treat conformation entropy effects arising from restricted motion of the ligand.

The new XP Glide scoring function starts from the "standard" terms discussed above, though the functional form of the first three terms have been significantly revised and the parameterization of all terms is specific to our scoring function. In the remainder of this section, the functional form and physical rationale for the novel scoring terms we have developed are described with examples from pharmaceutically relevant test cases provided to illustrate how the various terms arise from consideration of the underlying physical theory and experimental data.

Form of the XP Glide Scoring Function. The XP Glide scoring function is presented in eq 2. The principal terms that favor binding are presented in eq 3, while those that hinder binding are presented in eq 4. A description of each of the following terms besides $E_{\mathrm{hb_pair}}$ and $E_{\mathrm{phobic_pair}}$, which are standard ChemScore-like hydrogen bond and lipophilic pair terms, respectively, follows.

$$\text{XP GlideScore} = E_{\mathrm{coul}} + E_{\mathrm{vdW}} + E_{\mathrm{bind}} + E_{\mathrm{penalty}} \quad (2)$$

$$E_{\mathrm{bind}} = E_{\mathrm{hyd_enclosure}} + E_{\mathrm{hb_nn_motif}} + E_{\mathrm{hb_cc_motif}} + E_{\mathrm{PI}} + E_{\mathrm{hb_pair}} + E_{\mathrm{phobic_pair}} \quad (3)$$

$$E_{\mathrm{penalty}} = E_{\mathrm{desolv}} + E_{\mathrm{ligand_strain}} \quad (4)$$

Improved Model of Hydrophobic Interactions: Hydrophobic Enclosure ($E_{\mathrm{hyd_enclosure}}$). The ChemScore atom−atom pair function, $E_{\mathrm{phobic_pair}}$ described above, assigns scores to lipophilic ligand atoms based on summation over a pair function, each term of which depends on the interatomic distance between a ligand atom and a neighboring lipophilic protein atom. This clearly captures a significant component of the physics of the hydrophobic component of ligand binding. It is assumed that the displacement of water molecules from areas with many proximal lipophilic protein atoms will result in lower free energy than displacement from areas with fewer such atoms. As a crude example, it is clear that if the ligand is placed in an active-site cavity, as opposed to on the surface of the protein, the lipophilic atoms of the ligand are likely to receive better scores. If they are located in a "hydrophobic pocket" of the protein, scores should be better than in a location surrounded primarily by polar or charged groups. Furthermore, these improved scores are likely to be correlated with improvements in ligand binding affinity.

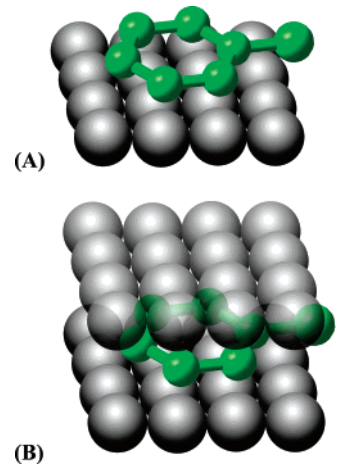

Figure 1. Schematic of a ligand group interacting with two distinct hydrophobic environments: above a hydrophobic "plane" (A) and enclosed in a hydrophobic cavity (B).

However, a function dependent only on the sum of interatomic pair functions is potentially inadequately sensitive to details of the local geometry of the lipophilic protein atoms relative to the ligand lipophilic atom in question. As an example, consider the two model distributions shown in Figure 1. In one case (A), a lipophilic ligand group is placed at a hydrophobic "wall" with lipophilic protein atoms on only a single face of the hydrophobic group. In the second case (B), the lipophilic ligand group is placed into a tight pocket, with lipophilic protein atoms contacting the two faces of the ligand group. As suggested above, one would normally expect a larger contribution to binding in the second case than in the first. However, this does not fully settle the question, which at root is whether the atom−atom pair contribution for a given ligand-atom/protein-atom distance should be identical when the ligand atom is enclosed by protein hydrophobic atoms, as opposed to when it is not, or whether there can be expected to be nonadditive effects.

From a rigorous point of view, the answer depends principally upon the free energy to be gained by displacing a water molecule at a given location. This in turn depends on how successfully that water molecule is able to satisfy its hydrogen-bonding requirements at that location, while retaining orientational flexibility. In the extreme case in which a single water molecule is placed in a protein cavity that can accommodate only one water molecule and is surrounded on all sides by lipophilic atoms that cannot make hydrogen bonds, the enthalpy gain of transferring the water to bulk solution is enormously favorable. In such a case it is not clear that a water molecule would occupy such a cavity in preference to leaving a vacuum, despite the statistical terms favoring occupancy. However, this is a rare situation not particularly relevant to the binding of a large ligand, whereas structural motifs similar to the examples in Figure 1 are quite common.

There have been a large number of papers in the literature studying, via molecular dynamics simulations, the behavior of

water in contact with various types of hydrophobic structures, including flat and curved surfaces,[20−22] parallel plates,[23,24] nanotubes,[25] and recently more realistic systems such as the hydrophobic surfaces of a protein or the interface between two protein domains.[26−28] There have also been attempts to develop general theories as to how the hydrophobic effect depends on the size and shape of the hydrophobic structure presented to the water molecules.[25,29]

A number of concepts that are clearly related to the proposals in the present paper have emerged from this work: evacuation of water (dewetting), under the appropriate conditions, from regions between two predominantly hydrophobic surfaces[1,2,9] and a model for the curvature dependence of the hydrophobic energy in which concave regions are argued to have greater hydrophobicity than convex ones.[29] However, while this work provides useful ideas and general background, development of a scoring function that can be used to quantitatively predict protein−ligand binding in the highly heterogeneous and complex environment of a protein active site requires direct engagement with a critical mass of experimental data as well as extensive parameterization and investigation of a variety of specific functional forms. In what follows, we describe the results of our investigations along these lines.

A large number of computational experiments involving modifications of the hydrophobic scoring term designed to discriminate between different geometrical protein environments have been performed. The criterion for success in these experiments is the ability of any proposed new term to fit a wide range of experimental binding free energy data and yield good predictions in enrichment studies. Key findings are summarized as follows:

(1) Ligand hydrophobic atoms must be considered in groups, as opposed to individually. The free energy of water molecules in the protein cavity is adversely affected beyond the norm primarily when placed in an enclosed hydrophobic microenvironment that extends over the dimension of several atoms. If there are individual isolated hydrophobic contacts, the water will typically be able to make its complement of hydrogen bonds anyway by partnering with neighboring waters as in clathrate structures surrounding small hydrocarbons in water.[28] After empirical experimentation, the minimum group size of connected ligand lipophilic atoms has been set at three.

(2) When a group of lipophilic ligand atoms is enclosed on two sides (at a 180 degree angle) by lipophilic protein atoms, this type of structure contributes to the binding free energy beyond what is encoded in the atom−atom pair term. We refer to this situation as *hydrophobic enclosure* of the ligand. There is some analogy here to the parallel plate, nanotube (with some sets of parameters), and protein systems in which dewetting has been observed, although the length scale of the region under consideration is smaller and (likely) more heterogeneous. The pair hydrophobic term in eq 1 is generally fit to data from a wide range of experimental protein−ligand complexes. As such, it represents the behavior of individual lipophilic ligand atoms in an "average" environment. Our new terms utilize specific molecular recognition motifs and are designed to capture deviations from this average that lead to substantial increases in potency for lipophilic ligand groups of types that are typically targeted in medicinal chemistry optimization programs. That is, placing an appropriate hydrophobic ligand group within the specified protein region leads to substantial increases in potency. Indeed, the data enabling development of this term was primarily obtained from a wide range of published medicinal chemistry efforts that provided examples of lipophilic groups that yielded exceptional increases in potency, as well as those yielding minimal increases. Our objective has been to explain these results on the basis of physical chemical principles and to develop empirical scoring terms that captured the essential physics while rejecting false positives, even with imperfect docking and the neglect of induced fit effects.

Calculation of the hydrophobic enclosure score, $E_{\mathrm{hyd_enclosure}}$, is summarized below with a more detailed description of the algorithm provided in Supporting Information:

(1) Lipophilic protein atoms near the surface of the active site and lipophilic ligand atoms are divided into connected groups. There are a set of rules specifying which atoms count as lipophilic and what delimits a group.

(2) For each atom in a group on the ligand, lipophilic protein atoms are enumerated at various distances.

(3) For each lipophilic ligand atom, the closest lipophilic protein atom is selected and a vector is drawn between it and the ligand atom. This is the protein "anchor" atom for that ligand atom. Vectors for all other suitably close lipophilic protein atoms are drawn to the ligand atom and their angles with the anchor-atom vector are determined. To be considered on the "opposite side" of the anchor atom, the angle between vectors must exceed a cutoff value that depends on the pair distance, with shorter distances requiring that the angle be closer to 180°. If the angle is close to zero degrees, the atom is on the "same side", and is at right angles to the anchor if the angle is close to 90°. When the angle between lipophilic protein atoms is close to 180°, we have argued this leads to an especially poor environment for waters.

(4) Each lipophilic ligand atom is assigned a score based on the number of total lipophilic contacts with protein atoms, weighted by the angle term. If no protein atom is greater than 90 degrees from the anchor atom, the angle term is zero and the atom contributes zero to the group's $E_{\mathrm{hyd_enclosure}}$ term. The overall score for a group is the sum over all atoms in that group of the product of the angular factor and a distance dependent factor.

(5) If the score for any ligand group is greater than 4.5 kcal/mol, the penalty is capped at 4.5 kcal/mol. This was an empirical determination based on investigating many test cases and comparing the results with experimental data. The capping is rationalized by arguing that if a very large region of this type leads to a score greater than 4.5 kcal/mol, there is probably some ability of the water molecules to compensate by interacting with each other.

An experimentally validated example of the gain in binding affinity from placing a large hydrophobic group in a pocket in which lipophilic protein atoms are present on both sides of the pocket (rings in both cases) is shown in Figure 2. Here, replacing a phenyl substituent with a naphthyl group was shown[30] to result in a 21-fold improvement in experimentally measured affinity (K_d). The naphthyl is required to fully occupy the hydrophobic pocket depicted in Figure 2.

As indicated above, the surrounding of ligand lipophilic atoms or groups by lipophilic protein atoms is referred to as hydrophobic enclosure. Our contention, here and in much of the following discussion of hydrogen bonding, is that proper treatment of hydrophobic enclosure is the key to discrimination of highly and weakly potent binding motifs and compounds. The underlying mathematical framework for describing enclosure, discussed above, could be cast in other forms, but the essential idea would remain unchanged. Detailed optimization of the numerical criteria for recognizing enclosure, and assigning

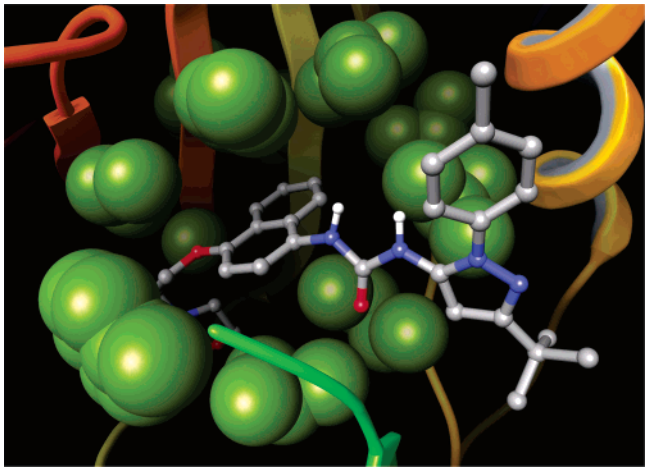

Figure 2. Boehringer active for 1kv2 bound to human p38 map kinase. The naphthyl group receives a −4.5 kcal/mol hydrophobic enclosure packing reward.

a specific contribution to the binding affinity for each motif is vital to developing methods with predictive capability.

Improved Model of Protein-Ligand Hydrogen Bonding. In developing a refined model of hydrogen bonding, we divide hydrogen bonds into three types, neutral−neutral, neutral−charged, and charged−charged. The analysis of each type of hydrogen bonding is different due to issues associated with the long-range solvation energy (Born energy) of charged groups. An initial step is to assign different default values (assuming optimal geometric features) to each of the three types of hydrogen bonds. The default values assigned are neutral−neutral, 1.0 kcal/mol, neutral−charged, 0.5 kcal/mol, and charged−charged, 0.0 kcal/mol. These assignments are based on a combination of physical reasoning and empirical observation from fitting to reported binding affinities of a wide range of PDB complexes.

The rationale for rewarding protein−ligand hydrogen bonds at all is subtle, because any such hydrogen bonds are replacing hydrogen bonds that the protein and ligand make with water. At best the net number of total hydrogen bonds on average will remain the same in the bound complex as compared to solution. However, the liberation of waters to the bulk can be argued to result in an increase in entropy, and liberation of waters around a polar protein group requires that a protein−ligand hydrogen bond with similar strength be made for a desolvation penalty to be avoided. This analysis is most plausible when both groups are neutral. The formation of a salt bridge between protein and ligand involves very different types of hydrogen bonding from what is found in solution. The thermodynamics of salt bridge formation in proteins has been studied extensively, both theoretically and experimentally,[31−34] and depends on many factors such as the degree of solvent exposure of the groups involved in the salt bridge. The default value of zero that we assign is based on the presence of many protein−ligand complexes in the PDB with very low binding affinities in which solvent-exposed protein−ligand salt bridges are formed. Assigning the contributions of these salt bridges to the binding affinity would lead to systematically worse agreement with experimental enrichment data. In XP scoring, certain features of a salt bridge are required for this type of structure to contribute to binding affinity in XP scoring. Finally, the charged−neutral default value represents an interpolation between the neutral−neutral and charged−charged value that appears to be consistent with the empirical data.

Hydrogen-bond scores are diminished from their default values as the geometry deviates from an ideal hydrogen-bonding geometry, based on both the angles between the donor and acceptor atoms and the distance. The function that we use to evaluate quality is similar to that used in ChemScore.

In what follows, specialized hydrogen-bonding motifs are described in which additional increments of binding affinity are assigned in addition to those from the ChemScore-like pairwise hydrogen-bond term. Our investigations indicate that these situations can arise for neutral−neutral or charged−charged hydrogen bonds, but not for charged−neutral hydrogen bonds. The exclusion of charged−neutral hydrogen-bond special rewards has principally been driven by our failure to date to identify motifs of this type that help to improve the agreement with experimental data. One can speculate that the lack of charge complementarity in charged−neutral hydrogen bonding precludes such structures from being major molecular recognition motifs, though further investigations with larger data sets will be needed to resolve this issue.

Special Neutral−Neutral Hydrogen-Bond Motifs ($E_{hb_nn_motif}$). In this section, neutral−neutral hydrogen-bonding motifs are described that were identified, based on both theoretical and empirical considerations, as making exceptional contributions to binding affinity. Such "special" hydrogen bonds represent key molecular recognition motifs that are found in many if not most pharmaceutical targets. Targeting such motifs is a central strategy in increasing the potency and specificity of medicinal compounds. Identifying such motifs through their incorporation in the scoring function should enable a dramatic improvement in both qualitative and quantitative predictions.

The critical idea in our recognition of special hydrogen bonds is to locate positions in the active-site cavity at which a water molecule forming a hydrogen bond to the protein would have particular difficulty in making its complement of *additional* hydrogen bonds. Forming such a hydrogen bond imposes nontrivial geometrical constraints on the water molecule. This is the basis for the default hydrogen-bond score, but such constraints become more problematic when the environment of the water molecule is challenging with respect to making additional hydrogen bonds such as those found in the bulk environment.

Our previous analysis of hydrophobic interactions suggests that the environment will be significantly more challenging if the water molecule has hydrophobic protein atoms on two faces, as opposed to a single face, and if few neighboring waters are available to readjust themselves to the constrained geometry of the protein−water hydrogen bond. Geometries of this type are identified using a modified version of the hydrophobic enclosure detection algorithm described previously. Replacement of such water molecules by the ligand will be particularly favorable if the donor or acceptor atom of the ligand achieves its full complement of hydrogen bonds by making the single targeted hydrogen bond with the protein group in question so that satisfaction of additional hydrogen bonds is not an issue. An example of a suitable group would be a planar nitrogen in an aromatic ring binding for example to a protein N−H backbone group. This has been observed to be essential to achieving high potency experimentally in the 1bl7 ligand binding to p38 MAP kinase, as shown in Figure 3. Here, the Met 109 hydrogen bond is known to be important for potency. Analogous hydrogen bonds have been found to be important in other kinases. In the absence of rigorous physical chemical simulations, we have used the experimental data from a significant number of diverse protein−ligand complexes to guide the development of a set of empirical rules, outlined below, for the types of ligand and

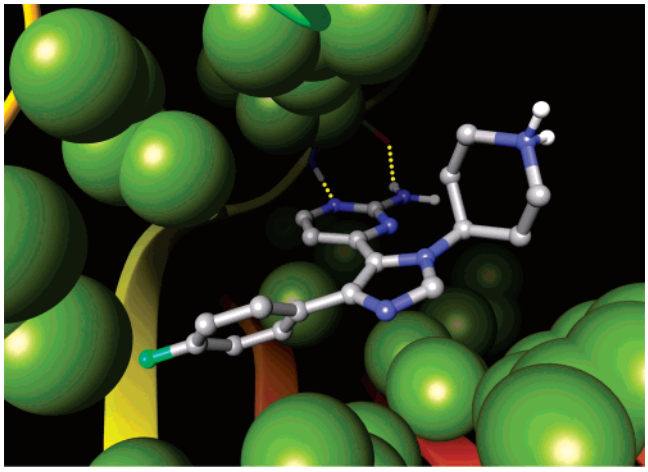

Figure 3. The 1bl7 ligand interacts with p38 MAP kinase through a neutral−neutral hydrogen bond between the ligand's aromatic nitrogen and the Met 109 N−H group.

receptor chemistries that receive this type of reward. These rules will likely evolve as more data is considered, and further simulations are undertaken.

The scoring-function term outlined above enables such hydrogen bonds to be detected automatically in advance of experimental measurement. Of equal importance, false positives, which superficially share some characteristics of the required structural motif but lack a key component, can be rejected automatically as well. Rejection of false positives has been optimized by running a given variant of the scoring function, identifying high scoring database ligands with special hydrogen bonds in locations not seen in known actives, examining the resulting structure, and altering the recognition function to eliminate the reward for such test cases.

In designing a detailed set of rules to implement the ideas outlined above, we have attempted to generalize results obtained from a wide variety of ligand−receptor systems, while at the same time avoiding false positives and respecting the basic physical chemistry principles that form the basis of the model. The detailed algorithm for detecting a single hydrogen bond in a hydrophobic environment is outlined as follows. In our implementation, the donor or acceptor atom must be in a ring with the exception of nitrogen, which is allowed to be a nonring atom. If the ligand atom is in a donor group, then all other donors of the group (e.g., the two hydrogen atoms of NH_2) must be hydrogen bonded to the protein. Only backbone protein atoms can participate in this type of special hydrogen bond. The unligated protein donor or acceptor must have fewer than three first-shell solvating waters where the waters are placed as outlined in section 3. If these criteria are met, the sum of the angular factor of the hydrophobic enclosure packing score described in this section is made over the hydrogen-bonded ligand heavy atoms and the carbon atoms attached to this ligand atom. The hydrophobic enclosure packing score used in this sum contains only the angular weight of the score described earlier and not the distance-based weight. If the absolute hydrophobic enclosure packing sum is above a defined cutoff, the hydrogen bond is considered to be in a hydrophobically constrained environment, and a special hydrogen-bond reward of 1.5 kcal/mol is applied to this hydrogen bond. In some instances, it is found that a small perturbation of the ligand can move the hydrophobic sum above the cutoff. Therefore, if the hydrophobic sum is below the cutoff, small rigid body perturbations of the ligand are made of 0.3 Å in magnitude. At each perturbed geometry, the sum is recalculated and the reward is applied if at some geometry the sum exceeds the cutoff. This procedure helps to avoid discontinuities inherent in the use of a cutoff.

Table 1.

conditions for not applying the special pair hydrogen-bond scores
ignore poor quality hydrogen bonds ($<0.05\ E_{hb_pair}$)
ignore pairs involving the same neutral protein atom
ignore pairs involved in a salt bridge if the electrostic potential at either ligand atom is above the cutoff
ignore salt bridge pairs if either protein atom is involved in a protein−protein salt bridge
ignore ligand donor/donor pairs that come from NH_x, where $x \geq 2$ groups, the nitrogen atom is not in a ring and has no formal charge
ignore formally charged protein atoms with more than eight second-shell waters in the unligated state
ignore charged−neutral hydrogen bonds unless the protein atom is in a salt bridge
ignore pairs of different neutral acceptor atoms on the ligand
neutral hydrogen-bond pairs must satisfy ring atom and hydrophobicity environment criteria as outlined in section 2
ignore ligand hydroxyl to protein hydrogen bonds if the protein atom has zero formal charge

The situation described above identifies a structure in which a single hydrogen bond should be assigned a "special" reward. A second situation occurs when there are multiple correlated hydrogen bonds between the protein and the ligand. The physical argument is that the organization of water molecules to effectively solvate a structure of this type in the confined geometry of the active site can be even more problematic than that for the single hydrogen-bond situation described above. However, this will occur only if the waters involved in such solvation are in a challenging hydrophobic environment, with hydrophobic groups on two sides. The coupling of the special hydrogen-bond identification with the hydrophobic enclosure motif is critical if false positives are to be rejected. Correlated hydrogen bonds are routinely formed in docking with highly solvent-exposed backbone pairs, but there is no evidence from the experimental data we have examined that such structures contribute to enhanced potency.

Pair-correlated hydrogen bonds are defined as a donor/acceptor, donor/donor, or acceptor/acceptor pair of ligand atoms (referred to as "ligand atom pair") that are separated by no more than one rotatable bond (hydroxyl groups count as nonrotatable in this calculation). Several restrictions on the types of pairs that can be considered are made as detailed in Table 1. If the ligand atoms of the pair individually have zero net formal charge, they must satisfy the following hydrophobicity criterion to achieve a special hydrogen-bond reward. First, the ligand atoms must be part of the same ring or be directly connected to the same ring. Assuming the pair satisfies these restrictions, the hydrophobicity of the hydrogen-bond region is detected in a manner similar though not identical to that for a single special hydrogen bond. A sum of the hydrophobic enclosure packing score described previously is made for the pair of hydrogen-bonding ligand heavy atoms and the ring atoms directly connected to the ligand pair atoms. If a ligand atom of the pair is not a ring atom but is connected to a ring, the sum includes atoms of the ring that are nearest neighbors to the nonring ligand atom. If the absolute hydrophobic sum is above a cutoff, the hydrogen-bonded pair is given a special reward of 3 kcal/mol. Note that double counting of pair and single special hydrogen bonds is avoided by checking pairs first and excluding any rewarded hydrogen bonds found from single hydrogen-bond consideration.

Finally, the special hydrogen-bond rewards are linearly reduced with the quality of the hydrogen bond. The hydrogen-

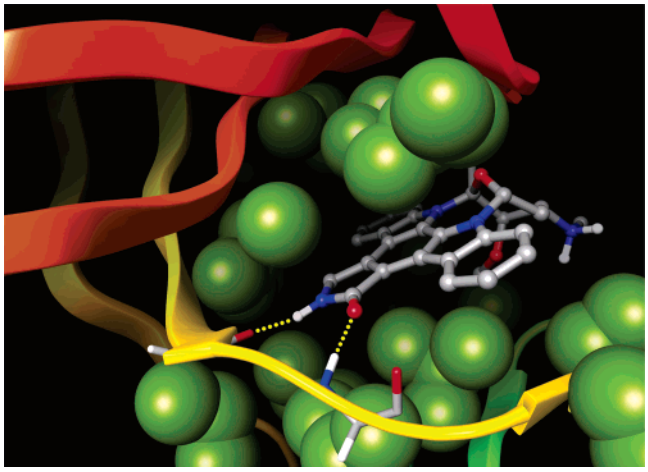

Figure 4. Staurosporine bound to human cyclin dependent kinase (CDK2). The pair of correlated hydrogen bonds receives a −3 kcal/mol reward, while the central component of the ring is given a −3 kcal/mol hydrophobic enclosure packing reward.

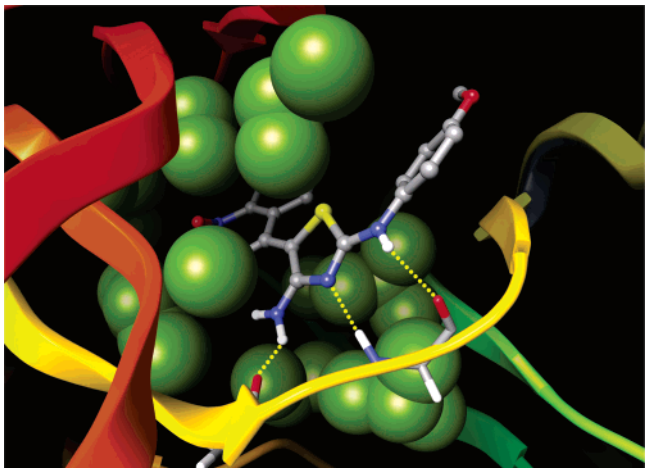

Figure 5. AG12073 bound to human cyclin dependent kinase (CDK2). The three correlated hydrogen bonds receive a −4.2 kcal/mol reward.

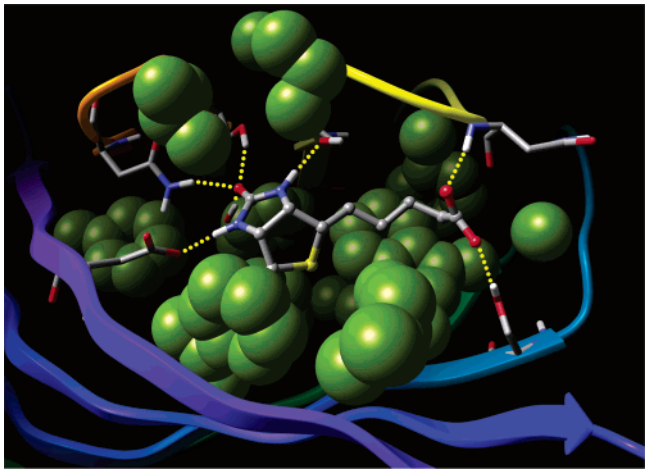

Figure 6. Biotin bound to streptavidin. The identification of a triplet of correlated hydrogen bonds in the ring in a hydrophobically enclosed region, and the three hydrogen bonds to the ligand carbonyl within that ring each contribute −3 kcal/mol rewards to this tightly bound complex ($\Delta G_{exp} = -18.3$ kcal/mol, XP binding = −18.2 kcal/mol).

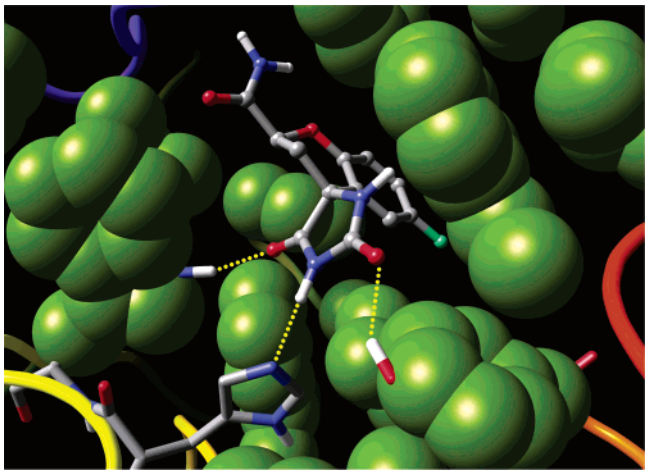

Figure 7. Fidarestat bound to aldose reductase. The triplet of special hydrogen bonds to the ring contributes −5.0 kcal/mol to the binding energy of this 9 nM inhibitor.

bond quality is measured in the sense of the pair hydrogen-bond score using the donor/acceptor distance and the angle made by the donor-heavy-atom-H vector and the H-acceptor vector. For ligand acceptor atoms in rings, the extent to which the acceptor lone pair vector is aligned with the donor-heavy-atom-H vector is evaluated. A detailed description of the algorithm for scaling the special hydrogen-bond reward is given in Supporting Information.

A substantial number of protein−ligand complexes in which motifs containing correlated hydrogen bonds that satisfy the above criteria, including the requisite hydrophobic enclosure, have been identified. A number of examples are shown below. Figure 4 depicts the 1aq1 structure of staurosporine bound to human cyclin-dependent kinase (CDK2). This type of correlated pair is also found in a number of other kinases. Some of the CDK2 actives, such as AG12073, make three correlated hydrogen bonds; this structure is shown in Figure 5. A second example is streptavidin bound to the 1stp structure of biotin (Figure 6), with three correlated hydrogen bonds in a hydrophobically enclosed region. To our knowledge, no empirical scoring function has explained the exceptionally large binding affinity of streptavidin to biotin. However, once the correlated hydrophobically enclosed hydrogen-bonding motif is recognized and assigned an appropriate score (a reward of 3 kcal/mol, consistent with other examples), the deviation between calculated and experimental binding affinity, using a docked structure, is only 0.1 kcal/mol. It should be noted that the high accuracy of this prediction is fortuitous and is not intended to suggest an ability to rigorously rank order compounds. Instead, the intent is to contrast the qualitatively reasonable prediction with that of alternative scoring functions, which typically yield results for this complex in error by ∼5−10 kcal/mol, for example, as reported in ref 35. The triply correlated, enclosed hydrogen-bonding motif also explains the low nanomolar binding affinity in the binding of fidarestat to the 1ef3 structure of aldose reductase (Figure 7) relative to a large number of ligands that achieve similarly large lipophilic scores in the highly hydrophobic active site, yet have only micromolar affinity.

In our studies of various pharmaceutically relevant targets, the combination of hydrophobic enclosure with one to three correctly positioned hydrogen bonds is characteristic of every "special" neutral−neutral hydrogen-bond motif that leads to an exceptional increase in experimentally measured potency. However, additional characteristics are required to eliminate false positives. In particular, if the hydrogen-bond partner in the protein is highly solvent exposed, formation of a structure capable of solvating the group in question, while still allowing the waters involved to form a suitable number of additional

Table 2. Special Hydrogen-Bond Reward Values

hydrogen-bond moiety	reward (kcal/mol)
single bond to ring in hydrophobic environments	1.5
neutral pair in hydrophobic environments	3.0

hydrogen bonds, becomes easier. Thus, we require that the protein group(s) involved in the special hydrogen bond(s) have a limited number of waters in the first or second shell. Determination of the number of surrounding waters is carried out via the water addition code described later in this section.

The magnitude of the rewards associated with the special hydrogen bonds has been determined by optimization against a large experimental database containing a significant number of examples of each type of structure. Values are given in Table 2. It is conceivable that finer discriminations, depending upon the details of the donors and acceptors, hydrophobic environment, bound waters, and so on, could be developed, along with a correspondingly more elaborate scoring scheme. However, the present relatively simple scheme appears to work remarkably well, at least at the level of discriminating active from inactive compounds (as opposed to rank ordering, which we have not yet examined in detail) in a wide variety of test cases.

Special Charged−Charged Hydrogen-Bond Motifs ($E_{hb_cc_motif}$). We have identified the following features that signal enhanced binding affinity from charged−charged hydrogen bonds:

(1) The number of waters surrounding the protein component of the salt bridge. Charged groups that are fully exposed to solvent are unlikely to participate in enhanced charged−charged hydrogen bonding because the cost of displacing the solvent is simply too large. Solvent exposure is calibrated using our water scoring code, described later in this section (see E_{desolv}), by examining the number of waters in the first two shells surrounding the charged protein group.

(2) The number of charged−charged hydrogen bonds made by the charged ligand group. Three different types of salt-bridge structures have been observed: (a) Monodentate (single hydrogen bond) between the ligand group and a protein group. (b) Bidentate (two hydrogen bonds) between the ligand group and a protein group. An example of a bidentate salt bridge occurs in the 1ett structure of thrombin between a positively charged amine group and a recessed Asp 189 carboxylate in the relevant specificity pocket, as displayed in Figure 8. (c) Hydrogen bonds of one ligand group to two different protein groups. This requires having two like-charged protein groups in close proximity. This structure, which presumably creates strain energy in the apo protein, occurs with a greater frequency than might be expected. Figure 9 presents an example showing ligand Gr217029 binding to the tern N9 influenza virus of the neuramidase receptor (1bji) with a distance between carboxylate oxygen atoms of only 4.5 Å.

Empirical observations, such as the unexpectedly high potency of several neuramidase ligands including Gr217029 cited above, and physical chemical reasoning in that the electric field from the two nonsalt-bridged, proximate carboxylates is highly negative and interacts more favorably with a ligand positive charge than is typical for a salt bridge suggest that (c) provides less stabilization energy than (b), which in turn provides less stabilization energy than (a). Similarly, one would expect that a bidentate structure is more favorable electrostatically than a monodentate structure. Note, however, that unless consideration (1) is properly satisfied, none of the three structures is likely to be favorable from a free energy point of view. It is the combination of restricted water access for the protein group and an exceptionally strong electrostatic interaction between the ligand and protein group that creates the molecular recognition motif.

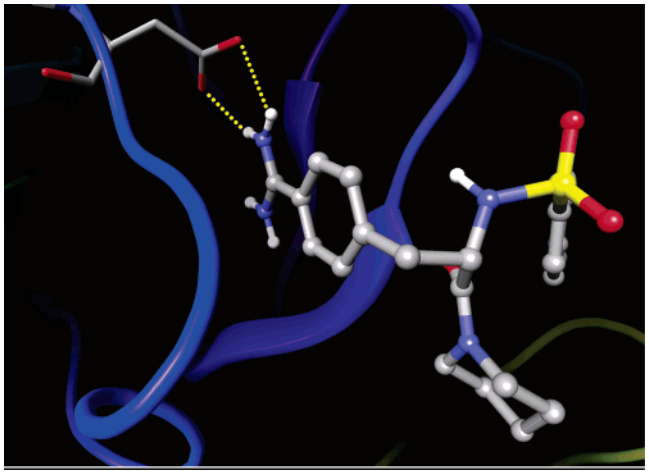

Figure 8. The bidentate hydrogen bonds in this thrombin complex bridge the ligand and Asp 189.

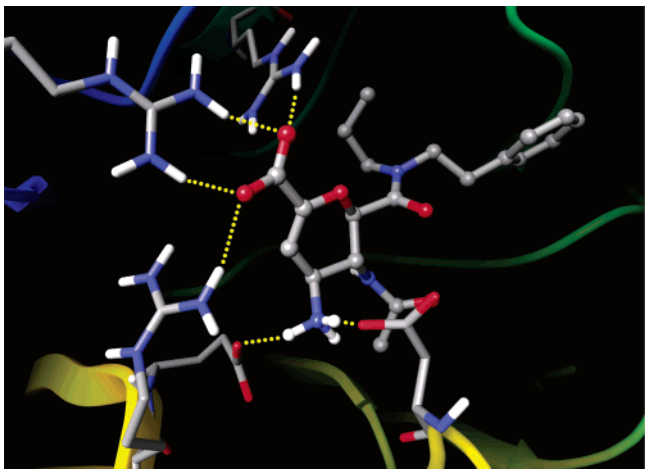

Figure 9. Gr217029 forms hydrogen bonds with two nearby Asp residues when bound to tern N9 influenza virus neuraminidase.

(3) Zwitterion ligands. A principal reason that the default value for charged−charged hydrogen bonds is set at zero is that, in forming a salt bridge, both the protein and ligand must surrender long-range contributions to the Born energy (i.e., those beyond the first shell). Satisfying the first-shell complement of hydrogen bonds is quite possible when forming a salt bridge, but the replacement of bulk water with the protein, or bound waters, clearly reduces the possible dielectric response to the ion. For a monovalent ion, the unscreened Coulomb field decreases as 1/r. Even though past the second shell dielectric screening substantially reduces further contributions, long-range effects make a nontrivial contribution to the total solvation-free energy. However, for a zwitterion, the fields from the positive and negative charges to some extent cancel at long range, yielding a dipolar field for which the second- and higher-shell contributions to the solvation free energy are significantly reduced. This cancellation depends on the separation of the two charged groups. Thus, formation of two salt bridges by the zwitterion, particularly if the two oppositely charged moieties in the ligand are spatially proximate, should be more favorable than binding a single ion. An example of a zwitterion binding in this fashion is shown in Figure 9.

Table 3. Electrostatic Rewards; Note that Double Counting Is Avoided

charge interaction	reward (kcal/mol)
charged ligand atom in low electrostatic potential environments	1.5
zwitterion configuration, range of rewards increasing with electrostatic attraction	3.0 to 4.7
positive ligand group binding to weakly solvated negative protein group	0.5
ligand CO_2- group hydrogen bound to multiple proximate positive protein residues	1.0
salt bridge pair in low solvation environment (less than nine second-shell waters about pre-ligated charge protein atom)	2.0

Table 4. 4.0 XP Binding Energies for Docking into Various GluR2 Receptors Compared to Experiment[a]

ligand	XP score	ΔG_{exp} (kcal/mol)
1ftl	−8.9	−8.3
1pwr	−12.6	−13.0
1ftj	−6.6	−8.5
1mm7	−10.7	−12.7
1mqi	−11.9	−11.9
1ftm	−11.3	−8.8
1n0t	−6.6	−5.7
1m5b	−9.3	−9.3
1m5c	−9.4	−9.3
1m5e	−6.4	−6.3

[a] These systems were used to calibrate the charged−charged hydrogen-bond recognition motif ($E_{hb_cc_motif}$).

(4) Cases where the ligand is positively charged and the protein is negatively charged are distinguished from those in which the charge states are reversed.

(5) Strength of the electrostatic field at the ligand. An enhanced binding affinity for a salt bridge is assigned if the site at which the ligand charge is placed is sufficiently electrostatically favorable. The electrostatic field at the ligand site is summed using constant and distance-dependent dielectric models, and cutoffs are imposed for assigning rewards based on empirical optimization over our suite of test cases. These cutoffs help reduce the number of false positives receiving special charged−charged rewards.

Table 3 enumerates the various special charged−charged rewards for motifs based on the five categories discussed above. The numerical values have been optimized based on fitting to our entire test suite.

Table 4 displays the XP active scores for a series of Glur2 receptors versus the experimental binding energies. Along with neuramididase, this is a system for which electrostatic interactions are particularly important. As such, it provides an important contribution to the training set. The good agreement displayed was achieved by using a combination of the electrostatic terms discussed above.

Other Terms. A number of other types of specialized terms have been investigated. These include terms rewarding pi stacking and pi-cation interactions (E_{PI}), rewards for halogen atoms placed in hydrophobic regions, and an empirical correction enhancing the binding affinity of smaller ligands relative to larger ones. These parameterizations were in many cases performed using limited data, and we do not view them yet as fully mature. As such, details will not be presented in the present publication. These terms are relatively small compared to the enclosure and charged−charged terms discussed above, but can have a nontrivial impact in specific cases. For example, the pi-cation term, for which a reward of 1.5 kcal/mol is assigned, is important for the acetylcholinesterase test case discussed below.

Terms that Penalize Binding in the XP Scoring Function. The most important physical effects that oppose binding are strain energy of the ligand, protein, or both, loss of entropy of ligand and protein, and desolvation of the ligand or protein. The penalty terms developed are targeted in all three areas, although terms addressing strain energy and entropic loss do not necessarily represent a significant advance as compared to previous terms described in the literature. In developing the penalty terms, some fundamental limitations arise from the rigid-receptor approximation and the use of empirical scoring functions rather than full energy expressions. One consequence is that, in our view, it is not possible to completely reject false positives with an empirical approach. However, significant improvements are possible as compared to previous efforts, as we shall demonstrate below.

Water Scoring: Rapid Docking of Explicit Waters (E_{desolv}). A number of approaches to incorporating desolvation penalties into a high throughput docking code have been presented in the literature.[35,36] However, the methods in these papers are based on continuum-solvation approaches. For computing protein−ligand binding affinities, the role of individual waters can be critical, and continuum models often provide poor results in treating bound waters in a confined cavity. Therefore, we have chosen to implement a crude explicit water model that can be rapidly evaluated yet captures the basic physics of solvation within the confines of the protein−ligand complex active-site region.

The approach employed is to use a grid-based methodology and add 2.8 Å spheres, approximating water molecules, to high-scoring docked poses emerging from the initial round of XP docking. In principle, this methodology is similar to that used in the GRID program, though, algorithmic details have been optimized to achieve speed, critical in the present application. The CPU time for water addition to a single pose is 3−8 CPU seconds (AMD Athelon MP 1800+ processor running Linux) on average depending on ligand size.

Once waters have been added, statistics are tabulated with regard to the number of waters surrounding each hydrophobic, polar, and charged group of the ligand and active site of the protein. When a polar or charged ligand or protein group is judged to be inadequately solvated, an appropriate desolvation penalty is assessed. Additionally, the environment of each active-site water is itself probed to search for cases in which waters make an unusual number of hydrophobic contacts. A penalty is assigned if the number of such contacts for an individual water molecule exceeds a given threshold. Water scoring statistics are also used to determine whether special hydrogen-bonding rewards should be assigned, as was discussed previously.

Contact Penalties (E_{lig_strain}). Penalizing strain energy in rigid-receptor docking is probably the single most difficult component of an empirical scoring function. The problem arises from the fact that, in a typical cross-docking situation, the ligand has to adjust to fit into an imperfect (from its point of view) and rigid cavity. This often requires ligands to adopt higher-energy, nonideal torsion angles. Considering the rigid-receptor approximation that is made, it is difficult to determine whether strained ligand geometries would arise if induced-fit effects were properly accounted for or whether strained ligand geometries would be a true requirement for docking to that receptor. Furthermore, even native ligand geometries have been found to exhibit high strain energies.[37] Given this limitation, the function used to penalize poses with close internal contacts is fairly lenient and only looks for severe cases of bad internal

contacts. One function simply counts the number of intramolecular heavy atom contacts below roughly 2.2 Å and rejects a pose entirely if there are more than three such contacts. A second, more sophisticated function assembles the contacts into groups and evaluates a penalty based on the size of the contacting groups, the range of contacts, and the extent to which the groups lie on the periphery of the molecule. Empirically, it has been found that peripheral groups are more difficult to penalize for intraligand contacts than are more centrally located groups.

Implementation Issues. Application of XP penalty terms, particularly those related to desolvation, imposes hurdles that make it difficult for random database ligands to achieve good scores in virtual screening. These hurdles do not exist in alternative programs that ignore desolvation effects and strain energy. On the other hand, if the terms are inaccurately defined, they will adversely affect active compounds. Furthermore, a definition that would be "accurate" for a high-resolution structure may function poorly for docked structures, particularly in the rigid-receptor framework, due to inaccuracies in sampling. This issue arises routinely in practice, such as when an active ligand could avoid a penalty by moving a few tenths of an Angstrom in some direction but is blocked by the rigid protein. Similarly, the sampling algorithm may simply fail to find the superior pose.

We have found that extensive sampling to enable ligands to avoid penalties when possible is an essential component of Glide XP scoring. If the penalties are due to limitations in the rigid-receptor approximation, the only solutions are (a) to adjust the parameters so that they allow the test case in question to escape penalization or (b) to accept that the active compound does not "fit" into the particular version of the receptor being used and to dock into multiple structures and/or employ induced-fit methodologies. For many cases, however, better sampling in the relevant phase space can locate ligand geometries that are able to avoid the penalties. The XP Glide sampling algorithm was explicitly designed with this objective in mind.

XP Glide Sampling Methodology. XP Glide sampling begins with SP Glide docking, as described in refs 1 and 2, but using a wider "docking funnel" so that a greater diversity of docked structures is obtained. For XP docking to succeed, SP docking must produce at least one structure in which a key fragment of the molecule is properly docked. This has been the case in the great majority of systems that have been investigated.

The second step in XP sampling is to assign various fragments of the molecule as "anchors" and to attempt to build a better-scoring pose for the ligand starting from each anchor. Typical anchors are rings, but can be other rigid fragments as well. Various positions of the anchors are clustered, representative members of each cluster are chosen, and the growing of side chains from relevant positions on the anchor is initiated.

The growing algorithm proceeds one side chain at a time, thereby avoiding the combinatorial explosion of total molecular conformations that occurs when all side chains are considered together. Because the anchor fragment is already positioned in the protein, most side-chain conformations can be trivially rejected based on steric clashes. The Glide "rough scoring" function is used to screen the initial side chain conformations. Most importantly, these conformations can be grown at extremely high resolution (4 degrees for each rotatable bond) because the total number of conformations considered at any one time is being constantly pruned via screening and clustering algorithms. It is this high-resolution sampling that enables difficult cross-docking cases to be effectively addressed, and that ultimately allows penalties to be avoided when possible.

After the individual side chains are grown, a set of candidate complete molecules is selected by combining high-scoring individual conformations at each position and eliminating structures with significant steric clashes between side chains. Candidate structures are minimized using the standard Glide total-energy function, which employs a distance-dependent dielectric to screen electrostatic interactions and are ranked according to the Emodel Glide pose-selection function,[1] comprising the molecular mechanics energy plus empirical scoring terms. Then, the grid-based water addition technology is applied to a subset of the top scoring structures, penalties are assessed as discussed above, and the full XP scoring function is computed.

At this stage, structures with the largest contributions from terms that promote binding may have penalties of various types. The next stage of the algorithm, critical to obtaining suitable results, is to attempt to evade penalties by regrowing specific side chains from such poses. The side chain that is the cause of the penalty can be targeted and, by focusing on this region of the ligand exclusively, a much larger number of candidate structures covering this region of phase space can be retained and minimized. The algorithm results in a significant increase in the density of poses and locates penalty-free structures when possible, despite the fact that the penalty terms are not at present encoded in the energy gradient. Finally, a single pose is selected based on a scoring function that combines weighted Coulomb/van der Waals protein−ligand interaction energies, the terms favoring binding affinity, and the various penalty terms.

XP Glide Parameterization: Philosophy and Implementation. The novel terms that we have described above have been developed via a combination of reasoning from basic physical chemistry principles and examining a large set of empirical data, as discussed further below. Because the terms are calculated via fast empirical functions (as opposed to rigorous atomistic simulations), extensive parameterization is required to obtain results in reasonable agreement with experiment. The large number of parameters employed in turn necessitates the use of a large number of examples in the training set; to avoid overfitting, the training set must be substantially larger than the number of parameters that are adjusted.

The total number of parameters in the current XP scoring function is on the order of 80; this includes parameters for desolvation penalties, hydrophobic enclosure, special neutral−neutral and charge−charge hydrogen bonds, and pi-cation and pi-stacking interactions. Parameters are required to convert various geometrical criteria into specific scores. The PDB complexes below, as well as the enrichment studies in the training set, were used to develop the parameter values. False positives, as well as known actives, were incorporated into the optimization protocol, so the data in the training set exceeds the total number of PDB complexes and known actives by a considerable margin (although a precise calculation of the total number of data points in the training set is very difficult to produce, as for example not every database ligand was competitive with known actives in ranking, and noncompetitive compounds played no role in parameter optimization.

Because the scoring function contains nonlinear functional forms, a rigorous optimization algorithm would also be nonlinear (rather than a simple least-squares fit); furthermore, constraints would be imposed based on physically reasonable values of the various parameters. The present set of parameters was in fact determined by a heuristic approach; a small number of

paradigmatic test cases were identified for each type of term, initial values were fit to yield reasonable results for these parameters, and the parameter set was then tested on the entire training set. Problem cases were then identified by these tests, and reoptimization was carried out to improve the worst outliers. A fully numerical optimization protocol would quite likely improve results for the training set, but it is unclear whether a corresponding improvement in the test set would result (note that test set results were not generated until the current parameter set was frozen in the Glide 4.0 release). As we develop a larger test set, we will investigate the use of more automated optimization algorithms, with likely quantitative improvements in the predictive capabilities of the scoring function.

3. Structural and Binding Affinity Prediction Results for PDB Complexes

A set of 268 complexes from the PDB, which we have previously used to assess the docking accuracy and scoring capabilities of Glide SP,[1] have been selected, of which 198 have reliable experimentally determined binding affinities as determined by our extensive examination of the literature for each case. This set of complexes displays a wide range of active-site cavities and protein−ligand interactions. The parameters of the scoring function were simultaneously optimized to reproduce the experimental binding affinity data and yield quality enrichment factors/binding affinities for the database screening tests that are discussed below.

An evaluation of the docking accuracy of Glide SP was presented in ref 1, and the XP results from docking MMFFs[38] optimized ligand structures shown in Table 5 are very similar. This suggests that the sources of error in docking accuracy are due to issues other than those addressed by the XP scoring-function modifications. In some cases, near symmetry in the ligand leads to docked poses that are functionally equivalent to those in the crystal structure, for example in terms of protein−ligand contacts, but that exhibit a large RMSD; such cases are identified in Table 5. A principal source of errors in pose RMSD appears to be the charge distribution of the ligand, which, in a standard force field representation, may not accurately distribute formal ionic charges and does not incorporate polarization effects. Cho and co-workers have demonstrated that low RMSD ligand poses can be reliably generated by utilizing more accurate polarized charges, where the ligand charge distribution is computed in the protein environment via QM/MM methods.[39] We may incorporate this methodology into future XP Glide releases for optional use. The quality of structural prediction shown by XP Glide is sufficient for the great majority of ligands to enable an adequate assessment of the scoring function to be made.

In our optimization protocol, we employed docked protein−ligand complexes, retaining only those with protein−ligand contacts that predominantly agree with those in crystal structures. In this fashion, we avoid corrupting the fitting process with irrelevant data such as would be provided by a grossly incorrect pose, yet include a realistic level of variation in the input structure. This is particularly critical in optimization of the penalty terms. If the sampling algorithm cannot avoid incorrect penalties in self-docking, it is unlikely to be able to do so in a much more challenging cross-docking situation. By incorporating docked poses of PDB complexes into the optimization process, the penalty function can be tuned to improve the agreement with experimental binding affinities while avoiding inappropriately penalizing active compounds, keeping in mind that there are also cases where the penalty terms are in fact appropriate.

Before discussing binding affinity predictions, a key point that has generally been neglected previously should be noted. An empirical scoring function that considers only protein−ligand interactions with no a priori information concerning the apo structure of the protein cannot, by definition, take into account the reorganization energy of the protein required to accommodate the ligand. In many cases, the protein is relatively rigid, the ligand fits without major rearrangements, and neglect of this term is acceptable. However, there are cases where it is overwhelmingly likely that the induced-fit energies are substantial. The most obvious cases are those in which an allosteric pocket is created to accommodate the ligand. This occurs, for example, in nonnucleoside reverse transcriptase inhibitors (NNRTIs) of HIV reverse transcriptase (HIV-RT) and is also manifested in the large-scale motion of the activation loop in p38 MAP kinases required to produce the DFG-out conformation to which inhibitors such as BIRB796 bind. However, there are many more subtle cases in which side chains or backbone groups alter their positions nontrivially, and this should introduce some energetic cost.

The goal of an empirical scoring function should be to predict the binding affinity of the ligand to the *structure with which it is presented.* That is, given the formal impossibility of predicting reorganization energy in any such scheme, this value should be removed from the experimental binding affinity. Otherwise, one will be fitting to an incorrect experimental number, given the goal of the exercise. A perfect scoring function of this type will correctly rank-order candidate ligands in their ability to bind to the structure at hand. If this structure is known to have a low reorganization energy, compounds with good scores when docked into that structure should yield satisfactory experimental binding affinities. For allosteric pockets and other sites with larger reorganization energies, one would expect that more favorable scores would be needed to yield the desired experimental binding affinity. The problem of comparing scores between ligands docked into different conformations of the receptor can then be treated separately. The problem is highly nontrivial, requiring either a heuristic procedure incorporating experimental information or the brute force ability to compare free energies of different protein conformations.

A qualitative observation that we have made, confirmed in a large number of examples, is that a large hydrophobic enclosure score is a signature of significant protein rearrangement and possibly creation of an allosteric pocket. Ligands binding tightly to both the HIV-RT NNRTI site and p38 DFG-out conformation, for example, generally receive maximal hydrophobic-enclosure scores. Furthermore, the total XP scores of these ligands are substantially higher in absolute terms than their experimental binding affinities would mandate. This is completely consistent with the ideas discussed above, in which the reorganization energy of the protein must be subtracted from the empirical binding affinity score to produce the correct experimental binding affinity.

In the initial Glide XP parameterization with PDB cocrystallized structures, systems with exceptionally large protein reorganization energies, as predicted by large hydrophobic enclosure contributions, have been intentionally omitted. This omission includes allosteric sites such as the HIV-RT and p38 structures mentioned above, and complexes such as staurosporine/CDK2 (PDB code 1aq1), in which the CDK2 pocket must expand substantially to accommodate the unusually large staurosporine ligand. If one believes that the scoring function is accurate, the protein reorganization energy can be inferred from the computed rigid receptor and experimental binding

Table 5. 4.0 XP and SP Binding Scores and Heavy Atom RMS (Å) Values for Docking into PDB Entries[a]

		GlideScore (kcal/mol)			RMS (Å)				GlideScore (kcal/mol)			RMS (Å)	
PDB	ΔG_{exp}	XP	XP-corr	SP	XP	SP	PDB	ΔG_{exp}	XP	XP-corr	SP	XP	SP
1aaq	−11.5	−10.6	−10.6	−11.5	2.01	1.40	1e5i		−7.4	−7.4	−11.2	0.28	0.17
1abe	−8.9	−7.8	−7.8	−8.8	0.31	0.40	1eap	−8.5	−12.6	−11.8	−9.2	0.65	2.38
1abf	−7.4	−8.0	−8.0	−9.3	0.17	0.14	1ebg	−14.8	−11.0	−11.0	−18.3	0.34	0.26
1acj	−10 0	−11.4	−9.9	−8.4	2.81	4.61	1ecv	−6.6	−7.4	−7.4	−9.4	0.24	0.18
1acm	−10 3	−12.2	−12.2	−13.1	0.40	0.32	1eed	−6.5	−12.0	−12.0	−8.3	11.29	1.58
1aco	−4.9	−2.3	−2.3	−10.1	0.34	0.29	1ejn	−7.7	−9.8	−7.8	−10.0	0.34	0.12
1add	−9.2	−10.3	−8.3	−9.9	0.83	0.70	1ela	−8.7	−8.4	−8.4	−5.1	0.39	6.01
1adf	−6.2	−7.9	−7.9	−11.8	3.03	9.92	1elb	−9.8	−6.3	−6.3	−7.0	5.42	4.29
1aha		−7.9	−7.9	−8.2	0.36	0.11	1elc	−9.4	−7.3	−7.3	−6.0	6.53	7.92
1ake		−14.0	−14.0	−3.9	15.45	14.95	1eld	−9.1	−6.5	−6.5	−4.8	0.32	3.94
1apb	−7.9	−7.7	−7.7	−9.4	0.06	0.10	1ele	−9.3	−8.0	−8.0	−6.4	0.36	0.38
1apt	−12.8	−11.9	−11.9	−11.0	1.48	1.24	1epb		−13.6	−11.1	−8.0	2.24	1.89
1apu	−10.2	−8.7	−8.7	−7.6	0.61	1.24	1eta		−3.9	−3.9	−3.5	8.69	1.85
1apv	−12.3	−11.2	−11.2	−8.5	0.63	0.48	1etr	−10.1	−11.7	−9.7	−9.4	0.71	0.68
1apw	−10.9	−11.0	−11.0	−9.0	0.97	0.32	1ets	−11.2	−13.7	−11.7	−11.7	1.32	1.44
1atl	−8.6	−10.6	−10.6	−7.8	0.87	3.47	1ett	−8.0	−12.6	−10.6	−9.5	0.62	0.58
1avd		−16.4	−16.4	−10.4	0.82	0.55	1ezq	−12.3	−11.4	−9.4	−12.6	0.75	0.21
1azm		−5.1	−5.1	−6.2	1.89	2.51	1f0r	−10.4	−13.3	−11.3	−10.1	2.11	0.59
1b6j	−10.8	−18.3	−18.3	−15.8	2.98	0.43	1f0s	−10.6	−11.3	−11.3	−9.4	2.08	0.35
1b6k	−11.9	−14.8	−12.8	−13.8	1.04	1.06	1f0t	−8.2	−7.7	−7.7	−9.8	0.38	0.24
1b6l	−11.3	−11.1	−11.1	−8.7	0.92	1.18	1f0u	−9.8	−10.8	−8.8	−10.5	1.56	1.54
1b6m	−11.5	−13.9	−11.9	−11.4	0.73	3.17	1fen		−12.8	−12.1	−8.4	1.10	0.40
1baf		−9.5	−9.5	−8.3	1.17	1.08	1fh8	−9.4	−10.7	−10.7	−10.3	0.20	0.20
1bap	−9.3	−8.2	−8.2	−9.1	0.39	0.38	1fh9	−8.8	−6.0	−6.0	−5.3	2.11	1.99
1bbp		−14.3	−12.4	−11.8	5.28	5.11	1fhd	−9.3	−8.1	−8.1	−8.7	5.37	0.46
1bkm		−11.6	−11.6	−14.6	4.77	2.36	1fjs	−13.6	−13.6	−11.6	−12.5	2.06	2.42
1bma	−6.3	−7.9	−7.9	−7.5	0.68	1.94	1fkg	−10.9	−13.1	−12.6	−8.4	1.21	1.33
1bra	−2.5	−5.5	−3.5	−7.8	2.26	0.32	1fki	−9.5	−10.2	−10.2	−7.7	1.30	1.29
1byb	−19.0	−14.2	−14.2	−11.3	0.56	0.46	1fq5	−11.5	−17.0	−17.0	−14.0	1.96	2.43
1c1b		−18.2	−15.7	−10.8	0.91	0.45	1fvt		−13.2	−13.2	−8.4	0.88	0.88
1c3i		−12.6	−12.6	−11.9	0.61	0.43	1g45	−11.8	−8.2	−8.2	−6.0	7.88	4.02
1c5p	−6.4	−6.2	−6.2	−8.6	0.27	0.25	1g46	−12.1	−8.2	−8.2	−6.3	8.06	4.50
1c83	−6.6	−7.3	−7.3	−9.9	0.17	0.14	1g48	−11.5	−7.0	−7.0	−5.9	1.88	3.77
1c84	−6.8	−7.5	−7.5	−8.8	0.26	0.32	1g4j	−11.9	−6.9	−6.9	−6.9	5.56	3.51
1c86	−6.4	−8.4	−8.4	−10.9	0.19	0.18	1g4o	−11.3	−7.6	−7.6	−6.0	3.39	4.21
1c87	−5.7	−8.1	−8.1	−10.9	0.28	0.21	1g52	−13.0	−8.2	−8.2	−5.7	8.01	4.26
1c88	−7.2	−7.5	−7.5	−11.8	0.25	0.22	1g53	−12.3	−8.4	−8.4	−6.4	7.88	4.53
1c8k		−8.1	−8.1	−7.3	3.28	5.50	1g54	−12.0	−7.2	−7.2	−5.4	8.45	5.14
1cbs	−9.8	−7.5	−7.5	−7.5	0.63	0.39	1ghb	−1.7	−3.2	−3.2	−9.7	0.45	0.30
1cbx	−8.7	−7.8	−7.8	−13.1	0.28	0.48	1glp		−4.5	−4.5	−8.1	0.75	0.32
1cde		−15.2	−15.2	−11.5	1.62	1.71	1glq		−6.2	−6.2	−9.7	0.46	0.32
1cdg	−3.3	−2.4	−2.4	−4.8	6.48	9.83	1gsp		−7.7	−7.7	−7.8	1.10	2.79
1cil	−12.9	−5.4	−5.4	−6.2	3.61	3.92	1hbv	−8.7	−12.1	−12.1	−5.5	2.19	2.07
1cnx	−10.0	−6.8	−6.8	−7.5	6.54	6.53	1hdc	−8.2	−9.8	−9.8	−8.3	0.56	0.35
1com	−5.4	−7.3	−7.3	−9.0	0.55	3.74	1hdy	−7.8	−4.8	−4.8	−4.2	1.65	1.70
1coy		−8.7	−8.7	−9.1	0.44	0.29	1hef	−12.1	−11.2	−11.2	−10.3	6.24	6.43
1ctr	−5.8	−7.5	−7.5	−5.7	2.27	2.59	1hfc	−7.5	−8.9	−8.9	−9.4	2.25	2.26
1ctt	−6.2	−6.9	−6.9	−7.0	0.60	5.04	1hgg	−3.4	−6.4	−6.4	−7.2	1.37	1.47
1d3d	−12.4	−12.7	−12.7	−10.4	1.52	2.74	1hgh	−3.9	−5.7	−5.7	−6.2	4.70	0.49
1d3p	−8.9	−9.6	−9.6	−10.3	1.72	2.05	1hgi	−3.7	−3.9	−3.9	−6.9	0.37	0.23
1d7x		−7.2	−7.2	−9.1	0.58	0.44	1hgj	−2.3	−4.8	−4.8	−4.0	0.64	0.34
1d8f		−6.8	−6.8	−6.8	4.25	4.31	1hih	−11.0	−11.8	−11.8	−11.0	1.26	1.29
1dbb	−12.3	−13.2	−13.0	−10.3	0.37	0.41	1hps	−12.6	−12.2	−12.2	−12 8	11.93*	2.09
1dbj	−10.4	−15.9	−13.4	−9.5	0.32	0.21	1hpv	−12.6	−9.3	−9.3	−10 0	1.05	0.93
1dbk	−11.0	−15.2	−12.9	−9.0	0.57	0.40	1hpx	−12.7	−11.5	−11.5	−13.1	3.34	3.31
1dbm	−12.9	−15.3	−13.1	−9.2	1.95	1.97	1hri	−5.9	−8.9	−8.9	−7.3	10.09	2.26
1dd6		−13.5	−13.5	−9.1	1.36	8.27	1hsg	−12.8	−12.5	−12.5	−13.3	0.41	0.35
1dds	−11.3	−10.8	−10.8	−7.4	1.75	2.38	1hsl	−9.8	−8.4	−6.4	−9.4	1.31	1.31
1dhf	−10.1	−8.7	−8.7	−8.5	6.31	5.62	1hte	−8.6	−9.9	−9.9	−9.6	7.32*	7.36
1did	−4.8	−3.8	−3.8	−6.6	3.27	4.15	1htf	−9.3	−10.9	−10.9	−9.4	2.19	2.74
1die	−2.9	−5.8	−5.8	−6.9	0.34	0.77	1hti	−7.0	−4.5	−4.5	−5.7	4.40	1.60
1dih	−7.8	−8.2	−8.2	−14.3	2.62	1.78	1hvr	−13.0	−13.7	−13.7	−7.8	1.60	1.75
1dm2		−13.9	−13.9	−10.3	0.66	0.69	1hyt		−8.0	−8.0	−10.7	2.65	0.43
1dog	−5.5	−6.5	−6.5	−9.5	3.77	3.77	1icn		−10.3	−7.8	−1.7	9.06	2.04
1dr1	−7.4	−7.6	−7.6	−7.0	0.37	1.46	1ida	−11.9	−13.1	−13.1	−12.2	1.95	2.12
1dwb	−4.0	−6.0	−4.0	−7.7	0.29	0.32	1igj	[b]	−9.2	−9.2	−7.0	0.72	0.46
1dwc	−10.3	−9.3	−9.3	−8.8	2.06	0.89	1imb	−5.7	−6.6	−6.6	−10.0	1.84	1.64
1dwd	−11.4	−13.2	−11.2	−11.2	0.47	1.43	1ivb		−5.2	−5.2	−7.0	3.27	0.47
1ivc		−3.4	−3.4	−5.0	2.05	1.89	1wap		−8.1	−8.1	−10.2	0.23	0.19
1ivd	−4 3	−6.3	−6.3	−6.2	0.73	0.73	1xid		−5.7	−5.7	−6.6	4.02	4.32
1ive		−3.6	−3.6	−5.6	5.11	5.17	1xie		−4.8	−4.8	−6.8	2.60	3.91
1ivf		−7.7	−7.7	−6.8	0.61	0.59	2ack		−9.3	−9.3	−7.0	1.07	0.88
1lah	−10.3	−7.3	−7.3	−9.6	0.52	0.19	2ada		−9.5	−9.5	−9.0	0.59	0.46
1lcp	−9.1	−7.7	−7.7	−8.8	1.82	1.06	2cgr	−9.9	−6.5	−6.5	−10.8	0.56	0.52
1ldm	−7.4	−6.3	−6.3	−7.3	1.34	1.35	2cht	−7.5	−7.7	−7.7	−11.4	0.48	0.51

Table 5. Continued

PDB	ΔG_{exp}	GlideScore (kcal/mol)			RMS (Å)		PDB	ΔG_{exp}	GlideScore (kcal/mol)			RMS (Å)	
		XP	XP-corr	SP	XP	SP			XP	XP-corr	SP	XP	SP
1lic		−6.1	−6.1	−5.3	3.96	4.99	2cmd	−6.2	−8.2	−8.2	−10.3	0.65	0.34
1lmo		−7.6	−7.6	−6.8	8.40	0 87	2cpp	−8.3	−8.2	−8.2	−6.8	0.15	3.04
1lna		−6.6	−6.6	−6.8	1.50	0.90	2ctc	−5.3	−7.4	−7.4	−10.1	1.43	1.58
1lst		−6.1	−4.1	−7.7	0.75	0.27	2dbl	−11.8	−12.2	−11.0	−8.8	2.40	0.81
1mbi	−2.6	−3.9	−3.9	−4.1	1.92	1 65	2gbp	−10.1	−9.0	−9.0	−12.5	0.61	0.14
1mcr	−4.3	−8.3	−8.3	−7.5	5.82	4 33	2ifb	−7.4	−8.7	−6.2	−2.4	2.27	1.77
1mdr	−5.4	−8.2	−8.2	−9.6	1.95	0 54	2lgs		−3.4	−3.4	−9.5	0.88	0.53
1mfe	−7.2	−8.0	−8.0	−7.4	6.09	1.78	2mcp	−7.1	−5.6	−5.6	−5.6	1.54	1.25
1mld		−6.6	−6.6	−10.6	0.25	0.22	2phh	−6.4	−7.9	−7.9	−8.7	0.47	0.41
1mmq	−10.3	−13.8	−13.8	−11.0	0.67	0 33	2pk4	−5.9	−8.8	−6.8	−5.8	0.65	0.85
1mnc	−12.3	−11.9	−11.9	−11.1	0.33	0.73	2plv		−14.3	−11.8	−7.3	1.78	1.90
1mrg		−7.8	−7.8	−7.8	0.15	0.12	2r04	−8.0	−10.6	−8.1	−8.9	1.44	0.75
1mrk	−6.2	−11.3	−11.3	−9.4	1.21	1.17	2r07		−10.9	−8.4	−8.7	0.92	0.67
1mup		−9.3	−7.6	−6.2	4.50	4 05	2sim	−4.7	−7.1	−7.1	−10.5	0.82	0.94
1nco	−10.6	−10.4	−10.4	−12.1	0.60	0 33	2tmn	−8.0	−10.1	−10.1	−10.4	0.65	0.50
1nis	−4.1	−4.0	−4.0	−7.5	0.26	0.45	2tpi	−5.9	−8.6	−6.6	−9.5	0.26	1.15
1nnb	−7.2	−5.5	−5.5	−8.7	1.39	0.25	2upj	−14.2	−11.1	−11.1	−10.7	3.24	2.69
1nsc	−4.1	−5.7	−5.7	−10.5	0.66	0.24	2xis	−7.9	−4.9	−4.9	−7.3	2.05	2.40
1nsd	−7.2	−6.3	−6.3	−9.0	0.74	0.22	2yhx		−4.3	−4.3	−5.6	1.91	2.19
1odw		−11.5	−11.5	−5.7	3.91	2 59	3cla	−6.7	−6.4	−6.4	−5.4	5.09	6.06
1okl	−8.2	−7.2	−7.2	−5.8	0.38	3.14	3dfr	−14.0	−15.3	−13.3	−11.3	0.51	0.70
1pbd		−11.4	−11.4	−9.9	0.32	0.26	3hvt		−11.9	−11.4	−9.0	0.72	0.79
1pgp	−7.8	−6.3	−6.3	−8.8	2.23	1 83	3mth		−5.4	−5.4	−5.9	1.23	5.62
1pha		−12.2	−12.2	−8.8	1.04	0 52	3ptb	−6.1	−6.5	−6.5	−8.8	0.23	0.16
1phd		−7.4	−7.4	−6.5	1.13	0 30	3tpi	−5.9	−9.8	−7.8	−9.0	0.47	0.51
1phf	−6.0	−8.2	−8.1	−6.3	1.46	1.11	4aah		−7.6	−7.6	−11.0	0.25	0.24
1phg	−11.8	−10.9	−10.9	−8.1	4.29	1.20	4cts		−4.6	−4.6	−8.8	0.25	0.19
1poc	−10.7	−8.4	−8.4	−10.1	1.44	1 52	4dfr	−11.8	−10.1	−10.1	−9.3	0.92	5.17
1ppc	−8.4	−11.9	−9.9	−9.8	1.40	6 31	4fab	−11.0	−13.7	−12.9	−9.1	4.44	0.82
1pph	−8.1	−10.5	−8.5	−10.4	0.70	0 58	4fbp		−9.7	−9.7	−12.1	2.03	0.55
1ppi		−16.7	−14.7	−8.5	1.01	2 80	4fxn		−17.3	−17.3	−13.2	0.50	0.49
1ppk	−10.4	−9.9	−9.9	−10.1	0.73	0.27	4hmg	−3.5	−6.6	−6.6	−6.7	0.54	0.67
1ppl	−11.7	−10.3	−10.3	−11.6	0.70	2 55	4phv	−12.5	−14.5	−14.5	−11.8	0.55	4.22
1ppm	−7.9	−12.5	−12.5	−11.7	0.99	0 62	4tim	−2.9	−5.3	−5.3	−10.2	1.32	1.32
1pro	−15.4	−15.6	−15.6	−12.7	1.50	1 51	4tln	−5.1	−7.1	−7.1	−6.6	1.43	2.67
1pso	−14.1	−10.9	−10.9	−9.6	5.15	6.12	4tmn	−13.9	−11.9	−11.9	−12.1	1.29	0.73
1sbg	−10.6	−11.1	−11.1	−10.9	0.88	0.40	4tpi	−4.0	−6.6	−6.6	−8.5	0.88	0.56
1slt		−5.8	−5.8	−6.2	1.03	0 57	4ts1	−6.7	−8.6	−8.6	−8.8	0.89	0.85
1snc	−9.1	−9.6	−9.6	−9.8	2.06	1.12	5abp	−9.1	−7.6	−7.6	−8.5	0.11	0.20
1sre	−5.3	−8.9	−8.5	−10.0	0.30	0 36	5cpp	−8.0	−8.2	−8.2	−6.8	0.11	2.65
1srj		−13.7	−12.9	−10.0	0.47	0.49	5cts		−3.2	−3.2	−8.1	0.27	0.27
1stp	−18.3	−18.2	−18.2	−10.7	0.60	0 58	5p2p		−9.0	−9.0	−5.4	4.95	6.18
1tdb		−7.9	−7.9	−7.6	7.34	7 50	5tim	−3.1	−3.2	−3.2	−7.3	1.32	0.69
1thy		−7.0	−7.0	−7.2	1.98	4.21	5tln	−8.7	−12.4	−12.4	−9.9	2.37	1.01
1tka		−7.7	−7.7	−11.5	2.28	2.28	5tmn	−11.0	−12.0	−12.0	−12.1	2.87	2.50
1tlp	−10.3	−10.0	−10.0	−10.4	7.70	7 33	6abp	−7.7	−7.7	−7.7	−8.8	0.33	0.36
1tmn	−10.0	−10.2	−10.2	−10.4	3.65	1.90	6cpa	−15.7	−10.8	−10.8	−10.8	3.93	4.29
1tng	−4.0	−3.9	−3.9	−8.3	0.26	0.21	6rnt	−3.2	−3.7	−3.7	−8.2	0.63	0.63
1tnh	−4.6	−3.9	−3.9	−8.7	0.39	0.28	6tim	−4.4	−5.9	−5.9	−8.3	0.59	0.42
1tni	−5.5	−3.7	−3.7	−6.7	2.12	2 03	6tmn	−6.9	−10.4	−10.4	−11.6	2.53	2.57
1tnj	−2.7	−4.1	−4.1	−7.8	0.43	0 36	7abp	−8.6	−8.0	−8.0	−9.4	0.15	0.17
1tnk	−2.0	−3.8	−3.8	−6.9	1.17	0.98	7cpa	−19.0	−13.1	−13.1	−10.5	3.91	2.87
1tnl	−2.6	−4.0	−4.0	−7.4	0.54	0.24	7cpp	−5.2	−6.2	−6.2	−5.7	1.99	3.23
1tph	−3.1	−5.3	−5.3	−7.5	0.22	0.23	7tim	−7.4	−5.4	−5.4	−7.7	0.20	0.19
1tpp		−7.1	−5.1	−7.9	0.43	1 07	8abp	−10.7	−7.6	−7.6	−8.4	0.10	0.21
1trk		−9.6	−9.6	−11.2	1.63	2.15	8atc	−10.3	−9.5	−9.5	−10.7	0.41	0.38
1tyl		−5.2	−5.2	−6.7	5.20	1 08	8gch		−7.9	−7.9	−9.6	0.32	0.30
1ukz		−11.3	−11.3	−13.6	0.55	0.41	9abp	−10.9	−7.2	−7.2	−10.4	0.23	0.13
1ulb	−6.0	−6.7	−6.7	−8.7	0.34	0 35	9hvp	−11.4	−11.9	−11.9	−12.9	1.44	1.47

[a] RMS values are computed relative to the native cocrystallized ligand. Scores are compared to available experimental binding energies, ΔG_{exp} (in kcal/mol). Note that the XP scores have been adjusted to cap the hydrophobic enclosure packing term at a value of −2.0 to bring this term closer to an absolute energy scale. Larger RMS values are indicated with a (*) in cases where a nearly chemically symmetric solution was found by XP docking. [b] The pdbbind dataset,[42] v2002, lists a binding affinity of −13.6 kcal/mol, but this is for the entire antigen−antibody complex, whereas the structure provided is for a fragment of the complex that appears to lack essential antigen−antibody interactions.

affinities. This idea could form the basis for an approach enabling scores achieved in different protein conformations for a given receptor to be related based on experimental calibration of reorganization energy. Performance of the methodology when there is substantial reorganization is addressed in a preliminary fashion by the enrichment studies in section 4, where a number of such ligand−receptor pairs are considered. In these cases, knowledge of the reorganization energy of the receptor is not necessary to rank order the binding of compounds to a particular form of the receptor.

To address less extreme yet still nontrivial reorganization effects present in our data set, we define "average" adjustable parameters to convert calculated empirical scores into predicted experimental binding affinities. In particular, we estimate the

Table 6. RMS and Average Absolute Deviations (Avg) in Predicted Binding Affinities for the XP 4.0 and SP 4.0 Scoring Functions[a]

comparison	number	XP 4.0 RMSD	XP 4.0 avg	SP 4.0 RMSD	SP 4.0 avg
all ligands with ΔG_{exp}	198	2.26	1.75	3.18	2.51
all well-docked ligands	136	1.73	1.34	2.80	2.23
all poorly docked ligands	62	3.02	2.49	3.70	3.11

[a] RMS and average absolute deviations are presented in kcal/mol. Well-docked ligands are defined as those having an RMSD to the cocrystallized pose of 2.5 Å or less, plus those identified in Table 5 as being docked appropriately but in a symmetry-related orientation.

Table 7. Training Set Used to Characterize XP Virtual Screening[a]

PDB code	description	no. actives	no. well-docked actives
1e66	acetylcholinesterase	20	20
1bji	neuraminidase	9	9
1fjs	factor Xa	13	8
1kv2	human p38 map kinase	10	10
1bl7	p38 map kinase	36	27
1rt1	HIV-RT	29	23
1cx2	cyclooxygenase-2	13	13
1aq1	human cyclin dep. kinase	10	6
1ett	thrombin	16	15
1hpx	HIV-1 protease	14	9
3ert	human estrogen receptor	10	8
1qpe	lck kinase	121	87
1m17	EGRF tyrosine kinase	117	106
1tmn	thermolysin	6	5
1kim	thymidine kinase	4	4

[a] All correctly docked ligands have experimental activities <10 μM except those for neuraminidase.

hydrophobic-enclosure reorganization energy by capping the maximum assignable enclosure energy at a value of 2.0 kcal/mol, as compared to the maximum scoring-function value of 4.5 kcal/mol. It also appears as though the special bidentate charged−charged reward when the ligand group is positively charged involves some degree of protein reorganization, presumably to enable the bidentate salt bridge to be made properly. Thus, this parameter, which normally is 2.0 kcal/mol, was set to 1.0 kcal/mol for comparison of predicted binding affinities to experiment. Other such average effects could be defined, but due to the limited experimental data considered in this paper, no further attempts were made to determine additional parameters.

Tables 6 and 7 present results for both 4.0 XP and 4.0 SP Glide for structural RMSDs (taken with respect to the refined structure of the native ligand generated by our standard protein-preparation procedure,[1,40] using heavy atoms only) and for binding affinities of the various complexes discussed herein. Table 6 further divides the comparison according to whether the XP complex is roughly correctly docked, that is, has a structural RMSD of 2.5 Å or less. Over all 198 complexes examined, the average RMSD in predicted binding affinity for the 4.0 XP scoring function is 2.3 kcal/mol and the average unsigned error is 1.8 kcal/mol. When only well-docked ligands are examined, the average RMSD for the 4.0 XP scoring function is 1.7 kcal/mol and the average unsigned error is 1.3 kcal/mol. This is a significant improvement over the performance of the 4.0 SP scoring function and is also an improvement over the performances of other empirical scoring functions in the literature of which we are aware. In only 11% (15 of 136) of well-docked ligands were errors greater than 3 kcal/mol, suggesting, at least when the appropriately fitting protein structure is presented to the ligand and the ligand is well-docked, the present scoring function has a respectable ability to distinguish weak (mM), moderate (μM), and strong (nM) binders. This capability is essential to the principal task in virtual screening, yet is only marginally present in prior scoring functions.

To improve beyond this level with regard to precision, we believe that receptor flexibility must be introduced and an additional level of detail with regard to the protein−ligand interactions must be incorporated. Robustness is another matter; new moieties and chemistries seem to emerge as additional receptors are added to the test suite. Thus, we cannot claim to have reached convergence in this regard with our current data sets. On the other hand, the amount of detailed medicinal-chemistry information incorporated into the current scoring function is a substantial advance as compared to alternative scoring functions in the literature.[8−19]

A final important point is that the accuracy cited above may quantitatively degrade in cross-docking calculations, even when the ligand is able to assume a qualitatively correct pose in the receptor, as of course would the accuracy of other docking/scoring methods, for similar reasons. The enrichment studies presented below address this question to some extent, but at present do not consider the relative rankings of different active compounds. We have carried out some preliminary investigations (data not shown) that suggest qualitatively reasonable results can be obtained in a limited number of test cases for ranking active compounds when cross docking is required, but these results are not yet robust across a wide range of receptors, and the quantitative precision with which this can be done is not yet clear.

4. Enrichment Studies

The goal of the present work is to optimize a scoring function that will properly assign binding affinities to active compounds if the compound is docked in a sufficiently native-like pose and that will minimize the number of inactive database ligands that score well. The question of what is sufficiently native-like is a heuristic one, but we would argue that precision in drawing the line is not critical. If a few additional compounds are included, or excluded, this will have minimal effect on the overall optimization process. Therefore, visual inspection has been employed to construct data sets of active compounds that "fit" into the particular versions of the receptors employed. Typically, 60−100% of the available active compounds per receptor fell into this category, so the current protocol is not simply cherry-picking of a small number of compounds. On the other hand, the fact that poorly fitting compounds are not included in the data set must be taken into account when comparing with other results reported in the literature.

We have divided our data set into a training set and a test set. The training set screens have been used in parametrizing the XP scoring function, while the test set screens have not. The training set contains 15 receptors and various numbers of ligands for each receptor, as enumerated in Table 7. A large and diverse training set is essential to address the range of chemical motifs identified in the XP Glide scoring function, as noted previously. Our present test set is relatively small and less diverse than the training set with regard to the number of new receptors considered. Therefore, the validation implied by the results must be considered preliminary. Six receptors are considered in the test set, four of which were not included in the training set: vascular endothelial growth factor receptor 2 (Vegfr2), peroxisome proliferators activated receptor γ (PPARγ), β-secretase (BACE), and blood coagulation factor VIIa (factor VIIa). For two training set receptors, CDK2 and thrombin, we

have located a significant number of additional ligands to include as part of the test set.

Training Set Results. Table 7 lists, for the training set, the receptors investigated, the number of known active ligands available with affinities better than 10 μM, and the number of ligands deemed to dock correctly into the chosen receptor. There is one exception; active neuraminidase ligands are relatively weak binders that do not have activities better than 10 μM. As suggested above, we assume that a random database will contain relatively few compounds with potency greater than this value. Given the process defined above, a key objective of the paper is to demonstrate the improvement that is obtained from employing the new terms defined above. Comparison of version 4.0 XP Glide is made with the following:

(1) Version 4.0 SP Glide, which has been optimized using a similar training set. The 4.0 SP Glide scoring function includes some XP terms, but with very small coefficients. In addition to terms such as those in ChemScore, from which SP Glide was originally derived, this scoring function also includes contributions from the Coulomb and vdW protein−ligand interaction energies and from Schrödinger's "active-site mapping" technology.

(2) A preliminary version of XP Glide (v2.7). This version of XP had a nonoptimal set of penalty terms, an initial version of the hydrophobic-enclosure term, a crude representation of the special hydrogen-bond reward term without the crucial coupling to the hydrophobic-enclosure term, and a sampling methodology significantly inferior to that in Glide XP 4.0. This comparison enables assessment of the impact of the improved sampling and scoring methodologies used in Glide XP 4.0.

Protein and Ligand Preparation. Because the XP Glide scoring function is based on enforcement of physical chemical principles to a much greater degree than is employed in many other scoring functions, appropriate protein and ligand preparation is particularly critical. In practical applications, it is often necessary to carry out such preparation without prior knowledge of the binding mode of the complex. However, for the present purposes, the objective is to optimize and evaluate the scoring function for correctly docked compounds. Therefore, we have endeavored to use all available information in preparing ligand and protein structures.

The most problematic aspect of protein and ligand preparation is the assignment of protonation states of ligand and protein in the protein active site (note that we neutralize ionizable residues distant from the active site to mimic the effects of dielectric screening by solvent and counterions). In the absence of structural data, correctly making such assignments can be a very challenging task. However, when structural data is available, the most likely protonation states of both protein and ligand can usually be deduced from the structure of the complex. In cases for which we do not have a PDB structure, the correct binding mode of the ligand can typically be inferred by analogy. With a binding mode and solution-phase pK_as of the ligand, the correct protonation states can then be assigned. It should be emphasized that the results shown here require accurate protonation-state assignment, and substantial degradation can result from incorrect assignments in unfavorable cases.

A second aspect of protein preparation is relaxation of the receptor structure so that it at least accommodates the native ligand. We employ the standard Schrödinger protein preparation utility[1,40] for this purpose. A related issue is the use of van der Waals scaling of nonpolar ligand and protein atoms to take minor induced-fit effects into account in an approximate fashion. Various scalings have been examined though, with the exception of the human estrogen receptor (3ert), which used a scaling of 0.8 on the ligand and 0.9 on the protein, the "standard" scaling of 0.8 on the ligand and no protein scaling has been applied. Only active ligands that succeed in 4.0 XP docking with the chosen scaling parameters have been retained in our enrichment studies.

Comparison Database. Our methods for generating comparison databases are outlined in ref 1. Molecules are selected from a purchasable-compound library of about one million compounds that have been filtered for predicted pharmacokinetic properties using the QikProp program.[41] Selection protocols are then applied to ensure a distribution of rings, acceptors, donors, molecular weight, and so on in line with averages determined for drug-like molecules.

Computed Binding Affinities of Known Actives for a Wide Range of Targets. As stated above, our expectation is that only a small number of database ligands will be competitive with active compounds whose experimental binding affinities are better than 10 μM. An ensemble of such compounds can, therefore, be used to optimize the scoring function. Because of the wide range of novel terms that have been incorporated, it has been necessary to perform optimizations using a wide variety of receptors and active compounds. The data set used to date is far from complete in covering the range of potential binding motifs, but is significantly larger and more diverse than any previous data set used in the literature for this purpose.

The optimization process consists of adjusting parameters so that as few database ligands as possible achieve better scores than the tight-binding (better than 10 μM) known actives for the term or terms under consideration. In practice, it is not possible to achieve perfect rank ordering in this regard, for the reasons discussed previously. The average error in binding affinity prediction in the PDB data set is about 1.7 kcal/mol for properly docked ligands, with some outliers with errors of more than 3 kcal/mol. These errors may be somewhat larger when cross docking, rather than self-docking, is performed. On the basis of this analysis, some database ligands with experimental binding affinities in the 10−100 μM range are likely to be computed as having low micromolar or even nanomolar binding affinity, while at the same time some of the active compounds will be underpredicted by similar amounts. As a result, overpredicted database ligands can score ahead of underpredicted active compounds. The crucial goal, however, is not to achieve perfection, but rather to eliminate systematic errors that can lead to large numbers of false positives and little enrichment. Neglect of special neutral−neutral hydrogen-bond terms such as those discussed in section 2 is an example of a systematic error of this nature. One would predict that any method that neglects this term should exhibit poor enrichment factors for kinases such as CDK2, where such hydrogen bonds are a crucial component of the molecular-recognition motif. This is because we have found that large hydrophobic compounds with a few strategically placed polar groups can "fit" into the active site and form structures that, based on the usual pair-scoring terms, are highly competitive with the known actives.

Another, and perhaps the ultimate, measure of performance of the scoring function is whether high-ranking database ligands are in fact active. The best way to address this issue is via experimental testing of such database ligands. We are in the process of carrying out such tests.

Training Set Composition. Our training set, primarily focused on pharmaceutical targets of current interest, is presented in Table 7. This suite represents a wide variety of

Table 8. Average Attractive Components of 4.0 XP Score from Eq 3 of Correctly Docked Active Ligands in the Training Set[a]

screen	$\langle E_{phobic_pair} \rangle$	$\langle E_{hb_pair} \rangle$	$\langle E_{hb_nn_motif} \rangle$	$\langle E_{hb_cc_motif} \rangle$	$\langle E_{hyd_enclosure} \rangle$	$\langle E_{PI} \rangle$	$\langle E_{bind} \rangle$
acetylcholinesterase	−12.1	−0.3	0.0	−1.9	−4.4	−1.1	−20.8
neuraminidase	−3.8	−0.7	−0.4	−3.6	0.0	0.0	−8.6
factor Xa	−8.3	−0.8	−1.8	−1.2	−0.8	−0.8	−14.1
human p38 map kinase	−10.2	−1.6	−0.3	0.0	−2.6	0.0	−14.7
p38 map kinase	−7.4	−1.2	0.0	0.0	−0.9	0.0	−9.5
HIV-RT	−9.1	−1.0	−1.4	0.0	−4.2	0.0	−15.7
cyclooxygenase-2	−8.9	−0.3	0.0	0.0	−3.2	0.0	−12.5
human cyclin dep. kinase	−7.8	−2.1	−3.2	0.0	−1.5	0.0	−14.7
thrombin	−8.4	−1.7	0.0	−3.0	−0.4	0.0	−13.1
HIV-1 protease	−10.2	−2.6	−0.2	0.0	0.0	0.0	−13.1
human estrogen receptor	−10.6	−1.2	−0.5	0.0	−2.0	0.0	−14.3
lck kinase	−8.1	−1.4	−1.3	0.0	0.0	0.0	−10.8
EGRF tyrosine kinase	−6.7	−1.3	−1.5	0.0	0.0	0.0	−9.5
thermolysin	−8.7	−2.3	0.0	−0.3	−0.1	0.0	−11.4
thymidine kinase	−5.5	−2.3	−3.0	0.0	−1.4	0.0	−12.6

[a] E_{phobic_pair} is the pair lipophilic term (eq 1), E_{hb_pair} is the Chemscore-like pair hydrogen-bond term, $E_{hb_nn_motif}$ is the term for neutral−neutral hydrogen bonds in a hydrophobically enclosed environment, $E_{hb_cc_motif}$ is the term for special charged−charged hydrogen bonds, $E_{hyd_enclosure}$ is the hydrophobic enclosure reward, E_{PI} is the pi-stacking/pi-cation reward, and E_{bind} is the sum of all terms in eq 3 that account for favorable binding affinities.

different types of active sites and binding motifs. A rough classification of these is as follows:

(1) Small hydrophobic sites: HIV-RT (1rt1) and cyclooxygenase-2 (Cox-2, 1cx2). The HIV-RT NNRTI site is an allosteric pocket that opens to accommodate the ligand; the Cox-2 site does not display as dramatic a structural rearrangement, but is also highly hydrophobic. The dominant terms of the scoring function are the pair hydrophobic term and the hydrophobic-enclosure term. The hydrophobic enclosure is quite large in both cases for known active compounds. In HIV-RT, the active compounds typically make a single hydrogen bond to a backbone carbonyl that receives the special single neutral−neutral hydrogen-bond reward discussed in section 2. In Cox-2, there is also typically a single hydrogen bond, though it does not receive a special reward.

(2) Medium-sized sites making a single special hydrogen bond. This category includes EGFR tyrosine kinase and the 1bl7 form of p38 MAP kinase. This motif is one of the two typical kinase binding motifs, in which there is a special hydrogen-bonding site in the hinge region of the kinase. These systems allow only a single hydrogen bond in this site, typically involving a ring nitrogen atom, whereas other kinases form a correlated pair or triplet of hydrogen bonds, discussed in (3) below.

(3) Sites making correlated hydrogen bonds. This category includes thymidine kinase (TK), CDK2, and lck kinase (LCK). TK and LCK form a pair of correlated hydrogen bonds in the standard kinase hinge region, while CDK2 actives form either a pair or a triplet, and neutral actives binding to aldose reductase form a correlated triplet similar to that in the streptavidin/biotin pair mentioned above. TK is a small, relatively polar site, with the binding driven primarily by the special hydrogen bonding, whereas CDK2 and LCK are medium-sized and more hydrophobic, binding ligands in the 400−500 molecular weight range, although smaller ligands can bind as well.

(4) Large, buried predominantly hydrophobic sites. This category includes the human estrogen receptor and the 1kv2 conformation of p38 MAP kinase. The estrogen receptor binds large, flat steroid-type molecules and exhibits a medium-sized hydrophobic enclosure term, with the binding driven by this term and by the pair hydrophobic score. The 1kv2 active site is created by an allosteric rearrangement of the p38 activation loop. It has a large hydrophobic enclosure term, and many active compounds make one to two hydrogen bonds. The most active compound, ligand 1kv2, also makes a special hydrogen bond in the hinge region, but binding is primarily driven by hydrophobic terms. The p38 pocket in particular is quite large and, in the absence of the hydrophobic enclosure term, will display a significant number of false positives that bind alternative motifs in the cavity in database screening.

(5) Large, open sites with relatively shallow cavities in the active site pocket. This category includes thrombin (1ett), HIV protease (1hpx), and factor Xa (1fjs). Active ligands in this case are invariably large and fill multiple pockets in the active site. HIV protease actives make a substantial number of hydrogen bonds and derive their binding affinity from this and the hydrophobic pair term. Remarkably, there is no contribution from the hydrophobic enclosure. Thrombin ligands typically form a salt bridge with a buried Asp 189 carboxylate in one of the small available pockets and form other hydrogen bonds as well. Hydrophobic enclosure also makes no contribution for thrombin ligands. Factor Xa exhibits the same salt-bridge motif but also has an unusual hydrophobic location in which a ring moiety of the ligand is sandwiched between a number of aromatic rings in a location near the surface of the protein. This pocket provides some hydrophobic enclosure, although not to the same degree as a deeply buried pocket like that in 1kv2, 1cx2, or 1rt1 and can also accommodate pi-cation and pi stacking interactions. Discriminating false positives that display interactions with the protein surface rather than in binding pockets is important for these receptors, particularly factor Xa.

(6) Small hydrophilic sites in which strong electrostatic interactions are important. This category includes the neuramididase (1bji) and GluR2 receptors. In neuramididase, much of the binding affinity derives from salt bridges, and the various ligands serve as important test cases for the rules for charged−charged interactions laid out in section 2. In GluR2, an unusual set of charged ligands in which charge is distributed over a ring system, are buried in the active site. This system was used to calibrate the use of electrostatic terms to turn off buried-charge penalties and to assign an additional contribution to binding affinity in exceptional cases, as discussed in section 2.

(7) The acetylcholinesterase (1e66) receptor was included to incorporate the well-known pi-cation motif of the active compounds in the parameterization process.

Table 8 lists the average contribution of the hydrophobic enclosure and special hydrogen-bonding terms to the scores for known active compounds that bind to each of the above targets, as well as the average total score. Note that the total score cannot be directly translated into the predicted binding affinity in all cases because there are a number of targets where allosteric rearrangement of the binding pocket leads to substantial protein

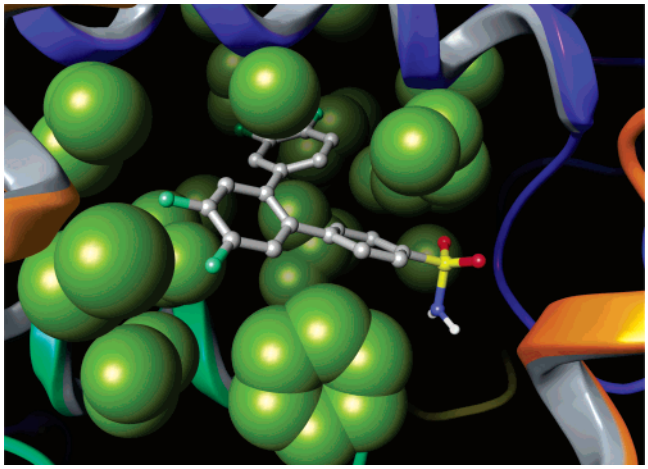

Figure 10. SCX-001 bound to cyclooxygenase-2. The three phenyl rings obtain a 3.7 kcal/mol hydrophobic enclosure packing reward for occupying the large hydrophobic pocket.

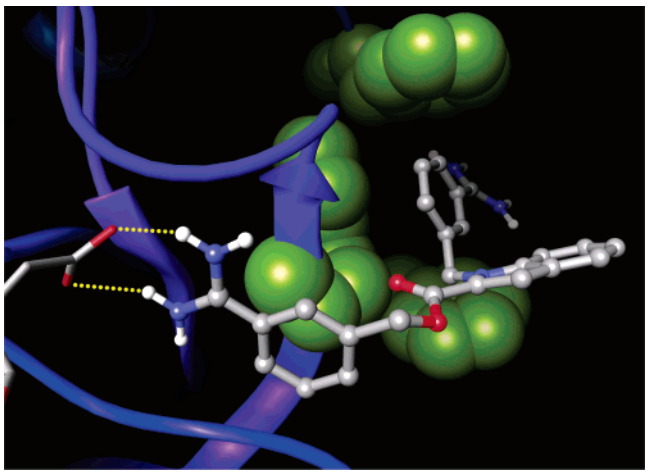

Figure 11. The 1lpk ligand bound to factor Xa. A special salt bridge pair with ASP 189, analogous to that in thrombin (Figure 8), is observed. A hydrophobic enclosure packing reward of −0.8 kcal/mol is achieved by the phenyl ring occupying the Tyr 99/Trp 215/Phe 174 pocket. Also, a pi-cation interaction is received by the charged end of the ligand.

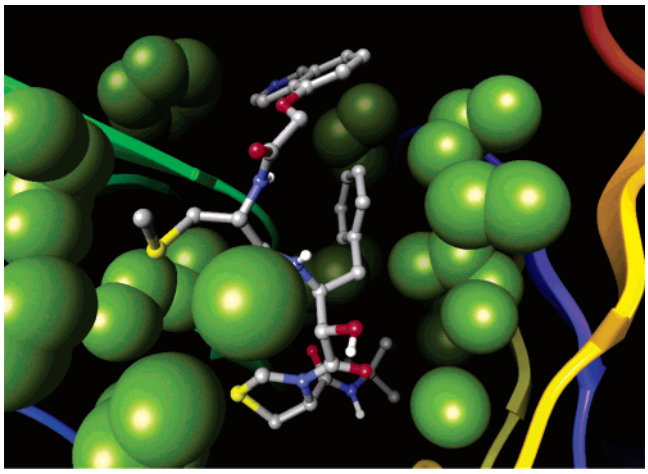

Figure 12. The 1hpx ligand bound to HIV-1 protease. Hydrophobic groups of the ligand are not hydrophobically enclosed and do not receive a hydrophobic enclosure packing reward. For example, the phenyl ring faces hydrophobic residues on only one face.

reorganization energy. Figures 2−13 provide illustrative examples of a "typical" active ligand binding to the various receptors illustrating features that contribute to the specialized scoring-function terms as appropriate.

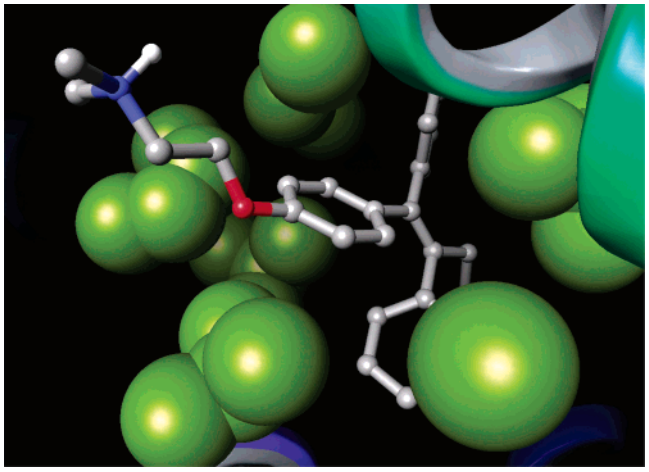

Figure 13. 4-Hydroxytamoxifen bound to the human estrogen receptor. Hydrophobic enclosure about the phenoxy group is illustrated by displaying lipophilic protein atoms as green spheres.

Table 9. Average Enrichments Defined as the Average Number of Outranking Decoy Ligands over Correctly Docked Actives in the Training Set[a]

	avg number of outranking decoys		
screen	v4.0 XP	v2.7 XP	v4.0 SP
acetylcholinesterase	111	580	344
neuramididase	25	411	37
factor Xa	1	196	187
human p38 map kinase	26	30	57
p38 map kinase	7	93	183
HIV-RT	11	26	83
cyclooxygenase-2	24	12	22
human cyclin dep. kinase	3	77	216
thrombin	2	57	70
HIV-1 protease	16	60	167
human estrogen receptor	14	2	23
lck kinase	13		157
EGRF tyrosine kinase	41	411	279
thermolysin	9	107	32
thymidine kinase	0	1	29

[a] Active ligands have at least 10 μM activity except those for neuraminidase, as described in Section 4.

Results of Training Set Enrichment Studies. Table 9 reports a measure of enrichment defined as the average number of database ligands outranking the active compounds in the database. Specifically, the number of database ligands with a GlideScore that is superior to each active is tabulated, these values are summed, and the result is then divided by the total number of active compounds in the data set. We believe that this metric is superior to standard definitions of enrichment, which punish active ligands when they are outranked by other active ligands; this is a particularly serious problem when the active test suite contains a large number of compounds. A "perfect" score based on this metric would thus be zero (no database ligands outranking any active compounds), and smaller numbers are better. These values are also presented for the older 2.7 XP and 4.0 SP Glide results. As noted previously, only active ligands that successfully docked in 4.0 XP Glide were considered. In a small number of cases, active ligands failed to dock with 4.0 SP or 2.7 XP. For a calculation of the number of outranking decoy ligands, such ligands were ranked lower than all successfully docked active and decoy ligands.

Table 10 is the corresponding table constructed using a more standard definition of enrichment that we have employed

Table 10. Standard Enrichment Factors for Recovering 40% of the Correctly Docked Active Ligands in the Training Set[a]

screen	enrichment factors		
	v4.0 XP	v2.7 XP	v4.0 SP
acetylcholinesterase	37	1	2
neuramididase	64	2	112
factor Xa	126	42	63
human p38 map kinase	81	58	51
p38 map kinase	35	18	11
HIV-RT	36	14	18
cyclooxygenase-2	35	56	78
human cyclin dep. kinase	168	168	8
thrombin	68	58	12
HIV-1 protease	90	32	112
human estrogen receptor	126	126	126
lck kinase	12		6
EGRF tyrosine kinase	6	1	2
thermolysin	50	134	201
thymidine kinase	251	251	17

[a] Active ligands have at least 10 μM activity, except those for neuraminidase, as described in section 4.

previously.[2] The results in Table 10 include the same sets of active ligands as in Table 9, that is, those whose activities are better than 10 μM and have been judged to fit more or less correctly into the specified conformation of the receptor. Table 11 presents results using all active compounds with binding affinities better than 10 μM, whether the binding mode is judged to be correct. Results are presented for recovering 40, 70, and 100% of the considered active ligands in each case. This type of analysis corresponds to the approach taken by us in ref 2, as well as to other work in the literature. While we believe that the analysis in Table 9 is the appropriate one to use in assessing the quality of a scoring function to be used in rigid-receptor docking, the results presented in Table 11 enable a direct connection to be made with alternative viewpoints. In what follows, our discussion is focused on the results in Table 9 for the reasons given above.

The XP 4.0 results are nearly uniformly comparable to or better than those of either SP 4.0 or XP 2.7 and, in many cases, are significantly better, as is manifested with particular clarity using the new definition of enrichment. There is a slight degradation for the estrogen receptor from XP 4.0 for cyclooxygenase-2 relative to both XP 2.7 and SP 4.0, but all of the results for these test cases are very good. The real question with regard to scoring-function effectiveness is the ability to prevent false positives from ranking ahead of active compounds. XP 4.0 displays an ability to reduce the average number of false positives ranking ahead of actives in many cases by an order of magnitude and in some cases by nearly 2 orders of magnitude, as compared to both 2.7 XP and 4.0 SP. This same effect is also reflected in the more common definition of enrichment factor (Table 10), but the improvement is quantitatively obscured by the definition of enrichment employed, particularly for the data sets containing larger numbers of actives. For example, in EFGR kinase the number of actives is greater than 10% of the random database, and standard enrichment measures that effectively penalize active compounds for having other active compounds ranking ahead of other actives can yield enrichment factors of at most 10.3 for Table 10 and 9.5 for Table 11.

The results shown in Table 9 are not perfect. However, until intrinsic RMS fluctuations in the scoring function can be reduced from the present average of 1.7 kcal/mol for well-docked ligands, the scoring function seems unlikely to systematically perform significantly better without overfitting. The number of high-scoring database ligands reflected in this table is consistent with the estimated experimental population of low micromolar hits in a 1000 molecule random database of drug-like molecules. The acetylcholinesterase receptor appears to manifest the largest systematic errors. This is likely due to our inability to optimize the pi-cation and pi-stacking scoring function terms with high precision because we lack sufficiently diverse examples manifesting these terms. There also remain some difficulties associated with smaller, highly hydrophobic sites, such as Cox-2 and in medium-sized sites with a single special hydrogen bond, such as EGFR. Overall though, the results are reasonably robust across the entire data set and clearly represent a major advance over the results obtained using 4.0 SP or 2.7 XP. Direct comparisons with other codes would require using the same sets of actives and database ligands. Based on anecdotal reports from various sources and from comparison with published data,[1] Glide SP has generally performed at least as well in enrichment studies as, if not better than, alternatives such as GOLD and FlexX. One would therefore expect 4.0 XP to outperform these methods by a margin similar to that seen in Table 9 for 4.0 SP.

Results of Test Set Enrichment Studies. A summary of key data for our test set, including the receptor crystal structures used, and the number of known and well-docked actives again restricted to ligands with experimental binding affinities better than 10 μM are presented in Table 12. The test set includes two kinases (CDK2, Vegfr2), three proteases (thrombin, BACE, factor VIIa), and one nuclear hormone receptor (PPARγ) and, hence, is reasonably diverse with regard to function; all of the receptors in the test set are drug targets of current or recent interest. There is less diversity with regard to active site size and hydrophobicity than in the training set. As discussed above, validation with a larger test set will be addressed in future publications. All test set calculations were performed with the released versions of Glide 4.0 XP and SP, with no parameter adjustment being made to improve results for any targets.

For two of the receptors (Vegfr2 and PPARγ) we utilize two different forms of the receptor structures. These are highly flexible active sites, and a significant fraction of ligands in both cases can be divided into groups that clearly fit better into one version of the receptor or the other. For PPARγ, for example, one class of ligands requires opening of an allosteric pocket (primarily via motion of a phenylalanine residue), while the second class is smaller and does not protrude into this pocket. Comparing scores of these two ligand classes using a single receptor structure does not make sense. If the pocket is closed, the larger ligands will not fit at all, whereas if the pocket is open, the larger ligands will unfairly score better, as the reorganization energy of the receptor required to engender the needed side chain motion will not have been included. There is no overlap between the ligands associated with the two receptor forms. This partitioning is meant as an introductory exploration of enrichment studies using multiple receptor conformations, a topic we intend to pursue more intensively in the future.

Enrichment metrics to recover well-docked active ligands based on number of outranking decoys and standard enrichment measures (as for the training set) are presented in Tables 13 and 14, respectively. For all known active ligands, standard enrichment measures are presented in Table 15. The same comparison database of 1000 decoy ligands employed in training set enrichment studies has been used. As expected, there is some quantitative degradation of the XP results from the training set, but overall the results are qualitatively comparable to the training set results using the outranking decoy metric (which we have argued is the most meaningful for our purposes), and the improvements as compared to SP Glide are, on average, significant. For PPARγ, both methods do reasonably well; this

Table 11. Standard Enrichment Factors for Recovering 40, 70, and 100% of Known Active Ligands in the Training Set, *Including* Misdocked Cases

	enrichment factors								
	v4.0 XP			v2.7 XP			v4.0 SP		
screen	40%	70%	100%	40%	70%	100%	40%	70%	100%
acetylcholinesterase	37	19	1	1	1	1	2	2	1
neuramididase	64	34	6	2	1	1	112	23	4
factor Xa	78	58	3	10	1	1	22	9	1
human p38 map kinase	81	59	4	58	11	9	51	21	3
p38 map kinase	27	14	1	13	3	1	3	2	2
HIV-RT	27	19	3	12	12	0	11	6	0
cyclooxygenase-2	35	29	7	56	29	0	78	58	9
human cyclin dep. kinase	81	20	2	51	4	2	5	3	2
thrombin	64	64	2	54	7	0	11	10	3
HIV-1 protease	72	32	14	27	16	2	72	60	1
human estrogen receptor	101	101	2	101	88	2	101	17	2
lck kinase	9	8	2				4	3	1
EGRF tyrosine kinase	5	6	1	1	1	1	2	2	1
thermolysin	42	52	23	112	37	2	168	34	7
thymidine kinase	251	251	201	251	188	201	17	24	17

Table 12. Test Set Used to Validate XP Virtual Screening[a]

PDB code	description	no. actives	No. well-docked actives
1m4h	BACE	77	34
1dan	factor VIIa	93	40
1fm6	PPARγ (closed form)	93[b]	32[c]
1fm9	PPARγ (open form)	93[b]	25[c]
1y6b	Vegfr2 (closed form)	111[b]	21[c]
1ywn	Vegfr2 (open form)	111[b]	26[c]
1aq1	human cyclin dep. kinase	253	143
1ett	thrombin	40	15

[a] All correctly docked ligands have experimental activities <10 μM. [b] When multiple forms of a receptor are utilized, the number of actives reported here is all known actives with affinities <10 μM. [c] Ligands are assessed as optimally fitting into either the open or closed form of the receptor. For example, in PPARγ, 61% of known active ligands (57 of the 93) were assessed as fitting into either the open or the closed form of the receptor.

Table 13. Average Enrichments Defined as the Average Number of Outranking Decoy Ligands over Correctly Docked Actives in the Test Set[a]

	avg no. of outranking decoys	
screen	v4.0 XP	v4.0 SP
BACE	35	342
factor VIIa	30	75
PPARγ (closed form)	45	48
PPARγ (open form)	44	76
Vegfr2 (closed form)	52	222
Vegfr2 (open form)	69	310
human cyclin dep. kinase	25	206
thrombin	9	52

[a] Active ligands have at least 10 μM activity.

is a case where SP scoring performs unexpectedly well, as opposed to suggesting a particular problem with the XP scoring function.

A number of caveats should be emphasized with regard to these results. The test set is small, and there are almost certainly cases where enrichment performance will not be as good as that indicated in Tables 13 and 14. Furthermore, high enrichment with few false positives can only be expected when the active ligands are properly docked. The fact that a significant fraction of actives are not well-docked, even when two receptor conformations are used, indicates that if a diverse set of active compounds is to be robustly separated from a random database without false positives or false negatives, significant work needs to be done to better treat receptor flexibility and ensure reliable docking accuracy. However, we believe that the attempt in this paper to separate scoring function accuracy, docking accuracy, and reorganization energy effects is an essential starting point if truly robust approaches are to be developed.

Table 14. Standard Enrichment Factors for Recovering 40% of the Correctly Docked Active Ligands in the Test Set[a]

	enrichment factors	
screen	v4.0 XP	v4.0 SP
BACE	30	25
factor VIIa	19	15
PPARγ (closed form)	9	19
PPARγ (open form)	14	40
Vegfr2 (closed form)	25	10
Vegfr2 (open form)	11	2
human cyclin dep. kinase	8	5
thrombin	64	25

[a] Active ligands have at least 10 μM activity.

Table 15. Standard Enrichment Factors for Recovering 40 and 70% of Known Active Ligands in the Test Set, *Including* Misdocked Cases

	enrichment factors			
	v4.0 XP		v4.0 SP	
screen	40%	70%	40%	70%
BACE	12	3	11	0
factor VIIa	10	6	7	5
PPARγ (closed form)	5	2	7	4
PPARγ (open form)	3	3	11	7
Vegfr2 (closed form)	2	1	1	1
Vegfr2 (open form)	3	2	1	1
human cyclin dep. kinase	4	3	2	2
thrombin	19	3	11	6

5. Conclusions

We have described a novel scoring function and an enhanced sampling algorithm, the combination of which constitutes the Glide 4.0 XP docking methodology. The methodology has been tested with a diverse set of ligands and receptors, and has produced large improvements in binding affinity prediction and database enrichment as compared to other scoring functions within Glide.

The potential for providing physical insight into the origins of enhanced binding affinity is, in our view, as important as quantitative improvement of enrichment factors. Visualization of XP Glide terms, as is presented in the Figures of the present paper, can be utilized by modelers and medicinal chemists in the design of new inhibitors. The success of design efforts along these lines in the context of lead optimization would provide

the most convincing evidence that the underlying model of molecular recognition proposed herein has substantial validity. Our hope is that the present paper will facilitate work along these lines by describing in considerable detail the theory that underlies the XP Glide implementation.

Acknowledgment. We thank Mark Murcko, Bob Pearlstein, and Barry Honig for reading preliminary versions of this manuscript and providing useful feedback. We also thank Mike Campbell for assistance in generating graphics.

Supporting Information Available: Detailed descriptions of the algorithms used in hydrophobic enclosure scoring and in scaling special neutral−neutral hydrogen-bond motif rewards. References to experimental binding affinities for all test and training set ligands are also included. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(2) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(3) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins* **2004**, *56*, 235−249.

(4) Krovat, E. M.; Steindl, T.; Langer, T. Recent Advances in Docking and Scoring. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 93−102.

(5) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem*. **2004**, *47*, 558−565.

(6) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534−553.

(7) Farid, R.; Day, T.; Friesner, R. A.; Pearlstein, R. A. New Insights about HERG Blockade Obtained from Protein Modeling, Potential Energy Mapping, and Docking Studies. *Bioorg. Med. Chem.* **2006**, *14*, 3160−3173.

(8) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolinine, 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.-Aid. Mol. Des.* **1997**, *11*, 425−445.

(9) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, L. J.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease−Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995**, *2*, 317−324.

(10) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olsen, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(11) 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.

(12) Bohm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein Ligand Complex of Known 3-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(13) Bohm, H. J. Prediction of Binding Constants of Protein Ligands: A Fast Method for the Prioritization of Hits Obtained from De Novo Design or 3D Database Search Programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.

(14) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(15) Jones, G.; Willet, 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.

(16) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

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

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

(19) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(20) Wallqvist, A.; Berne, B. J. Computer-Simulation of Hydrophobic Hydration Forces on Stacked Planes at Short-Range. *J. Phys. Chem.* **1995**, *99*, 2893−2899.

(21) Lum, K.; Chandler, D.; Weeks, J. D. Hydrophobicity at Small and Large Length Scales. *J. Phys. Chem. B* **1999**, *103*, 4570−4577.

(22) Nicholls, A.; Sharp, K. A.; Honig, B. Protein Folding and Association−Insights from the Interfacial and Thermodynamic Properties of Hydrocarbons. *Proteins* **1991**, *11*, 281−296.

(23) Zhou, R. H.; Huang, X. H.; Margulis, C. J.; Berne, B. J. Hydrophobic collapse in multidomain protein folding. *Science* **2004**, *305*, 1605−1609. Huang, X. H.; Zhou, R. H.; Berne, B. J. Drying and hydrophobic collapse of paraffin plates. *J. Phys. Chem. B* **2005**, *109*, 3546−3552.

(24) Cheng, Y. K.; Rossky, P. J. Surface Topography Dependence of Biomolecular Hydrophobic Hydration. *Nature* **1998**, *392*, 696−699.

(25) Hummer, G.; Rasaiah, J. C.; Noworyta, J. P. Water Conduction Through the Hydrophobic Channel of a Carbon Nanotube. *Nature* **2001**, *414*, 188−190.

(26) Liu, P.; Huang, X. H.; Zhou, R. H.; Berne, B. J. Observation of a dewetting transition in the collapse of the melittin tetramer. *Nature* **2005**, *437*, 159−162.

(27) Wallqvist, A.; Berne, B. J. Molecular-Dynamics Study of the Dependence of Water Solvation Free-Energy on Solute Curvature and Surface-Area. *J. Phys. Chem.* **1995**, *99*, 2885−2892.

(28) Lee, S. H.; Rossky, P. J. A Comparison of the Structure and Dynamics of Liquid Water at Hydrophobic and Hydrophilic Surfaces−A Molecular-Dynamics Simulation Study. *J. Chem. Phys.* **1994**, *100*, 3334−3345. Lee, C. Y.; McCammon, J. A.; Rossky, P. J. The Structure of Liquid Water at an Extended Hydrophobic Surface. *J. Chem. Phys.* **1984**, *80*, 4448−4455.

(29) Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. Reconciling the Magnitude of the Microscopic and Macroscopic Hydrophobic Effects. *Science* **1991**, *252*, 106−109.

(30) Regan, J.; Breitfelder, S.; Cirillo, P.; Gilmore, T.; Graham, A. G.; Hickey, E.; Klaus, B.; Madwed, J.; Moriak, M.; Moss, N.; Pargellis, C.; Pay, S.; Proto, A.; Swinamer, A.; Tong, L.; Torcellini, C. Pyrazole Urea-Based Inhibitors of p38 MAP Kinase: From Lead Compound to Clinical Candidate. *J. Med. Chem.* **2002**, *45*, 2994−3008.

(31) Hendsch, Z. S.; Tidor, B. Do Salt Bridges Stabilize Proteins−A Continuum Electrostatic Analysis. *Protein Sci.* **1994**, *3*, 211−226.

(32) Waldburger, C. D.; Schildbach, J. F.; Sauer, R. T. Are Buried Salt Bridges Important for Protein Stability and Conformational Specificity. *Nat. Struct. Biol.* **1995**, *2*, 122−128.

(33) Marqusee, S.; Sauer, R. T. Contributions of a Hydrogen-Bond Salt Bridge Network to the Stability of Secondary and Tertiary Structure in Lambda-Repressor. *Protein Sci.* **1994**, *3*, 2217−2225.

(34) Luo, R.; David, L.; Hung, H.; Devaney, J.; Gilson, M. K. Strength of Solvent-Exposed Salt-Bridges. Strength of Solvent-Exposed Salt-Bridges. *J. Phys. Chem. B* **1999**, *103*, 727−736.

(35) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: A Novel Scoring Function for Predicting Binding Affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.

(36) Wei, B. Q. 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.

(37) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.

(38) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20*, 720−729.

(39) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. A. Importance of Accurate Charges in Molecular Docking: Quantum Mechanical/Molecular Mechanical (QM/MM) Approach. *J. Comput. Chem.* **2005**, *26*, 915−931.

(40) *Schrödinger User Manuals*, Glide v3.0; Schrödinger, L.L.C.: New York, NY, 1994.

(41) *QikProp v2.3*; Schrödinger, L.L.C.: New York, NY, 2005.

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

JM051256O