Detecting the Native Ligand Orientation by Interfacial Rigidity: SiteInterlock

Short title: Identifying Ligand Binding Poses by Rigidity

Manuscript with Supporting Information
Proteins: Structure, Function, and Bioinformatics, 84, in press.

Sebastian Raschka[1], Joseph Bemister-Buffington[1], and Leslie A. Kuhn[1,2,3]

[1]Department of Biochemistry & Molecular Biology, Michigan State University, Biochemistry Building, 603 Wilson Road, East Lansing, MI 48824

[2]Department of Computer Science & Engineering, Michigan State University

[3]Corresponding author: Leslie A. Kuhn

E-mail: kuhnl@msu.edu

Website: http://www.kuhnlab.bmb.msu.edu

Telephone: 517-353-8745

Fax: 517-353-9334

Keywords: binding determinants, binding mode prediction, coupled interactions, docking, flexibility, interface features, ligand stabilization of proteins, ProFlex, protein stability, scoring functions

Abbreviations: DSF, differential scanning fluorimetry; HIV, human immunodeficiency virus; MD, molecular dynamics; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; pK_D, $-\log_{10}$ of the equilibrium dissociation constant; RMSD, root mean square deviation

ABSTRACT

Understanding the physical attributes of protein-ligand interfaces, the source of most biological activity, is a fundamental issue in biophysics. Knowing the characteristic features of interfaces also enables the design of molecules with potent and selective interactions. Prediction of native protein-ligand interactions has traditionally focused on the development of physics-based potential energy functions, empirical scoring functions that are fit to binding data, and knowledge-based potentials that assess the likelihood of pairwise interactions. Here we explore a new approach, testing the hypothesis that protein-ligand binding results in computationally detectable rigidification of the protein-ligand interface. Our SiteInterlock approach uses rigidity theory to efficiently measure the relative interfacial rigidity of a series of small-molecule ligand orientations and conformations for a number of protein complexes. In the majority of cases, SiteInterlock detects a near-native binding mode as being the most rigid, with particularly robust performance relative to other methods when the ligand-free conformation of the protein is provided. The interfacial rigidification of both the protein and ligand prove to be important characteristics of the native binding mode. This measure of rigidity is also sensitive to the spatial coupling of interactions and bond-rotational degrees of freedom in the interface. While the predictive performance of SiteInterlock is competitive with the best of the five other scoring functions tested, its measure of rigidity encompasses cooperative rather than just additive binding interactions, providing novel information for detecting native-like complexes. SiteInterlock shows special strength in enhancing the prediction of native complexes by ruling out inaccurate poses.

INTRODUCTION

Stabilization of protein complexes by ligand binding

Experimental methods that probe the relationship between protein order, stability, and ligand binding have proven increasingly useful in structure determination and ligand screening. For instance, thermal shift assays such as differential scanning fluorimetry (DSF) and calorimetry measure the temperature at which a protein gains or loses structural integrity. Taking advantage of the tendency for ligand binding to shift the unfolding equilibrium towards the native state and for ligand binding to increase the melting temperature[1,2], DSF has become important for high-throughput drug discovery[3] and the discovery of ligands that stabilize proteins for structure determination[4,5]. Nuclear magnetic resonance (NMR) studies have also shown that many intrinsically disordered protein domains adopt stable structures upon binding to their targets[6]. Theoretical models of protein folding indicate that proteins with greater thermal stability tend to have fewer major internal motions and less flexibility overall at constant temperature[7]. These principles have been used to design proteins with high-affinity, pre-specified ligand binding, by focusing on the principles of "energetically favorable hydrogen-bonding and van der Waals interaction with the ligand..., high overall shape complementarity to the ligand, and ... structural pre-organization in the unbound protein state, which minimizes entropy loss upon ligand binding"[8].

However, experiments have revealed that designing ligands by maximizing the number of non-covalent interactions in the binding interface does not always improve the affinity between a protein and its binding partner [9-10]. Theory tells us that the net enthalpic gain

of newly designed interactions may be overcome by the entropic cost of losing bond-rotational degrees of freedom due to the additional non-covalent constraints. Similarly, assuming the additivity and dominance of enthalpic contributions can be oversimplifications[11]. However, neither of these considerations rules out the possibility of there typically being localized rigidification at the site of interaction between the protein and ligand, which may be accompanied by compensatory flexibility elsewhere in the molecules. In this work, we test whether such a measure of interfacial rigidity, involving protein atoms close to the ligand, contains sufficient information to predict their binding mode.

Computational probes of protein rigidity and flexibility

Two computational approaches for identifying rigid (stable) and flexible regions in proteins based on their intramolecular contacts or bond networks, rather than force field calculations by methods such as molecular dynamics (MD), have become widely used in recent years. The aim of these methods is to simplify the analysis of coupled motions and access larger-scale, biologically relevant conformational changes. The pioneering atomistic elastic network models for proteins[12] evolved into faster, residue-based Gaussian network models[13,14]. These network models use normal mode analysis to identify the principal directions and amplitudes of motion at different frequencies within an oscillating spring system representing the protein, in which the spring force constants reflect the strength of non-covalent forces between atoms or residues. In contrast, ProFlex (initially named FIRST) evaluates protein flexibility by counting the bond-rotational degrees of freedom on a 3-dimensional graph of the covalent and non-covalent bond network[15]. This approach evolved from structural rigidity theory developed in the

1800's by James Clerk Maxwell for analyzing the distribution of flexible, rigid, and strained regions in bridges and other trusswork, based on the number and configuration of the struts[16]. Instead of struts, bonds are used to represent the covalent and non-covalent interactions in proteins, including hydrophobic contacts, strong hydrogen bonds, and salt bridges. The 3D constraint counting search on the graph representing the protein covalent and non-covalent bond network results in a decomposition of the protein structure into spatial subsets: regions that are overconstrained by bonds and are rigid; cooperatively flexible regions that are formed by a coupled network of rotatable and non-rotatable bonds; and entirely flexible regions, such as side chains and main chain termini that do not interact with other groups[15,17]. The temperature dependence of flexibility and the spatial hierarchy of flexible regions within a protein can also be evaluated with ProFlex[18,19]. The use of ProFlex by a number of research groups has shown its ability to reproduce main-chain crystallographic temperature factors and flexible regions identified by NMR for a number of proteins[15,18,20], as well as subtle long-range changes in flexibility, including accurately predicting how flexibility redistributes upon ligand binding in Ras/Raf and HIV protease complexes[15,21]. Interestingly, despite taking less than a second of computing time per protein on a standard desktop computer, ProFlex results substantially agree with the flexible regions identified by elastic network models[19] and computationally more expensive MD simulations[21]. For HIV protease (Fig. 1), ProFlex reproduces NMR, crystallography, and MD results[15,22–24] indicating that the flaps above the binding pocket rigidify upon ligand binding and that chemical asymmetry within a ligand induces asymmetry in the flexibility of the monomers forming the active site.

Computational detection of protein-ligand interfacial rigidification

Given the experimental support for a protein-stabilizing effect of ligand binding in many cases, and the availability of ProFlex, a tool uniquely suited to define the rigid and flexible regions in a protein-ligand complex, we tested the hypothesis that native ligand binding results in rigidification of the protein-ligand interface through cooperative interactions. Interfacial rigidification has not previously been evaluated theoretically or computationally as a predictor of protein-ligand binding. In the majority of cases, the ProFlex-based SiteInterlock rigidity measure can predict the native complex given a series of sampled conformations and orientations of the ligand. SiteInterlock also provides new information to combine with existing protein-ligand scoring potentials, given that it is not highly correlated with scoring functions that have been trained to predict the interaction energy. Rather than being trained with a particular set of proteins to predict a response variable such as $\Delta G_{binding}$, SiteInterlock directly evaluates the change in rigidity of the interfacial bond network upon complex formation.

MATERIALS AND METHODS

The SiteInterlock analysis can be summarized in three main steps: (1) sampling all low-energy conformations of each ligand by using a tool such as OMEGA v. 2.3.2 (OpenEye Scientific Software, Inc., Santa Fe, NM; http://www.eyesopen.com)[25,26] if this is not already done by the ligand docking/orientational sampling tool, (2) sampling and saving a variety of sterically allowed orientations of all ligand conformations in the protein site of interest by using a docking tool such as SLIDE[27] (http://www.kuhnlab.bmb.msu.edu/software/slide/index.html) *without* using the docking scoring function to filter the orientations, and (3) analyzing the structural rigidity of the protein-ligand binding interface for all docked ligand orientations with SiteInterlock, which employs ProFlex rigidity analysis[28].

Protein-ligand complexes analyzed

To test the efficacy of SiteInterlock in predicting native-like complexes, a set of 30 diverse protein complexes was prepared, including 25 enzymes and five receptors (Table I and Fig. 2). All are determined at crystallographic resolution of 2.5 Å or better, and are not listed as problematic structures in a quality analysis of protein-ligand fitting and refinement[29]. Water molecules, hydrogen atoms, ligands, and non-protein atoms were removed from the Protein Data Bank (PDB)[30] files prior to docking; however, metal ions were retained if they were part of the ligand binding pocket. The 30 protein targets can be distinguished further as holo or apo structures. Eleven apo structures, in which a ligand-free structure of the protein was used for docking, were included to represent the

additional challenge of not knowing the precise conformation of the protein bound to the ligand. For these 11 apo cases, the corresponding ligand-bound structures were available as separate PDB entries and used to provide an initial conformation of the ligand and also to validate the accuracy of the SiteInterlock-selected complex. For the 19 holo and 11 apo structures, the ligand of interest was extracted from the protein binding site and then conformationally sampled to reflect the realistic situation of not knowing the bioactive conformation of the ligand. The exact crystallographic ligand conformation was not included in docking for any of the 30 cases. This results in the "needle in the haystack problem" of having a large number of imperfect complexes (due to many orientations and many conformations of the ligand, plus protein conformational inaccuracies), challenging the scoring method to identify the most native-like.

Sampling complexes by molecular docking

After the ligands were extracted from their Protein Data Bank complexes (Table I), hydrogen atoms and partial charges were assigned via partial semi-empirical AM1 geometry optimization with bond charge correction[66] by using *molcharge* (v. 1.3.1) from the QUACPAC package (version 1.6.3.1, OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com). Up to 50,000 conformations were sampled for each ligand with OpenEye OMEGA version 2.3.2[25,26], and the most energetically favorable conformations (up to 200 conformers) were kept for docking. SLIDE, which docks ligands by exhaustive three-point pharmacophore matching between each conformer and the binding site and performs minimal protein side-chain and ligand single-bond rotations to allow van der Waals collision-free docking, was then used to sample a range of dockings for each complex. SLIDE version 3.4 was modified to output all sterically

allowed orientations of each ligand, given the OMEGA conformers as input. To assess the goodness of a docking, the root-mean-square deviation (RMSD) between non-hydrogen atom positions was calculated between each docking and the crystallographic ligand pose. Starting with this large set of ligand dockings labeled by RMSD, a series of dockings was selected to span the RMSD range between 0 and 3 Å (relative to the crystallographic position), representing a range of sterically feasible, near-native but otherwise un-scored dockings. For each complex, this series included the best-sampled docking (closest to 0 Å RMSD), the docking closest to 3 Å RMSD, and an average of 8 additional dockings distributed semi-uniformly in the 0-3 Å RMSD range. Ligand dockings in the range of 3-6 Å RMSD were also sampled, and several dockings with different RMSD values in that range were also kept for each complex as examples of poor dockings. For seven of the complexes, dockings in the 6-10 Å RMSD range were also observed and included. Ideally, evenly separated dockings would be selected over a specified RMSD interval for all targets (e.g., ligand dockings with ~0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2 Å RMSD, etc., relative to the crystallographic position). However, the RMSD space of possible dockings is remarkably restricted by the size, geometry and flexibility of the particular ligand as well as by the binding site geometry. This is found even with thorough ligand conformational sampling prior to docking. For each complex, the crystallographic ligand conformer was not included in pose prediction, because the bioactive conformation is not known *a priori* in a real world application. For all 30 complexes, the set of docking poses (reflecting both conformational and orientational sampling) and corresponding protein conformations (which may include SLIDE-rotated side chains) were presented to SiteInterlock and the other five scoring functions. All

resulting protein and ligand structural figures were rendered by PyMOL (version 1.5.0.4; Schrödinger, LLC; http://pymol.org).

Evaluating correlation between scoring functions

To assess the degree of monotonicity between two scoring functions (the extent to which they rank dockings in the same order), Spearman's rank correlation coefficient, ρ, was calculated as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where d_i^2 is the difference in the ranks of two poses x_i x_i and y_i for scoring functions x and y, and n is the number of docking poses. Spearman's ρ takes values in the range between -1 and 1, where a perfect monotonic relationship in ranks between two scoring functions exists when ρ=1, and a perfect inverse relationship exists when ρ=-1. A complete absence of correlation in ranking is indicated by ρ=0.

Rigidity analysis

To prepare the series of dockings for rigidity analysis by ProFlex version 5.2 (http://www.kuhnlab.bmb.msu.edu/software/proflex/index.html), hydrogen atoms were added to the protein structures via Reduce[67], and the coordinates of the ligand poses were converted to PDB format. The ligand atom hydrogen-bond donor and acceptor assignment was automated for each docking analyzed by ProFlex, based on the intermolecular interactions identified by SLIDE for that docking. This is more accurate

than assigning hydrogen-bonding rules prior to docking. For instance, a hydroxyl group could potentially act as a hydrogen-bond donor and/or an acceptor. SLIDE determines whether one or both occur, based on evaluation of interaction distances and angles between the protein and ligand for a given ligand orientation[27]. The steps in SiteInterlock (Fig. 3) were designed to test whether a ligand docking close to the known crystallographic orientation and conformation can be detected based on exhibiting greater protein-ligand interface rigidity than is found for incorrect dockings. The first step in the procedure is to select an energy for ProFlex rigidity analysis of the protein structure, determining which hydrogen bonds and salt bridges will be included in the bond network based on their energy values, which are measured as a function of atom type, distance, and angle. This selection of a suitable hydrogen-bond/salt-bridge energy threshold adjusts for the fact that protein structures in the PDB are solved at different temperatures and pressures, in different solvents, and with different fitting and refinement software, all of which affect the prevalence of non-covalent interactions that meet a given set of distance and angle criteria. The native state of most proteins is poised near the rigid to flexible transition energy[19], where the main chain remains structurally stable (mostly rigid) while also exhibiting some flexible regions, which are often relevant to ligand binding [28,20]. The HETHER (Hydrogen-bond Energy ThresHold Estimator for Rigidity analysis) software module developed here is designed to identify that native-like energy threshold. HETHER reads the results of the hydrogen-bond dilution function in ProFlex that mimics the thermal denaturation of a protein[18]. HETHER analyzes changes in the regions of the protein main chain that remain either independently rigid (able to move as rigid bodies relative to each other), mutually rigid, or flexible, as the ProFlex hydrogen-

bond energy (temperature) increases. As the energy increases, non-covalent interactions break, and regions that were rigid become flexible or less coupled to each other. ProFlex reports every energy value at which the size or number of rigid regions in the protein main chain has changed, as well as the non-covalent interactions included in that bond network. From the series of energy values at which main-chain rigidity differences were observed, HETHER selects the lower energy value (the more rigid state) between the two adjacent energy values (structural states) at which the number of independent rigid regions changed the most. This is called the energy threshold (or cutoff) for HETHER and SiteInterlock analysis. This energy threshold detects the point at which the protein is rapidly changing from a rigid to a flexible state[19], when the protein is also sensitive to changes in the interfacial bond network upon ligand interaction. For instance, if there are two independent rigid regions at one energy value, and four at the next higher energy (due to rigid regions breaking apart upon the loss of non-covalent interactions), then the increase in the number of rigid regions is two. If this is the greatest change in the number of rigid regions between any two consecutive energy values, then the bond network of the system with two independent rigid regions will be chosen by HETHER for SiteInterlock analysis of the protein-ligand complex. HETHER only considers the range of energy values at which the main chain is between 25 and 90% rigid (leaving out totally rigid or mostly flexible states), and HETHER defines rigid regions as those containing at least three alpha carbons to avoid including trivial rigid regions such as dipeptides containing proline as the second residue. The rigid-to-flexible transition energy threshold is identified by HETHER for the apo or de-ligated holo version of each protein complex, and then the same energy threshold is used to analyze each docked ligand complex of that

protein. An example of a hydrogen-bond dilution plot and illustration of the energy threshold chosen by HETHER for SiteInterlock analysis is provided in the Supporting Information (Fig. S1).

To quantify the degree of structural rigidity in a protein-ligand complex, we used the continuous flexibility index f_i, which ProFlex computes for each atom i. For atoms in rigid regions, the flexibility index quantifies the degree of rigidity of each atom based on the larger number of constraints in that region relative to the number needed for the region to be just barely rigid; this total-number-of-constraints value for the region is divided by the total number of bonds in that region to define the flexibility index for each atom in the region. The same calculation is done for atoms in flexible regions, which show fewer constraints than are needed for the region to be rigid. Following the rigid region decomposition by ProFlex, each atom is also assigned a rescaled flexibility score f_i' in the range from 0 to 100, where a value of 50 indicates that the atom belongs to an isostatically (just barely) rigid region, and atoms with a flexibility index below 50 or above 50 are part of a rigid region or a flexible region, respectively. This rescaling is done for the convenience of writing flexibility data in the crystallographic temperature factor column of PDB files, typically for 3D visualization with a color spectrum. It should be noted that ProFlex is sensitive to the stereochemical quality of the protein structure being analyzed, particularly the main-chain bond lengths and angles, because they are critical for defining the rigidity of the protein structure as a whole. Thus, we recommend using structural validation tools such as PROCHECK[68], MolProbity[69], and SWISS-MODEL[70] Structural Assessment to evaluate the stereochemical quality of any protein structure before using it as the basis for ProFlex or SiteInterlock analysis. An

example of a structure which is borderline in suitability for ProFlex analysis is a second PDB entry for HIV-1 protease bound to a different inhibitor (relative to that shown in Fig. 1). At the end of the third line of the ProFlex results on holo structures in Fig. 2, this second HIV-1 protease structure is assessed as mostly flexible at the ProFlex energy threshold selected by HETHER for use in SiteInterlock. To understand the basis for this unexpected flexibility relative to the other 29 proteins analyzed, PROCHECK was run. It showed this PDB entry to have a main-chain (Φ, Ψ) angle value distribution that is "unusual" for structures solved at this (1.8 Å) resolution, and its main chain bond angle and Ω (peptide bond planarity) angle distributions are "highly unusual". ProFlex is appropriately sensitive to main-chain stereochemistry, because the main-chain hydrogen bond network is essential for maintaining overall structural integrity. While the SiteInterlock ligand orientation results are reasonable for this protein, as detailed below, in general we would recommend considering an alternative PDB structure with better stereochemistry.

SiteInterlock interfacial rigidity score

Based on the rescaled flexibility index, f_i', the rigidity metric ProteinAvg was computed as the average over the f_i' values of all protein atoms (including hydrogens) within 9 Å of one or more heavy atoms in the docked ligand. Similarly, LigandAvg was calculated as the average of the f_i' values of all ligand atoms in the current docking. As for protein interfacial atoms, the ligand atoms' flexibility index values are influenced by the changes in non-covalent interactions as well as ligand and protein conformational differences

between the different dockings. The final SiteInterlock score was calculated as the average of ProteinAvg and LigandAvg scores:

$$\text{SiteInterlock score} = \frac{1}{2}(\text{ProteinAvg}_{sc} + \text{LigandAvg}_{sc}),$$

where Z-score standardization was first used to rescale ProteinAvg and LigandAvg to fall on the same scale, based on the mean score μ and standard deviation σ of ProteinAvg and LigandAvg values across the docking poses of a target:

$$x_{sc} = \frac{x - \mu}{\sigma}$$

Thus, the SiteInterlock score is an equal weighting of interfacial protein atoms' average rigidity (or flexibility) and interfacial docked ligand atoms' average rigidity (or flexibility), in units of standard deviations above or below the mean value for that set of dockings. This measure of rigidity considers any reorganization of protein and ligand groups upon docking, reflecting the cooperativity of the bond network in the interface. The workflow of the SiteInterlock software, including preparatory steps that may be done with user-preferred tools, and the roles of HETHER and ProFlex, is outlined in Fig. 3. The HETHER, ProFlex, and SiteInterlock software modules are available to academic researchers under GNU General Public License version 3 and to commercial entities by making licensing arrangements; for more information, please visit http://www.kuhnlab.bmb.msu.edu/software or contact the corresponding author.

Other scoring functions

Scoring functions for comparison with SiteInterlock were used with their respective default settings, unless noted otherwise. Values for the docking scoring function X-Score were computed by using X-Score version 1.3, which outputs binding affinities in pK_D units of the different ligand poses as the average of the X-Score scoring functions HPScore, HMScore, and HSScore[71]. DrugScore (DSX) version 0.88 was used[72]. LigScore was executed from the IMP package (version 2.2)[73], using the PoseScore module for ranking ligand orientations[74]. Protein PDB and ligand MOL2 files were prepared for DOCK6 Amber Score (DOCK6 v. 6.3)[75] via their prepare_amber.pl script, using the recommended parameter set in http://dock.compbio.ucsf.edu/DOCK_6/tutorials/amber_score/dock.in. For scoring protein-ligand complexes via AutoDock Vina (v. 1.1.2), protein and ligand files were prepared by using the prepare_ligand4.py and prepare_receptor4.py in the AutoDockTools utilities from the MGLTools package (version1.5.6)[76].

RESULTS AND DISCUSSION

Detecting structural rigidification upon protein-ligand complex formation

To assess whether the native ligand orientation results in a discernable rigidification of the protein-ligand interface, 30 different protein-ligand complexes were analyzed with SiteInterlock (Table I; Fig. 2). Nineteen of the cases were holo protein structures solved in complex with a ligand (Fig. 2A). The native ligand was deleted from the crystal structure, HETHER energy-based selection of hydrogen bonds was performed on the de-

ligated structure, and rigid region decomposition was performed by ProFlex on each of the docked complexes at the same energy threshold. First, our analysis focused on whether the native (crystallographic) complex exhibited greater rigidity in the protein-ligand interface with the ligand present versus absent. This tested whether there is a consistent trend towards rigidification upon complex formation for the ideal case with no significant conformational or orientational inaccuracies in the ligand or protein structure. To quantify the rigidity of a structure, the SiteInterlock score was computed as the equally-weighted sum of the averaged flexibility indices of ligand atoms and interfacial protein atoms (those within 9 Å of non-hydrogen atoms in the ligand). In the majority (17 out of 19) of the holo complexes, interfacial protein atoms were found to become more rigid in the presence of the ligand presented in the crystallographic binding mode (Supporting Information Fig. S2), due to cooperativity of the non-covalent bond network between the molecules. This is consistent with a previous analysis of protein-ligand complexes showing that 71% of protein atoms within 8 Å of ligand atoms in the holo structures have decreased mobility (lower crystallographic temperature factors) relative to their apo states[77]. This phenomenon is illustrated in Fig. 1 for HIV protease[15,22–24]. In two cases, the protein interface in the complex was equally rigid with and without the ligand (Supporting Information Fig. S2). In one of these cases, adenosine kinase (PDB entries 1bx4), π:π or π:cation interactions with the adenosine ring system in the ligand were not assigned as strong non-covalent interactions by ProFlex, suggesting an area for improvement. The possibility of an equally rigid protein site in the presence and absence of ligand also suggested that the role of ligand rigidification in complex formation be considered. The SiteInterlock score, which includes the LigandAvg component as well

as ProteinAvg, was therefore used for analyzing docked complexes. This combination scoring also has the practical advantage of breaking ties in rigidity values between different protein-ligand dockings that could be observed when using ProteinAvg or LigandAvg alone. An example of SiteInterlock rigidity analysis of the crystallographic binding mode versus an inaccurately docked pose is shown for chorismate mutase (Fig. 4). The protein backbone and ligand are colored by rigidity, and it is evident that both the protein and ligand are more rigid in the near-native (0.36 Å ligand RMSD) complex (Fig. 4A) than in the 3.56 Å RMSD ligand docking (Fig. 4B). Reorganization of protein side chains and ligand flexible groups to accommodate the mispositioned ligand yielded decreased rigidity of the protein binding site and flanking beta sheet, while the ligand remained flexible due to few stabilizing interactions. Across all 30 complexes, it was observed that a net decrease in flexibility of the combination of protein and ligand atoms at the interface (the SiteInterlock score) is a signature of native or near-native complexes, rather than both the protein and ligand individually becoming more rigid.

The SiteInterlock approach was then tested for the ability to discriminate and predict the native binding pose from a series of docked poses with increasing RMSD relative to the crystallographic position. Favorable ligand conformations from OMEGA were used as the input to sample a variety of binding poses with SLIDE for the 19 holo protein structures. Only sterically permissible dockings were retained, with no filtering of poses based on docking scores. To reflect the real-world case of protein complex prediction in which the ligand conformation and orientation and the conformations of interfacial protein side chains upon binding are all unknown, apo crystal structures for 11 proteins were also used as the basis for docking. The corresponding holo structures (Table I) were

used to provide the ligand structure as input to conformational sampling for docking and to assess the accuracy of the apo structure dockings selected by SiteInterlock and the other scoring methods.

For chorismate mutase, the range of sampled poses and corresponding SiteInterlock scores appears in Fig. S3A (Supporting Information), showing a funnel-like profile in which the protein-ligand interface becomes increasingly rigid as the ligand RMSD approaches 0 (the crystallographic pose). The prephenic acid ligand pose with the most rigid SiteInterlock score falls within 0.4 Å of the crystallographically observed position. The ability of SiteInterlock score to rank the docking poses from lowest to highest RMSD was then tested for all the complexes. A positive correlation was found between decreasing RMSD and greater rigidity (more negative SiteInterlock score) for 25 out of the 30 cases, which is also apparent when all the dockings are pooled (Fig. S3B). The Spearman rank correlation coefficient (median value of 0.55 across the 30 complexes) between the SiteInterlock score and the docked ligand RMSD indicates that SiteInterlock is well behaved in discriminating among poses across a broad RMSD range.

For predicting the native protein-ligand complex, when the ligand pose with the most rigid SiteInterlock value is identified for each of the 30 complexes (Fig. 5), it is found to be within 0.5 Å RMSD of the best-sampled pose for 14 of the complexes and within 1.5 Å RMSD for 11 others. A poor docking was identified only for the glutamate dehydrogenase complex (3.9 Å ligand RMSD; PDB entry 1bgv). SiteInterlock inclusion of both protein and ligand interfacial rigidity for identifying native-like dockings clearly outperforms using the protein interfacial rigidity value alone (ProteinAvg), especially in avoiding low-accuracy dockings (Fig. 5).

SiteInterlock was then compared with five commonly used methods for evaluating ligand binding to proteins – PoseScore, AutoDock Vina, DSX, DOCK6 Amber Score, and X-Score – which reflect a spectrum of commonly used knowledge-based, empirical and force field scoring functions. SiteInterlock performs competitively with the better of these methods (Fig. 6), performing particularly well in predicting most protein-ligand complexes to within 1-2.5 Å ligand RMSD. SiteInterlock also avoided selecting suboptimal dockings for all but one of the 30 complexes (PDB 1bgv, 3.9 Å RMSD), while four of the five other scoring functions selected poor dockings (5.4-9.3 Å RMSD) for one, two or three apo complexes, respectively (Table II). SiteInterlock also shows strength in avoiding inaccurate (high RMSD) ligand orientations when docking into an apo structure, where the protein is not pre-conformed to bind that ligand (Fig. 6B). Four of the other scoring functions selected poor-accuracy (5.4-9.3 Å RMSD) poses for between one and three of the apo cases, possibly because they were parameterized to favor interaction geometries found in holo structures. However, all scoring functions performed well on the holo structure set (Table II). These results suggest not only that SiteInterlock performs robustly on its own in selecting near-native dockings across a wide range of protein and ligand types, but also that it has unique strengths in ferreting out decoy poses.

Interfacial rigidity as a signature of native protein-ligand interaction

To assess the relationship between SiteInterlock and other scoring function rankings of the same ligand poses, scatter plots were made to compare all pairs of scoring function values (SiteInterlock, PoseScore, AutoDock Vina, DSX, DOCK6 Amber Score, and X-Score) for the same 331 dockings for the full set of complexes (Fig. 7). A narrow, linear

or flame-like pattern in a plot of scoring function *x* versus scoring function *y* values for the dockings indicates that the two scoring functions rank the dockings similarly, whereas a diffuse (globular or more scattered) pattern indicates that the two scoring functions measure different features of the complexes and rank the dockings only partly similarly. The similarity in trends of two scoring functions across the dockings can be summarized by a single number, the nonparametric Spearman rank correlation coefficient, ρ, as shown in Fig. 7. Unlike the Pearson linear correlation coefficient, the Spearman ρ does not assume a linear relationship between the scoring methods being compared. If two scoring methods rank all the dockings in the same order, a Spearman ρ of 1 will be assigned, whereas a value of -1 indicates the methods rank the dockings in exactly the opposite order, and a value of 0 indicates no correlation in their ranking. Most pairs of scoring functions evaluated here have a Spearman ρ value in the range of 0.5-0.8 (Fig. 7), while the correlation between SiteInterlock and other scoring functions is lower, ranging from 0.20-0.26. This indicates that SiteInterlock measures independent feature(s) of the complexes that are not measured by the other methods. SiteInterlock's rigidity measure is novel in that synergy between interactions (their spatial arrangement and coupling) is key to measuring rigidity, rather than just reflecting additive contributions of bonds. Furthermore, this coupling can extend throughout the ligand and binding site rather than being highly localized to the pairs of atoms and functional groups that interact directly. Thus, SiteInterlock can be considered to measure the degree of coupling between interactions in the binding sites, as well as depending on the presence of favorable individual interactions for that coupling to occur.

CONCLUSIONS

SiteInterlock, based on rigidity theory derived from structural mechanics, has been applied here to identify the native complex between a protein and ligand, given the protein structure in either the ligand bound or free conformation and the ligand molecule in a variety of conformations. Several results support the hypothesis that the native complex is characterized by enhanced interfacial rigidity involving both molecules:

- The majority of holo complexes (17 out of 19 diverse proteins) display increased protein rigidity at the interface when the protein is bound, while the remaining two appear equally rigid.
- Including ligand as well as protein interfacial rigidification improves discrimination of the native complex from misdocked complexes.
- SiteInterlock rigidity performs competitively with the best of five commonly used, well-developed docking scoring functions in discriminating near-native poses from a range of decoy poses.
- For the majority (29) of the complexes, SiteInterlock selects ligand poses that are within 2.8 Å RMSD of the native pose, when given a set of sampled (not crystallographic) ligand conformations. For 25 of the complexes, the best-scoring pose is within 1.5 Å RMSD of the best-sampled pose.
- SiteInterlock has the advantage of avoiding very poor dockings (5 Å or greater RMSD), which are an issue for four of the other scoring functions.

More fundamentally, this work shows that rigidification of the cooperative network of non-covalent bonds upon complex formation is a signature of binding interfaces that is sufficient to detect the native complex. This measure of interaction coupling between the protein and ligand, rather than purely additive interactions, may explain why SiteInterlock rigidity values for complexes have a modest correlation with the values of other scoring functions. Thus, SiteInterlock provides a new feature – interfacial rigidity – and a new way of assessing protein-ligand interfaces that can be used alone or in combination with other methods. We anticipate many useful applications of this interfacial rigidity method for structure-based ligand discovery, with the potential to also aid ligand fitting in crystallography for complexes with moderate resolution.

ACKNOWLEDGEMENTS

We would like to thank OpenEye Scientific Software (Santa Fe, NM) for providing academic licenses for the QUACPAC, OMEGA, and OEchem software packages used in this work.

REFERENCES

1. Niesen FH, Berglund H, Vedadi M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat Protoc 2007;2:2212–2221.

2. Brandts JF, Lin LN. Study of strong to ultratight protein interactions using differential scanning calorimetry. Biochemistry 1990;29:6927–6940.

3. Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, Graf E, Carver T, Asel E, Springer BA, Lane P, Salemme FR. High-density miniaturized thermal shift assays as a general strategy for drug discovery. J Biomol Screen 2001;6:429–440.

4. Ericsson UB, Hallberg BM, DeTitta GT, Dekker N, Nordlund P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. Anal Biochem 2006;357:289–298.

5. Vedadi M, Niesen FH, Allali-Hassani A, Fedorov OY, Finerty PJ, Wasney GA, Yeung R, Arrowsmith C, Ball LJ, Berglund H, Hui R, Marsden BD, Nordlund P, Sundstrom M, Weigelt J, Edwards AM. Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. Proc Natl Acad Sci 2006;103:15835–15840.

6. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331.

7. Tang KES, Dill KA. Native protein fluctuations: the conformational-motion temperature and the inverse correlation of protein flexibility with protein stability. J Biomol Struct Dyn 1998;16:397–411.

8. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, Baker D. Computational design of ligand-binding proteins with high affinity and selectivity. Nature 2013;501:212–216.

9. Velazquez-Campoy A, Todd MJ, Freire E. HIV-1 protease inhibitors: enthalpic versus entropic optimization of the binding affinity. Biochemistry 2000;39:2201–2207.

10. Chodera JD, Mobley DL. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. Annu Rev Biophys 2013;42:121–142.

11. Dill KA. Additivity principles in biochemistry. J Biol Chem 1997;272:701–704.

12. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996;77:1905.

13. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2:173–181.

14. Bahar I, Wallqvist A, Covell DG, Jernigan RL. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. Biochemistry 1998;37:1067–1075.

15. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins Struct Funct Bioinf 2001;44:150–165.

16. Maxwell JC. On the calculation of the equilibrium and stiffness of frames. London, Edinburgh, Dublin Philos Mag J Sci 1864;27:294–299.

17. Jacobs DJ, Hendrickson B. An algorithm for two-dimensional rigidity percolation: the pebble game. J Comput Phys 1997;137:346–365.

18. Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA. Identifying protein folding cores from the evolution of flexible regions during unfolding. J Mol Graph Model 2002;21:195–207.

19. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF. Protein unfolding: rigidity lost. Proc Natl Acad Sci USA 2002;99:3540–3545.

20. Zavodszky MI, Lei M, Thorpe MF, Day AR, Kuhn LA. Modeling correlated main-chain motions in proteins for flexible molecular recognition. Proteins Struct Funct Bioinf 2004;57:243–261.

21. Gohlke H, Kuhn LA, Case DA. Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. Proteins Struct Funct Bioinf 2004;56:322–337.

22. Goodman JL, Pagel MD, Stone MJ. Relationships between protein structure and dynamics from a database of NMR-derived backbone order parameters. J Mol Biol 2000;295:963–978.

23. Korn AP, Rose DR. Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. Protein Eng Des Sel 1994;7:961–967.

24. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res 1998;26:4280–4290.

25. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model 2010;50:572–584.

26. Hawkins PCD, Nicholls A. Conformer generation with OMEGA: learning from the data set and the analysis of failures. J Chem Inf Model 2012;52:2919–2936.

27. Zavodszky MI, Sanschagrin PC, Kuhn LA, Korde RS. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. J Comput Aided Mol Des 2002;16:883–902.

28. Jacobs DJ, Kuhn LA, Thorpe MF. Flexible and rigid regions in proteins. In: Rigidity theory and applications. Springer, 2002:357–384.

29. Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD. Essential considerations for using protein–ligand structures in drug discovery. Drug Discov Today 2012;17:1270–1281.

30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.

31. Thoden JB, Miran SG, Phillips JC, Howard AJ, Raushel FM, Holden HM. Carbamoyl phosphate synthetase: caught in the act of glutamine hydrolysis. Biochemistry 1998;37:8825–8831.

32. Conti E, Stachelhaus T, Marahiel MA, Brick P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. EMBO J 1997;16:4174–4183.

33. Song, Hyun Kyu, Se Hui Sohn and SWS. Crystal structure of deoxycytidylate hydroxymethylase from bacteriophage T4, a component of the deoxyribonucleoside triphosphate-synthesizing complex. EMBO J 1999;18:1104–1113.

34. Stillman TJ, Baker PJ, Britton KL, Rice DW. Conformational flexibility in glutamate dehydrogenase. J Mol Biol 1993;234:1131–1139.

35. Mathews II, Erion MD, Ealick SE. Structure of human adenosine kinase at 1.5 A resolution. Biochemistry 1998;37:15607–15620.

36. Lloyd SJ, Lauble H, Prasad GS, Stout CD. The mechanism of aconitase: 1.8 Å resolution crystal structure of the S642A:citrate complex. Protein Sci 2008;8:2655–2662.

37. Kleywegt GJ, Bergfors T, Senn H, Le Motte P, Gsell B, Shudo K, Jones TA. Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure 1994;2:1241–1258.

38. Mangani S, Carloni P, Orioli P. Crystal structure of the complex between carboxypeptidase A and the biproduct analog inhibitor L-benzylsuccinate at 2.0 A resolution. J Mol Biol 1992;223:573–578.

39. Reitzer R, Gruber K, Jogl G, Wagner UG, Bothe H, Buckel W, Kratky C. Glutamate mutase from Clostridium cochlearium: the structure of a coenzyme B12-dependent enzyme provides new mechanistic insights. Structure 1999;7:891–902.

40. Coll M, Knof SH, Ohga Y, Messerschmidt A, Huber R, Moellering H, Rüssmann L, Schumacher G. Enzymatic mechanism of creatine amidinohydrolase as deduced from crystal structures. J Mol Biol 1990;214:597–610.

41. Chook YM, Gray J V, Ke H, Lipscomb WN. The monofunctional chorismate mutase from Bacillus subtilis. Structure determination of chorismate mutase and its complexes with a transition state analog and prephenate,

and implications for the mechanism of the enzymatic reaction. J Mol Biol 1994;240:476–500.

42. Li J, Vrielink A, Brick P, Blow DM. Crystal structure of cholesterol oxidase complexed with a steroid substrate: implications for flavin adenine dinucleotide dependent alcohol oxidases. Biochemistry 1993;32:11507–11515.

43. Cappalonga AM, Alexander RS, Christianson DW. Structural comparison of sulfodiimine and sulfonamide inhibitors in their complexes with zinc enzymes. J Biol Chem 1992;267:19192–19197.

44. Collyer CA, Blow DM. Observations of reaction intermediates and the mechanism of aldose-ketose interconversion by D-xylose isomerase. Proc Natl Acad Sci U S A 1990;87:1362–1366.

45. Ala PJ, DeLoskey RJ, Huston EE, Jadhav PK, Lam PY, Eyermann CJ, Hodge CN, Schadt MC, Lewandowski FA, Weber PC, McCabe DD, Duke JL, Chang CH. Molecular recognition of cyclic urea HIV-1 protease inhibitors. J Biol Chem 1998;273:12325–12331.

46. Sawaya MR, Kraut J. Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence. Biochemistry 1997;36:586–603.

47. Dobrovetsky E, Khutoreskaya G, Seitova A, Cossar D, Edwards AM, Arrowsmith CH, Bountra C, Weigelt J, Bochkarev A. Metabotropic glutamate receptor mglur1 complexed with LY341495 antagonist. To be Publ.

48. Wu B, Chien EYT, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC, Hamel DJ, Kuhn P, Handel TM, Cherezov V, Stevens RC. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. Science 2010;330:1066–1071.

49. Davenport RC, Bash PA, Seaton BA, Karplus M, Petsko GA, Ringe D. Structure of the triosephosphate isomerase-phosphoglycolohydroxamate complex: an analogue of the intermediate on the reaction pathway. Biochemistry 1991;30:5821–5826.

50. Oakley AJ, Bello ML, Battistoni A, Ricci G, Rossjohn J, Villar HO, Parker MW. The structures of human glutathione transferase P1-1 in complex with glutathione and various inhibitors at high resolution. JMolBiol 1997;274:84–100.

51. Oakley AJ, Bello M Lo, Ricci G, Federici G, Parker MW. Evidence for an induced-fit mechanism operating in pi class glutathione transferases. Biochemistry 1998;37:9912–9917.

52. Ren J, Wang Y, Dong Y, Stuart DI. The N-glycosidase mechanism of ribosome-inactivating proteins implied by crystal structures of alpha-momorcharin. Structure 1994;2:7–16.

53. Achari A, Somers DO, Champness JN, Bryant PK, Rosemond J, Stammers DK. Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. NatStructBiol 1997;4:490–497.

54. Sevcik J, Hill CP, Dauter Z, Wilson KS. Complex of ribonuclease from Streptomyces aureofaciens with 2'-GMP at 1.7 A resolution. Acta Crystallogr,SectD 1993;49:257–271.

55. Hsieh-Wilson LC, Schultz PG, Stevens RC. Insights into antibody catalysis: structure of an oxygenation catalyst at 1.9-angstrom resolution. Proc Natl Acad Sci USA 1996;93:5363–5367.

56. Burmeister WP, Henrissat B, Bosso C, Cusack S, Ruigrok RW. Influenza B virus neuraminidase can synthesize its own inhibitor. Structure 1993;1:19–26.

57. Burmeister WP, Ruigrok RW, Cusack S. The 2.2 A resolution crystal structure of influenza B neuraminidase and its complex with sialic acid. EMBO J 1992;11:49–56.

58. Freitag S, Trong I Le, Klumb L, Stayton PS, Stenkamp RE. Structural studies of the streptavidin binding loop. Protein Sci 1997;6:1157–1166.

59. Holden HM, Matthews BW. The binding of L-valyl-L-tryptophan to crystalline thermolysin illustrates the mode of interaction of a product of peptide hydrolysis. JBiolChem 1988;263:3256–3260.

60. English AC, Done SH, Caves LS, Groom CR, Hubbard RE. Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol. Proteins Struct Funct Bioinf 1999;37:628–640.

61. Priestle J., Rahuel J, Rink H, Tones M, Grutter MG. Changes in interactions in complexes of hirudin derivatives and human alpha-thrombin due to different crystal forms. Protein Sci 1993;2:1630–1642.

62. Dekker RJ, Eichinger A, Stoop AA, Bode W, Pannekoek H, Horrevoets AJG. The variable region-1 from tissue-type plasminogen activator confers specificity for plasminogen activator inhibitor-1 to thrombin by facilitating catalysis: release of a kinetic block by a heterologous protein surface loop. J Mol Biol 1999;293:613–627.

63. Nair SK, Krebs JF, Christianson DW, Fierke CA. Structural basis of inhibitor affinity to variants of human carbonic anhydrase II. Biochemistry 1995;34:3981–3989.

64. James MN, Sielecki AR, Brayer GD, Delbaere LT, Bauer CA. Structures of product and inhibitor complexes of Streptomyces griseus protease A at 1.8 A resolution. A model for serine protease catalysis. J Mol Biol 1980;144:43–88.

65. Moult J, Sussman F, James MN. Electron density calculations as an extension of protein structure refinement. Streptomyces griseus protease A at 1.5 A resolution. J Mol Biol 1985;182:555–566.

66. Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem 2002;23:1623–1641.

67. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.

68. Laskowski RA, MacArthur MW, Moss DS, Thornton JM, IUCr. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.

69. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 2007;35:W375–W383.

70. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 2008;4:1–13.

71. 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.

72. Neudert G, Klebe G. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. J Chem Inf Model 2011;51:2731–2745.

73. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol 2012;10:e1001244.

74. Fan H, Schneidman-Duhovny D, Irwin JJ, Dong G, Shoichet BK, Sali A. Statistical Potential for Modeling and Ranking of Protein-Ligand Interactions. J Chem Inf Model 2011;51:3078–3092.

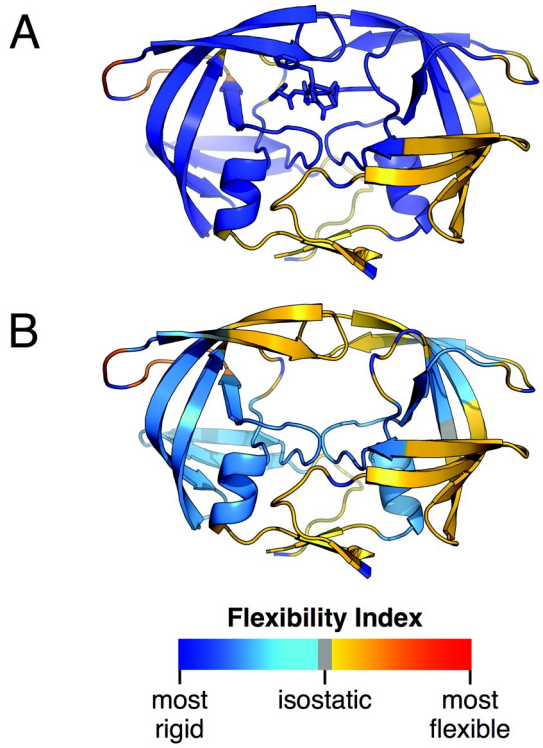
75. Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, Case DA, Kuntz ID, Rizzo RC. DOCK 6: Impact of new features and current docking performance. J Comput Chem 2015;36:1132–1156.

76. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 2009;30:2785–2791.

77. Chao-Yie Yang, Renxiao Wang A, Wang S. A Systematic Analysis of the Effect of Small-Molecule Binding on Protein Flexibility of the Ligand-Binding Sites. J Med Chem 2005; 18:5648-5650.

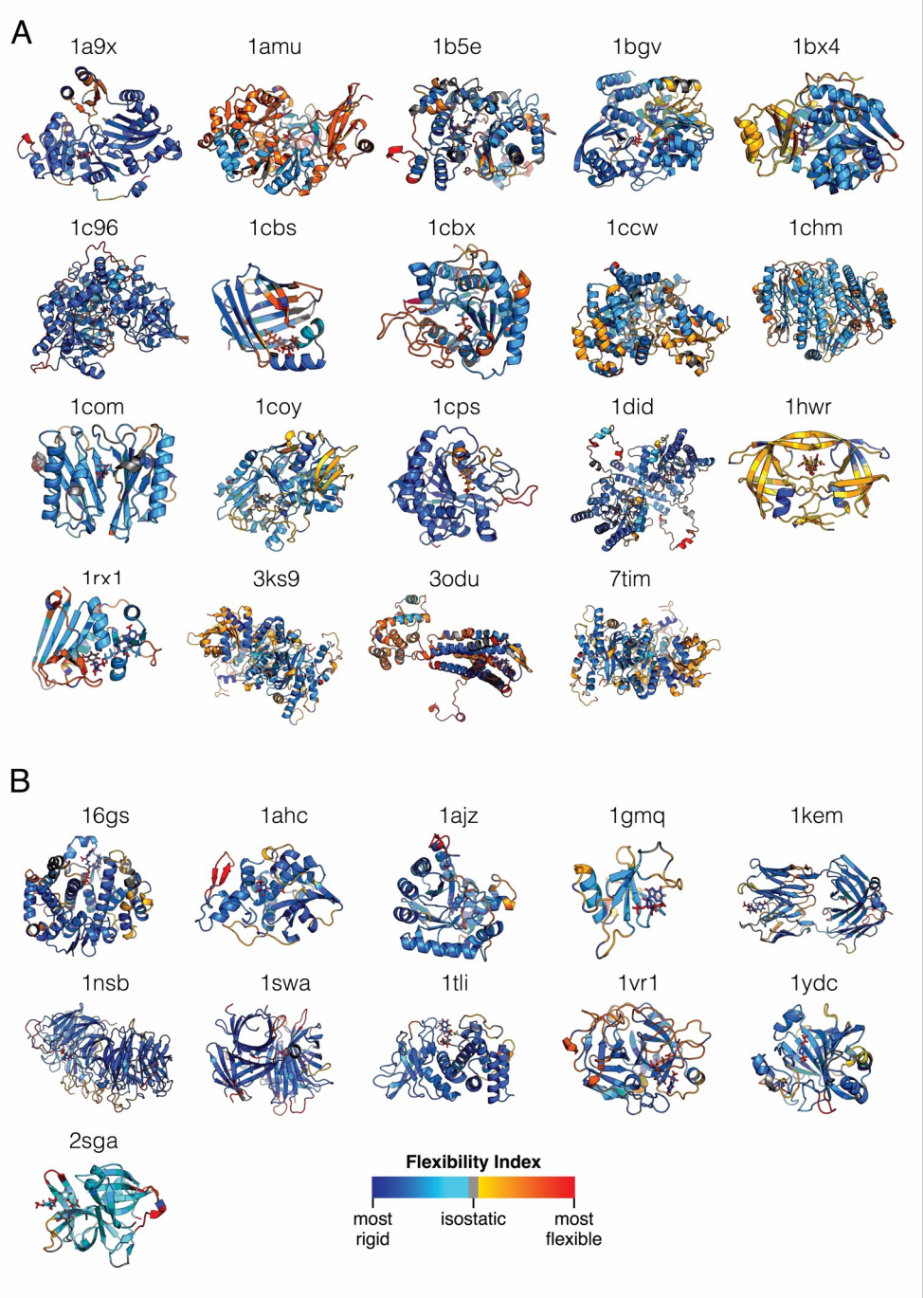
FIGURES & TABLES

Figure 1



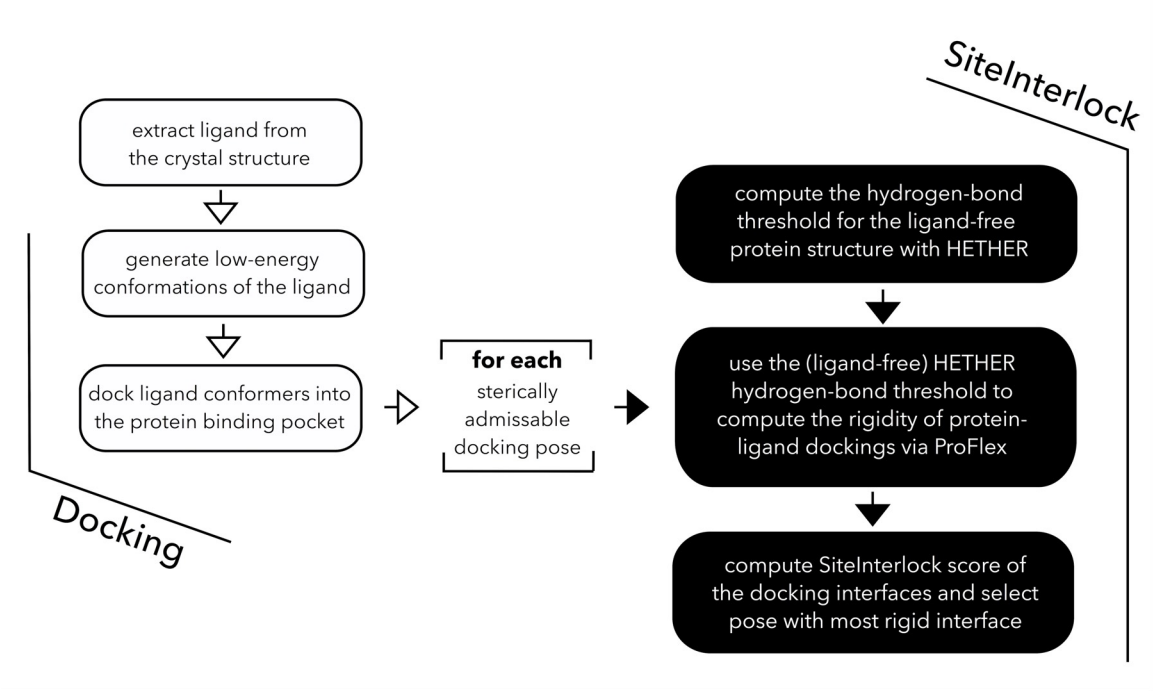
ProFlex assessment of the change in HIV protease flexibility upon inhibitor binding. (A) X-ray crystal structure of HIV-1 protease (PDB entry 1htg) in complex with a penicillin-derived, asymmetric inhibitor. The protein structure is shown in cartoon representation, with the ligand in stick representation in the central binding pocket. The protein main chain and the ligand heavy atoms are colored according to the flexibility indices measured by ProFlex. Note that the inhibitor has induced an asymmetry in flexibility between the two chains of HIV-1 protease, observed in the flexible beta strands to the right, while both halves of the dimer interface are similarly flexible (bottom center). (B) The same PDB structure was analyzed with the ligand removed (while reflecting ligand-induced conformational changes in the protein), indicating that interactions with the ligand in (A) are responsible for rigidifying the beta hairpin flaps (top center) over the ligand, while the flaps become flexible in the absence of the ligand (B).

Figure 2



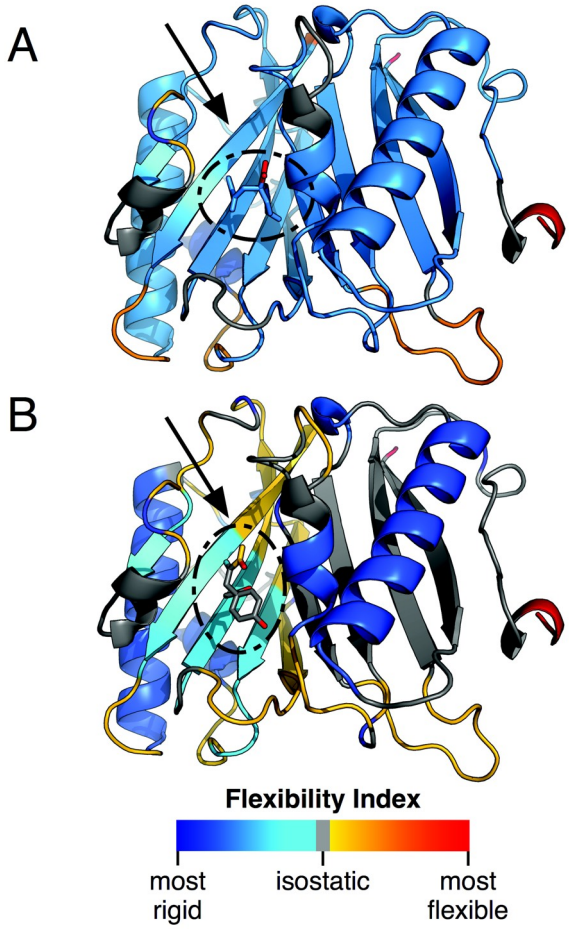
Flexible and rigid regions in 30 diverse protein crystal structures used to evaluate SiteInterlock and other scoring methods for their ability to detect the native ligand binding orientation. (A) Crystal structures of the 19 complexes in the holo structure set. (B) Crystal structures of the 11 apo protein structures. The protein structures (cartoon representation) and ligands (stick representation) are colored to reflect the degree of structural flexibility defined by ProFlex and SiteInterlock, as shown by the color spectrum below.

Figure 3



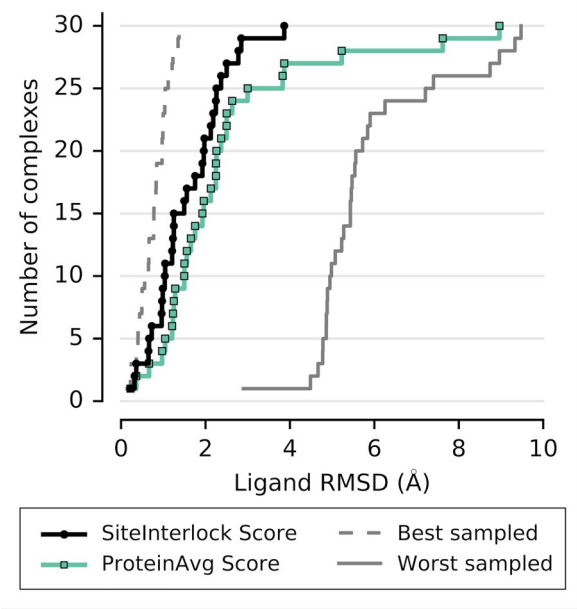
Flowchart of the preparation of input structures for SiteInterlock, followed by the steps of SiteInterlock analysis: HETHER selection of the ProFlex hydrogen-bond energy threshold for the protein in absence of the ligand, ProFlex analysis of protein-ligand interfacial flexibility/rigidity for each docking, and selection of the docking pose with the greatest interfacial rigidity.

Figure 4



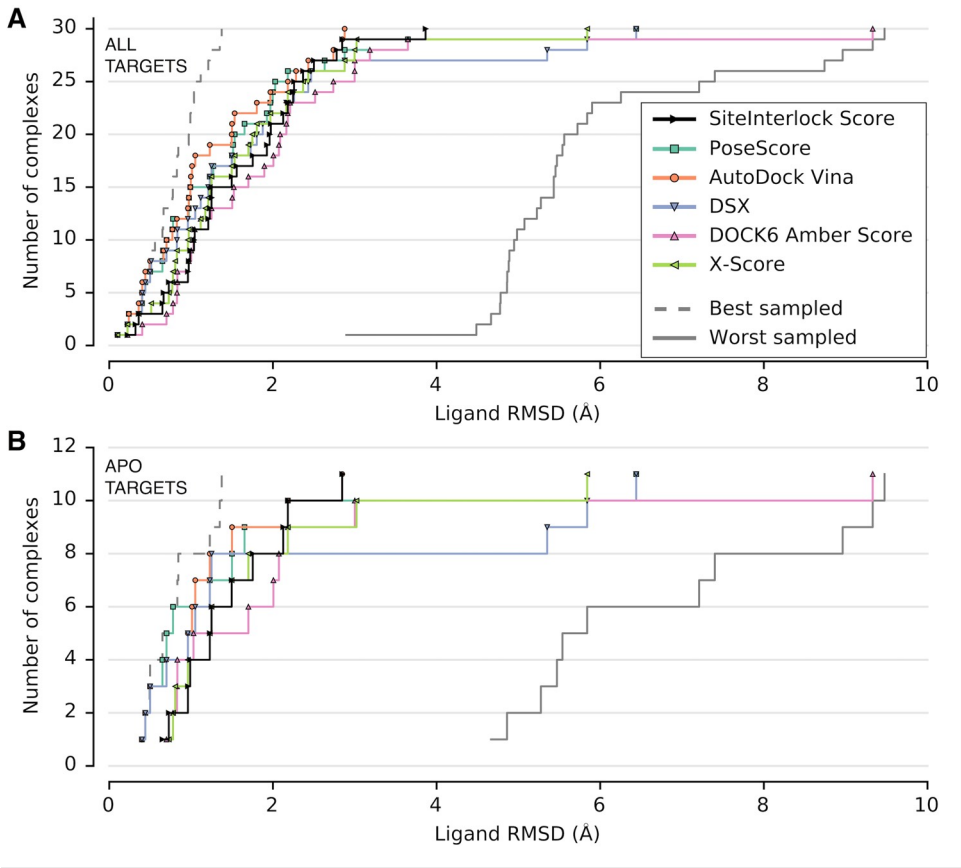
Changes in structural flexibility of the complex of monofunctional chorismate mutase with its enzymatic product, prephenic acid, depending on native-like (PDB entry 1com) versus non-native dockings of the ligand. Arrows point to prephenic acid in the binding site. (A) Near-native docking pose (ligand RMSD 0.36 Å). (B) Inaccurate docking pose (ligand RMSD 3.56 Å). Note the enhanced rigidity of both the binding site and the ligand in the native pose relative to the misdocked pose.

Figure 5



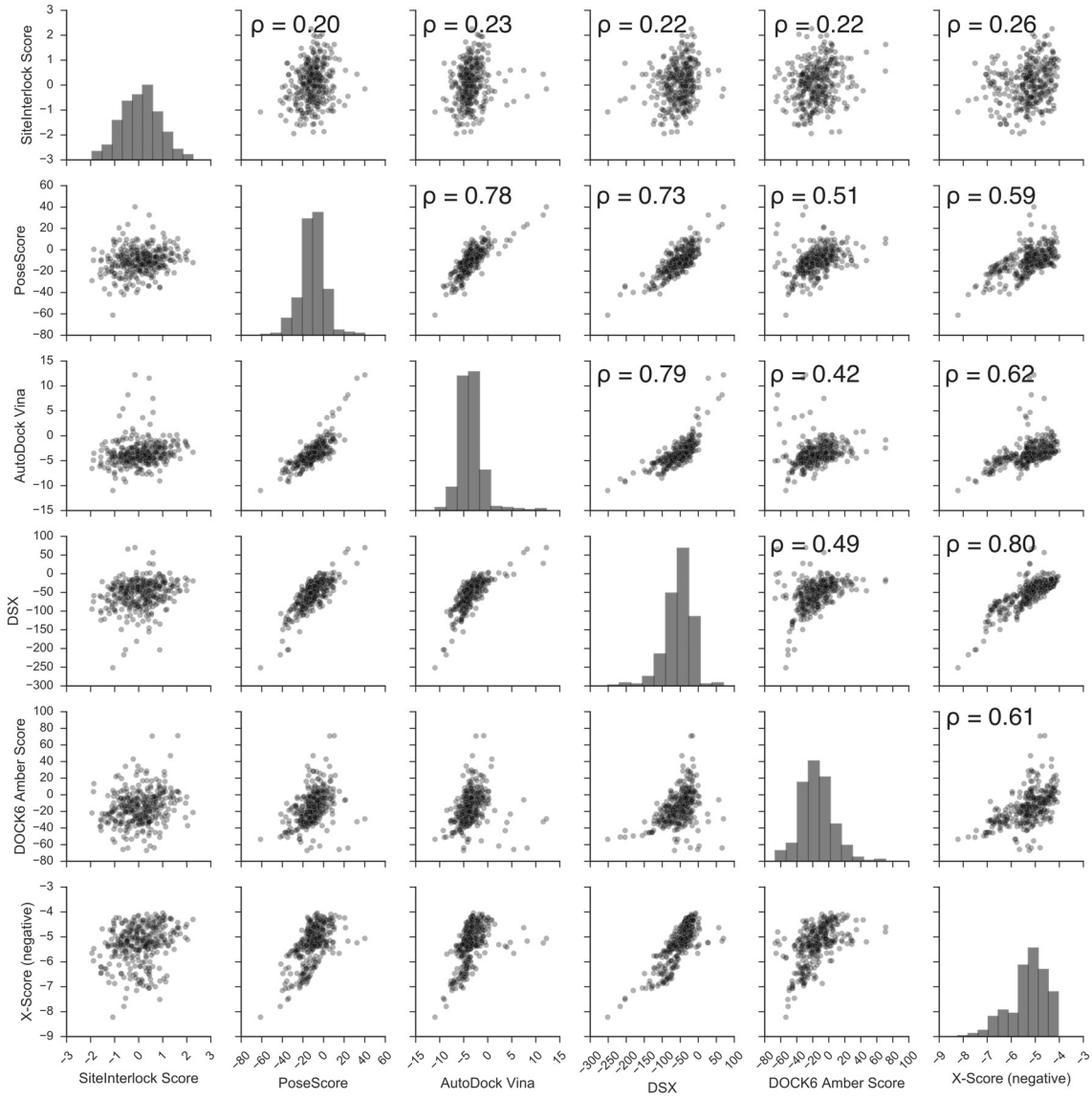
Enrichment plot comparing the SiteInterlock score with the ProteinAvg score for selecting near-native docking poses for the 30 targets. Here, the *y* axis value shows the number of complexes for which the best-scoring pose selected by SiteInterlock (black curve) and ProteinAvg (green curve) is within the ligand RMSD value shown on the *x* axis. For example, we see that the best-scoring ligand pose selected by SiteInterlock is under 3 Å RMSD in 29 of the 30 cases. The combination of protein and ligand interfacial rigidity in the SiteInterlock score is apparently a better predictor of native-like poses than protein rigidity alone (ProteinAvg). The gray dashed line indicates the best scoring performance possible, if the best-sampled pose were selected for each complex, and the solid dashed line indicates the worst possible performance, based on selecting the worst-RMSD pose for each complex.

Figure 6



Enrichment plot, as described in Figure 5, comparing the accuracy of pose selection of SiteInterlock (black line with square symbols) with five different docking scoring functions (see color legend on plot), bounded by the curves showing the best-sampled (dashed gray line) and worst sampled (solid gray line) poses for the complexes. (A) Performance for all 30 protein targets. (B) The 11 apo protein cases only, showing that four of the other scoring functions select poor-accuracy (5.4-9.3 Å RMSD) poses for between one and three of the apo cases, possibly because they were parameterized to favor interaction geometries found in holo structures.

Figure 7



Comparison of the values of different scoring functions for all docking poses (n = 331), as a matrix of pairwise scatter plots. Spearman's rank correlation coefficient, denoted as ρ, is provided for each scoring function pair in the upper triangle, measuring the extent to which the two scoring functions shown in each plot rank the poses in the same order. Along the diagonal appears the histogram of the number of docking poses as a function of score value for each scoring function. The standardization of SiteInterlock score components ProteinAvg and LigandAvg leads to a Gaussian distribution of scores, which helps to distinguish good from average from poor dockings. Some of the other scoring functions exhibit narrow distributions, making the discrimination of good protein-ligand orientations more challenging. To facilitate the comparisons here, X-Score values (last column and row) are presented multiplied by -1, so that more negative values appear as more favorable.

Table I

Protein-ligand complexes analyzed. The 19 complexes in which the holo conformation of the protein was used for docking and SiteInterlock analysis are listed first, followed by the 11 complexes in which the apo conformation of the protein was used. The binding site RMSD is based on main-chain superposition of the apo onto the holo structure, with the binding site atoms in the two structures defined as those within 9 Å of the ligand.

PDB entry (holo/apo)	Protein	Ligand	Resolution (Ångstrom)	Holo-apo binding site RMSD (Å)
1a9x[31] / -	carbamoyl phosphate synthetase	L-ornithine	1.80	-
1amu[32] / -	gramidicin synthetase 1	L-phenylalanine	1.90	-
1b5e[33] / -	deoxycytidylate hydroxymethylase	deoxycytidylic acid	1.60	-
1bgv[34] / -	glutamate dehydrogenase	L-glutamate	1.90	-
1bx4[35] / -	adenosine kinase	adenosine	1.50	-
1c96[36] / -	mitochondrial aconitase	citrate anion-iron/sulfur cluster	1.81	-
1cbs[37] / -	retinoic acid binding protein	retinoic acid	1.80	-
1cbx[38] / -	carboxypeptidase A	L-benzylsuccinic acid	2.00	-
1ccw[39] / -	glutamate mutase	D-tartaric acid	1.60	-
1chm[40] / -	creatine amidinohydrolase	carbamoyl sarcosine	1.90	-
1com[41] / -	chorismate mutase	prephenic acid	2.20	-
1coy[42] / -	cholesterol oxidase	dehydroepiandrosterone	1.80	-
1cps[43] / -	carboxypeptidase A	sulfodiimine	2.25	-
1did[44] / -	D-xylose isomerase	2,5-dideoxy-2,5-imino-D-glucitol	2.50	-
1hwr[45] / -	HIV-1 protease	Xk216	1.80	-
1rx1[46] / -	dihydrofolate reductase	NADP+	2.00	-
3ks9[47] / -	metabotopic glutamate receptor	Z99	1.90	-
3odu[48] / -	G-protein-coupled chemokine receptor	IT1t	2.50	-
7tim[49] / -	triosephosphate isomerase	phosphoglycolohydroxamic	1.90	
10gs[50] / 16gs[51]	glutathione S-transferase	L-cysteine amide	2.20 / 1.90	0.27
1ahb[52] / 1ahc[52]	alpha-momorcharin	formycin-5'-monophosphate	1.90 / 2.00	0.75
1aj2[53] / 1ajz[53]	dihydropteroate synthase	pterin diphosphate	2.20 / 2.00	0.64
1gmr[54] / 1gmq[54]	ribonuclease	guanosine-2'-monophosphate	1.77 / 1.80	0.46
1kel[55] / 1kem[55]	sulfide oxidase antibody	methylphosphonic acid	1.90 / 2.20	0.68
1nsc[56] / 1nsb[57]	influenza B neuraminidase	O-sialic acid	1.70 / 2.20	0.32
1swd[58] / 1swa[58]	streptavidin	biotin	1.90 / 1.90	0.52
3tmn[59] / 1tli[60]	thermolysin	tryptophan	1.70 / 2.05	0.69
1tmt[61] / 1vr1[62]	alpha-thrombin	D-phenylalanine	2.20 / 1.90	0.66
1ydb[63] / 1ydc[63]	carbonic anhydrase II	acetazolamide	1.90 / 1.95	0.30
5sga[64] / 2sga[65]	proteinase A	acetyl group	1.80 / 1.50	0.19

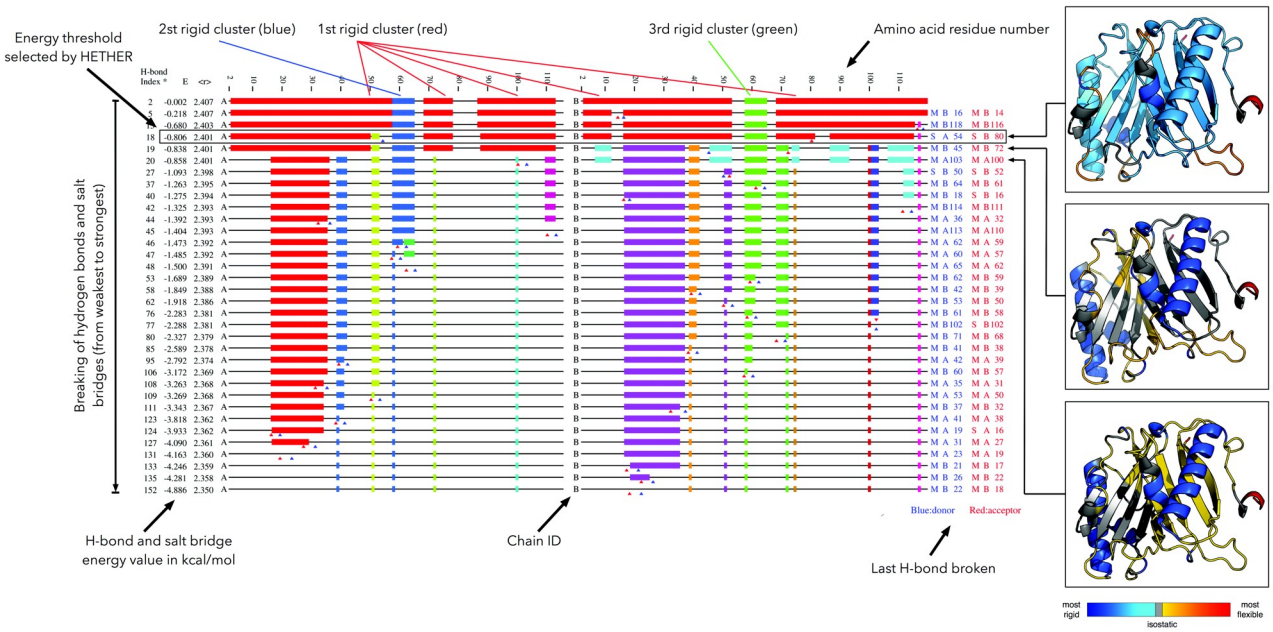
Table II

Ligand RMSD values of the best predicted docking poses

PDB entry	Holo/apo	Best sampled	SiteInterlock Score	PoseScore	AutoDock Vina	DSX	DOCK6 Amber Score	X-Score
1a9x	holo	0.66	0.66	3.65	0.66	1.88	3.65	3.00
1amu	holo	0.40	2.37	0.40	0.40	0.40	0.40	2.37
1b5e	holo	1.12	1.93	1.93	2.28	1.12	1.12	1.12
1bgv	holo	0.56	3.87	2.03	2.43	2.43	3.00	2.43
1bx4	holo	0.10	0.32	0.10	0.10	0.10	2.21	0.10
1c96	holo	1.04	1.04	2.88	2.88	2.88	1.04	2.88
1cbs	holo	1.00	2.25	1.27	1.00	2.47	1.50	1.75
1cbx	holo	0.78	0.97	0.97	0.97	2.26	1.52	0.97
1ccw	holo	0.83	1.97	2.63	0.83	0.83	2.09	0.83
1chm	holo	1.04	1.04	1.97	1.97	1.97	1.90	1.97
1com	holo	0.36	0.36	1.51	0.36	0.36	1.00	0.36
1coy	holo	0.24	1.96	0.24	0.24	0.51	3.19	0.51
1cps	holo	0.97	2.26	1.53	1.53	1.73	0.97	1.53
1did	holo	0.97	2.78	0.97	1.80	1.80	2.52	1.80
1hwr	holo	0.77	1.56	0.77	0.77	0.83	0.83	1.17
1rx1	holo	0.22	0.22	0.22	0.22	0.22	0.22	0.22
3ks9	holo	1.21	1.21	2.00	2.74	1.21	2.74	1.21
3odu	holo	0.99	2.50	0.99	0.99	2.16	2.16	2.16
7tim	holo	0.66	1.25	0.66	1.50	1.50	1.25	0.77
16gs	apo	0.78	1.75	0.78	1.05	1.05	0.78	0.78
1ahc	apo	0.84	1.25	1.50	1.50	1.25	3.00	1.25
1ajz	apo	1.35	2.85	6.44	2.85	6.44	9.33	3.02
1gmq	apo	1.23	1.23	1.23	1.23	1.23	2.07	1.23
1kem	apo	0.44	0.99	0.44	0.44	0.44	2.01	0.73
1nsb	apo	0.50	1.50	0.70	0.70	0.70	0.70	1.50
1swa	apo	0.50	2.13	0.50	0.50	0.50	1.70	1.70
1tli	apo	0.65	0.65	0.65	1.01	5.84	0.83	5.84
1vr1	apo	0.83	0.96	1.65	0.96	0.96	0.83	0.96
1ydc	apo	1.37	2.18	2.18	2.18	5.35	2.18	2.18
2sga	apo	0.40	0.73	0.40	0.40	0.40	1.03	0.81

SUPPORTING INFORMATION

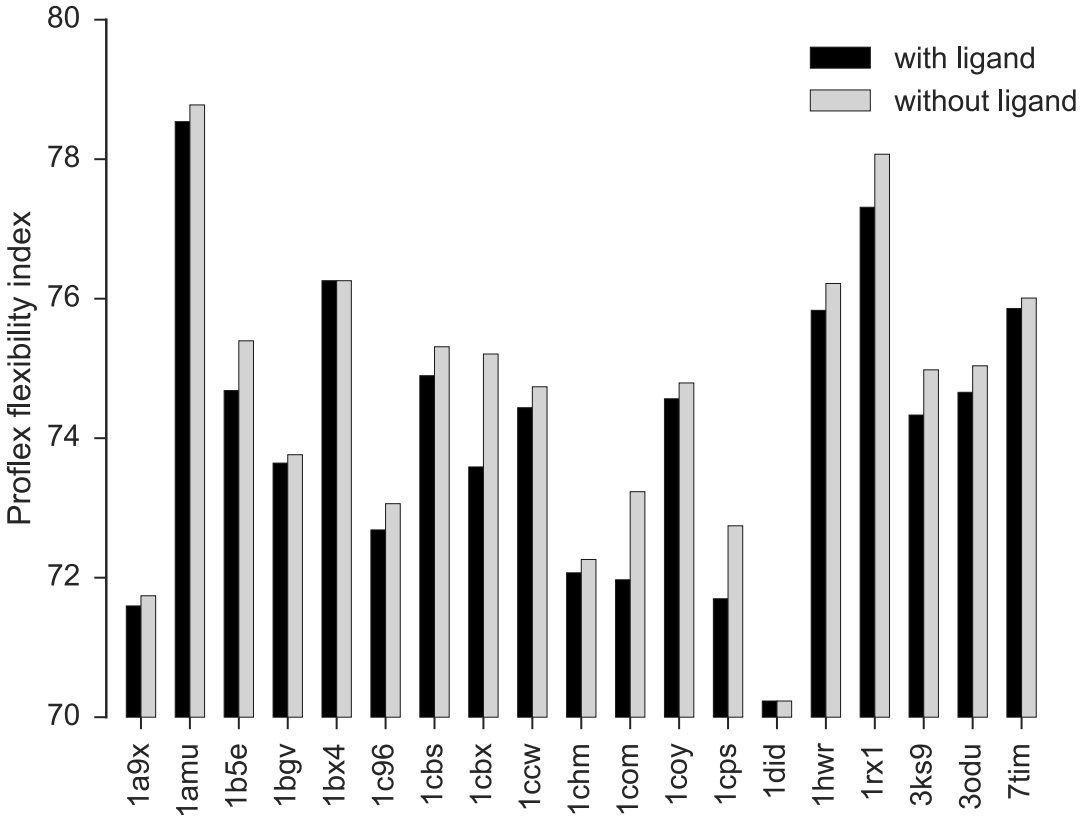
Supporting Information Figure 1



ProFlex hydrogen-bond dilution plot of the de-ligated protein structure of a monofunctional chorismate mutase from Bacillus subtilis (PDB code: 1com), showing the transition from mostly rigid to mostly flexible as hydrogen bonds and salt bridges are broken with increasing energy. The HETHER module of SiteInterlock is designed to identify the energy just before the protein becomes substantially flexible, as described below. The distinct lines in this plot show the rigid and flexible regions of the protein at different energy values, with successive lines representing increasingly flexible states of the protein as the energy level (temperature) increases. Residues of the protein chain are numbered from left to right at the top of the plot. At a given energy value, the thick, colored blocks in each row indicate the rigid clusters of the protein main chain, with a different color used for each independently rigid cluster of atoms. The thin, black lines correspond to intervening flexible regions observed in the protein bond network at that energy. A rigid region may be comprised of residues that are not contiguous in sequence; thus, blocks of residues with the same color indicate residues belonging to the same mutually-rigid region. The energy value for each row is listed in the second column from the left. The first row shows the predicted state of the protein when all hydrogen bonds and salt bridges are included in the bond network. The number of salt bridges and hydrogen bonds is listed in the leftmost column. The third column shows the average number of bonds connecting to each atom (averaged over all atoms in the protein) at that energy level, including covalent single and double bonds, bond-coordination constraints (constraining sp3 and sp2 centers in the correct geometry), hydrophobic tethers, hydrogen bonds, and salt bridges. For instance, the second row, at an energy value of -0.218 kcal/mol, shows the rigid and flexible regions in the protein when all hydrogen bonds and salt bridges with an energy of -0.218 kcal/mol or stronger are included in the bond network. Moving down the rows of the plot, the energy values increase and hydrogen bonds and salt bridges are incrementally broken (from weakest to strongest), resulting in an overall increase of flexible regions in the protein structures indicated by the intervening, black lines and fragmentation of rigid regions.

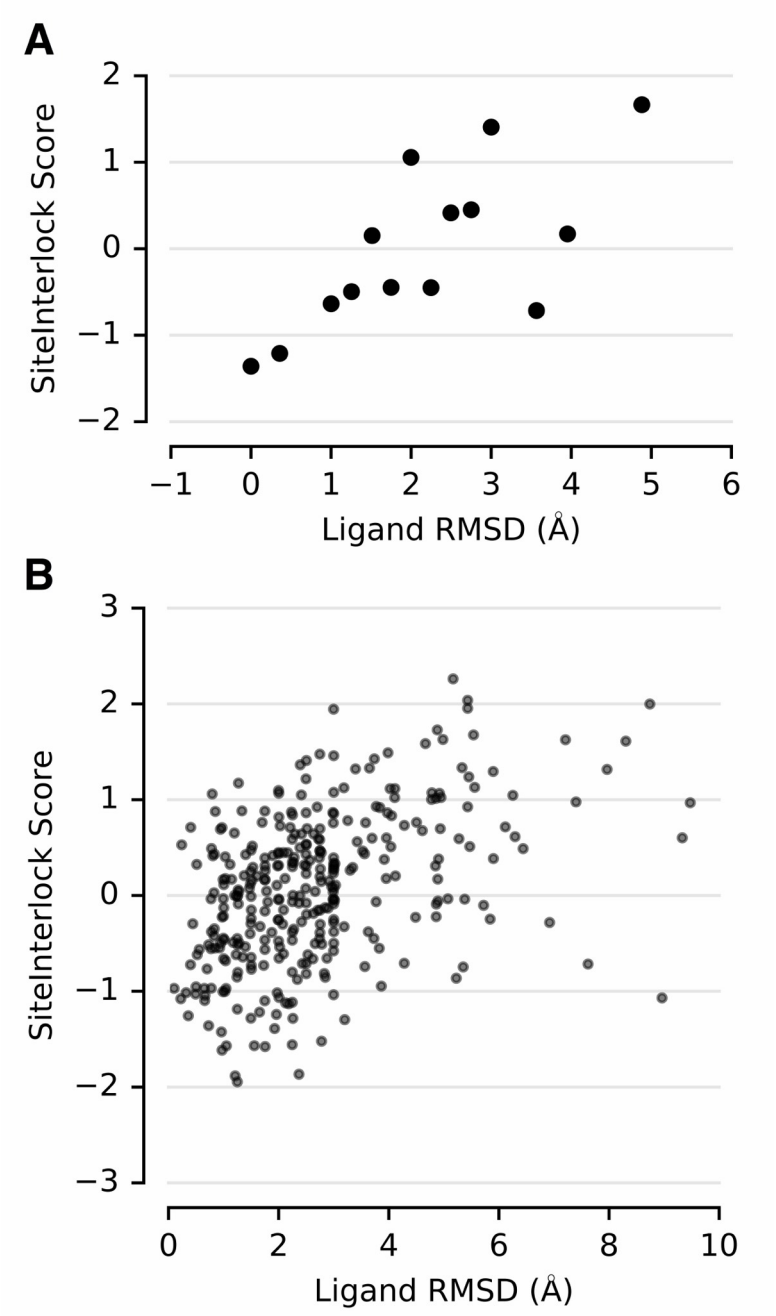
The energy value selected by HETHER is highlighted by the black frame shown at -0.806 kcal/mol, in which the main chain is mostly rigid (comprised by the large rigid region shown in red, plus two ~10-residue independent rigid regions colored in blue and green, and a very short rigid region in lime green appearing at residue 50). This state shows some residual flexibility that is sensitive to native-like ligand interactions, as described in the results in the main text. The rigid and flexible regions mapped onto the corresponding, ligand-free protein structure at different energy levels are shown at the far right, now colored by flexibility index (with colors defined in the spectrum bar shown beneath the structures). At the next energy step (-0.838 kcal/mol) above that chosen by HETHER, the protein structure decomposes into eight rigid clusters (red, yellow, blue, green, cyan, orange, lime green, and dark blue), which results in a structure with about one-third of the main chain being flexible. Thus, HETHER selected the last substantially stable state of the protein structure, as intended.

Supporting Information Figure 2



Rigidity of interfacial protein atoms (within 9 Å of ligand heavy atoms) in the presence (black bars) and absence (gray bars) of the crystallographic ligand pose for the 19 holo structures. Lower ProFlex values indicate greater rigidity. For 17 cases, the protein interface is more rigid in the presence of the ligand, and for 2 cases (PDB entries 1bx4 and 1did), it is equally rigid.

Supporting Information Figure 3



Relationship between the SiteInterlock score and ligand RMSD relative to the crystallographic pose for (A) dockings spanning the RMSD range of 0-5 Å for prephenic acid in complex with chorismate mutase (PDB entry 1com; also see Figure 4 for SiteInterlock results on two of these poses) and (B) 331 dockings from all 30 protein-ligand complexes. A funnel-like tendency is seen that discriminates more native-like dockings (closer to 0 Å RMSD) based on these dockings having more negative (rigid) SiteInterlock scores, particularly for dockings with RMSD values of ≤ 3 Å.