Docking Validation Resources: Protein Family and Ligand Flexibility Experiments

Sudipto Mukherjee,[†] Trent E. Balius,[†] and Robert C. Rizzo*[,†,‡]

Department of Applied Mathematics and Statistics, and Institute of Chemical Biology and Drug Discovery,
Stony Brook University, Stony Brook, New York 11794, United States

Received May 19, 2010

A database consisting of 780 ligand−receptor complexes, termed SB2010, has been derived from the Protein Databank to evaluate the accuracy of docking protocols for regenerating bound ligand conformations. The goal is to provide easily accessible community resources for development of improved procedures to aid virtual screening for ligands with a wide range of flexibilities. Three core experiments using the program DOCK, which employ rigid (RGD), fixed anchor (FAD), and flexible (FLX) protocols, were used to gauge performance by several different metrics: (1) global results, (2) ligand flexibility, (3) protein family, and (4) cross-docking. Global spectrum plots of successes and failures vs rmsd reveal well-defined inflection regions, which suggest the commonly used 2 Å criteria is a reasonable choice for defining success. Across all 780 systems, success tracks with the relative difficulty of the calculations: RGD (82.3%) > FAD (78.1%) > FLX (63.8%). In general, failures due to scoring strongly outweigh those due to sampling. Subsets of SB2010 grouped by ligand flexibility (7-or-less, 8-to-15, and 15-plus rotatable bonds) reveal that success degrades linearly for FAD and FLX protocols, in contrast to RGD, which remains constant. Despite the challenges associated with FLX anchor orientation and on-the-fly flexible growth, success rates for the 7-or-less (74.5%) and, in particular, the 8-to-15 (55.2%) subset are encouraging. Poorer results for the very flexible 15-plus set (39.3%) indicate substantial room for improvement. Family-based success appears largely independent of ligand flexibility, suggesting a strong dependence on the binding site environment. For example, zinc-containing proteins are generally problematic, despite moderately flexible ligands. Finally, representative cross-docking examples, for carbonic anhydrase, thermolysin, and neuraminidase families, show the utility of family-based analysis for rapid identification of particularly good or bad docking trends, and the type of failures involved (scoring/sampling), which will likely be of interest to researchers making specific receptor choices for virtual screening. SB2010 is available for download at http://rizzolab.org.

INTRODUCTION

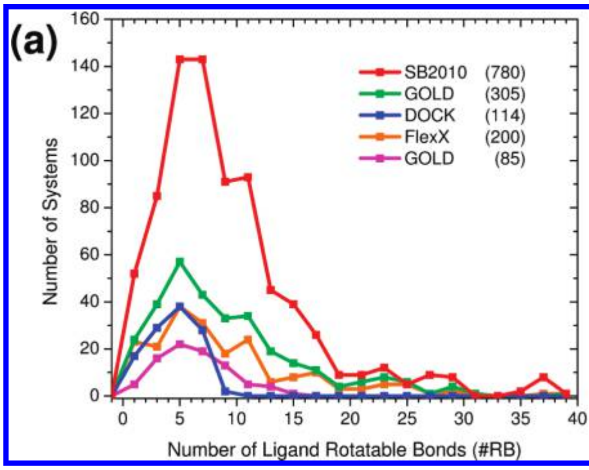
A central challenge in computational structure-based drug discovery is routine and robust prediction of the bound geometry and interactions of small organic molecules (ligands) with their biological targets (receptors). Computationally, the procedure is referred to as docking, and the field has seen widespread growth since the first program DOCK[1] was introduced in 1982. Since then, numerous docking programs have come into use, including Autodock,[2] updated versions of DOCK,[3−5] FlexX,[6] FRED,[7] Glide,[8,9] and GOLD,[10] among others. Although there are many success stories from academic and industrial groups,[11−15] for both the new and expert users alike, it would be desirable if docking methods were generally more reliable, more robust, and easier to use. In particular, validation controls to assess accuracy[16−18] are particularly important, as it is critical that each user assess their unique docking setup(s) and computational infrastructure(s) prior to embarking on a project. A primary focus of this work is the construction of a docking database to aid users in establishing the accuracy of their docking codebases and protocols.

In practice, docking is used to accomplish two primary objectives: (1) prediction of the binding geometry (pose) for a single molecule to a known target and (2) screening a virtual database of molecules to a target, filtering for a small subset of predicted actives. In both cases, good pose accuracy is important. For virtual screening, it is additionally important that active ligands score better than other decoy molecules (enrichment). Focusing on pose accuracy, the central idea is to evaluate how well a given docking method can recapitulate bound ligand conformations using crystallographically determined binding modes contained in the Protein Data Bank (PDB)[19] as a reference. Several PDB-derived databases that provide useful benchmarks have been previously described, usually being derived in conjunction with development of the docking programs themselves. A partial list includes databases associated with the programs GOLD,[10,20] FlexX,[6] and DOCK5−6.[4,5] Recently, there have been efforts to automate database construction, such as the notably large-scale DOCKBlaster ($N = 7755$)[21] study. Other relevant databases include, for example, DUD,[17] which provides sets of active and decoy ligands to evaluate enrichment, and Binding MOAD,[22] Pdbbind,[23,24] BindingDB,[25] and LPDB,[26] which include experimental binding energies for corresponding PDB entries to aid scoring function development.

Prompted by the need for a large versatile testset to aid method development and virtual screening projects ongoing in our laboratory, we have constructed a docking database termed SB2010 (Stony Brook year 2010) consisting of 780

* Corresponding author e-mail: rizzorc@gmail.com.
[†] Department of Applied Mathematics and Statistics.
[‡] Institute of Chemical Biology and Drug Discovery.

10.1021/ci1001982
Published on Web 10/29/2010



(b)

Pdb code	Drug/Inhibitor	#RB	Indication
8KME	SEL2770	29	anticoagulant
1AGM	acarbose	22	type 2 diabetes
1G7F	PNU177496	19	diabetes
1CNY	benzene sulfonamide	17	glaucoma
4TMN	peptidomimetic	16	antibacterial
1SDT	indinavir	16	antiviral HIV/AIDS
1HWK	atorvastatin	15	heart disease
1KPM	vitamin E	13	vitamin supplement
1H22	huperazine analog	13	alzheimers
1HFC	hydroxamate inhibitor	13	anticancer
1NJJ	geneticin	12	antibiotic
1A4G	zanamivir	10	antiviral influenza

Figure 1. (a) Ligand flexibility histogram for SB2010 in comparison with other docking databases. (b) Representative examples of FDA approved drugs and medically significant experimental inhibitors in SB2010 having ≥10 rotatable bonds. Ligand flexibility (the number of rotatable bonds) for all data sets was computed using DOCK6.4.

protein−ligand complexes derived from the PDB (Table S1, Supporting Information). Figure 1a compares SB2010 (red histogram) with four of the databases noted above, (i) GOLD305[10] (green, $N = 305$), (ii) FLEXX200 (blue, $N = 200$), (iii) DOCK114 (orange, $N = 114$), and (iv) GOLD85[20] (magenta, $N = 85$), in terms of overall ligand flexibility. The larger size of SB2010 compared with other databases leads to greater numbers of ligands that are more flexible. In particular, SB2010 contains 266 ligands with ≥10 rotatable bonds compared with the other sets that contain between 0 and 109 entries (Figure 1b red vs other color histograms). This is important, as many approved drugs and medically significant experimental inhibitors have more than 10 rotatable bonds, as illustrated in Figure 1b. Thus, the database will likely be of use to researchers wishing to calibrate virtual screening protocols for a wide range of ligand flexibilities.

Partitioning of the SB2010 database into different subsets was also performed to characterize how ligand flexibility and/or binding site environment (i.e., protein family) affects pose accuracy prediction. Entries were grouped into subsets having 7-or-less, 8-to-15, and 15-plus rotatable bonds. Alternatively, the database was arranged into specific protein families with seven or greater members ($N = 25$) or between two and seven members ($N = 25$), which together constitute 83.9% of the data (655/780). Family-based groupings additionally allow for assessment of cross-docking success. All entries in SB2010 were assessed to be sure there were no intermolecular clashes for cognate protein−ligand pairs, significant numbers of missing side-chain atoms near the binding site, or for which an intermolecular energy minimization substantially moved the ligand. Database files were prepared to be immediately compatible with the program DOCK[4] (MOL2, GRID, SPHGEN formats).

It is important to emphasize that the calculations in this study employ a rigid receptor approximation with the primary focus being on ligand-based sampling. However, for systems in which significant induced fit effects occur for the receptor or large conformational changes are observed, a rigid approximation is likely not appropriate. Previous work addressing protein flexibility in the context of docking include studies by Knegtel et al.,[27] Sandak et al.,[28] Cavasotto et al.,[29] Sherman et al.,[30] Moitessier et al.,[31] and Amaro et al.,[32] to name a few. In a general sense, the present study does account for receptor conformational variability through cross-docking experiments that employ family-based ensembles of crystal structures, somewhat similar, for example, to using ensembles derived from molecular dynamics. And, the calculations employ 6−9 van der Waals potentials, which have been shown to crudely account for partial receptor flexibility through softening of the intermolecular energy landscape.[33] In any event, despite the approximations, for a large number of cases use of a fixed receptor for docking is reasonable, especially when using binding site coordinates in which a representative (parent) ligand or substrate was cocrystallized.

A long-term objective of our work is development of improved sampling and scoring methods to enhance docking accuracy. The goals of this specific study are (i) to construct a docking database with a wide range of ligand flexibilities, (ii) to evaluate the accuracy of three distinct protocols for recapitulating experimentally observed binding poses using the program DOCK, and (iii) to characterize docking outcomes for the testset as a whole and subsets based on ligand flexibility, protein family, and cross-docking. To improve protocols and methods, continued evaluation of success and failure across a wide-range of systems is important. SB2010 is available for download at http://rizzolab.org

COMPUTATIONAL METHODS

Success, Sampling Failures, and Scoring Failures. The ability to predict how small molecules geometrically interact with protein and nucleic acid targets remains an important and challenging problem. For virtual screening to be most useful, an implicit assumption is that the geometric poses generated are accurate. In this report, unless otherwise noted, a docked pose within 2 Å heavy atom root-mean-square-deviation (rmsd) of a crystallographic pose is considered a successful match. Accuracy is gauged by three criteria, (1) success, (2) sampling failures, and (3) scoring failures, which together always sum to make up 100% of the possible docking outcomes. *Success* is defined when the best-scoring pose matches the crystallographic pose. This definition mirrors typical screening applications, which save only a single pose and thus is comparable across different docking programs and platforms. *Sampling failures* quantify the inability of a given protocol to generate at least one pose

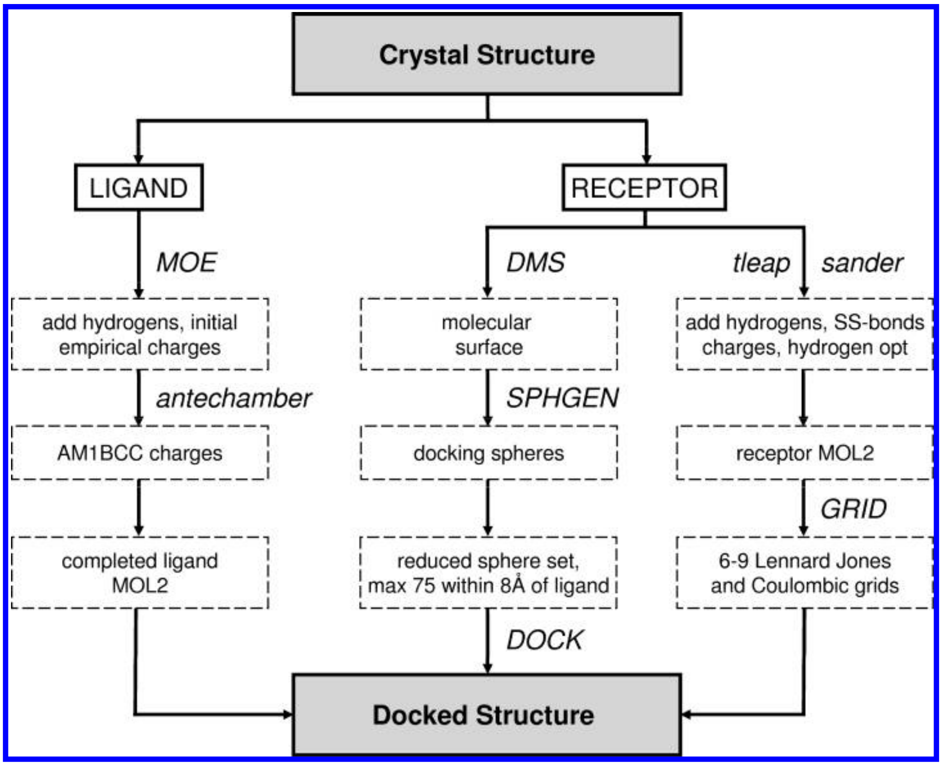


Figure 2. Flowchart depicting the protocol used for ligand and receptor preparation for each protein−ligand complex downloaded from the Protein Data Bank. Details of each step are discussed in the text.

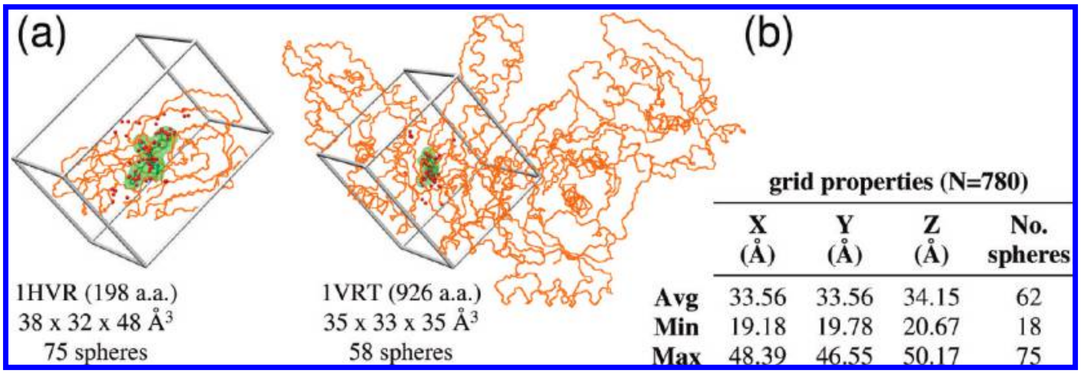
similar to the experimental structure within the ensemble of poses generated. Importantly, a docking experiment that fails sampling cannot benefit from rescoring results with a more accurate energy function because a native-like pose is not present in the ensemble. *Scoring failures* quantify the inability of the energy function to assign the best score to a correctly sampled (native-like) pose out of the ensemble generated. In such cases, more accurate energy functions could in principle improve the overall success.

Rigid (RGD), Fixed Anchor (FAD), and Flexible (FLX) Docking. Three distinct docking experiments were employed in this study to evaluate different portions of the DOCK sampling algorithm. Rigid docking (RGD) protocols test the ability to rigidly place and optimize the experimental pose back into the binding site through sampling the six degrees of rigid body translation and rotation. Although dihedral angles are not explicitly sampled, ligand torsions are allowed to gently adjust during RGD energy minimizations. Fixed anchor docking (FAD) tests regrowth of the molecule starting from each crystallographic ligand scaffold and progressively samples the torsional degrees of freedom. In a practical sense, FAD can be used to generate a series of overlapped poses for chemically related ligands when the position of a common scaffold is known or can be inferred. Flexible docking (FLX) uses the DOCK anchor-and-grow algorithm,[3] which starts by orienting ligand anchors (scaffolds) into the binding site, followed by on-the-fly flexible conformer growth and minimization, which for many users provides a convenient way to perform virtual screening. Alternative strategies, not pursued here, include pregeneration of conformationally expanded databases (flexibase approach)[34] for use with rigid docking protocols. RGD and FAD protocols employ orthogonal components of sampling, scaffold orientation, and growth respectively, while FLX involves both.

Testset Construction Details. Protein−ligand complexes were extracted from the PDB, separated into individual receptor and ligand files, and saved in MOL2 format using the program MOE[35] for subsequent processing as described below. PDB biological unit files were used during construction to account for binding sites occurring at the interface of protein multimers (i.e., homodimers as in HIV protease). Systems containing covalently bound ligands or those with cofactors, with the exception of monatomic ions, were not included. Figure 2 schematically outlines the workflow and primary software used for construction of the testset.

All ligands were processed with MOE, which was used to assign connectivities, bond orders, atom types, and add hydrogen atoms. Each ligand was processed and inspected manually with every attempt being made to assign ligand protonation states consistent with the original references describing the complex deposited in the PDB. Following visual inspection, empirical Gastieger−Marsili[36] charges were initially assigned using MOE, thus defining the formal charge, and the completed ligand was saved as a MOL2 file. Semiempirical AM1-BCC[37,38] charges were then computed using the AMBER8[39,40] suite of programs using the previously determined formal charge.

All protein receptors were processed with the AMBER8 tleap program, which was used to assign hydrogen atoms, create disulfide linkages, and assign force field parameters. Monoatomic ions were treated as part of the receptor if they were within ca. 10 Å from the binding site. Unless otherwise stated, all water molecules were removed. AMBER8 default protonation were used, resulting in Asp and Glu as negative and Lys and Arg as positive. Histidine residues were treated as neutral with hydrogen atoms added to either the ε or δ nitrogen, depending on the environment, i.e., which nitrogen was coordinated with ions and/or ligands. The prepared receptors were then subjected to a short AMBER energy



grid properties (N=780)

	X (Å)	Y (Å)	Z (Å)	No. spheres
Avg	33.56	33.56	34.15	62
Min	19.18	19.78	20.67	18
Max	48.39	46.55	50.17	75

Figure 3. (a) Prototypical docking setups showing energy grids as gray boxes, docking spheres as red balls, ligand as green surface, and protein backbone as orange tubes. (b) Average, minimum, and maximum grid sizes and number of spheres for the entire testset.

minimization (1000 steps) using a stiff 100 kcal/mol Å^2 restraint on all heavy atoms to allow only the added hydrogen atoms to adjust. It should be noted that crystallographic structures downloaded from the PDB often have incomplete side chains and there may not be enough atoms to unambiguously construct missing side-chain conformations. In several cases, tleap assignments yielded rebuilt side chains having intermolecular clashes. The energy minimization protocol is designed to fail in such situations, and very problematic complexes were removed from the test set. In some instances, side chains with missing electron density were treated as Ala, provided that such residues were distal from the binding site. The resultant AMBER crd and prm files were then used to prepare complete receptor files in MOL2 format with added hydrogen atoms and partial atomic charges and in PDB format without hydrogen atoms for subsequent binding site preparation calculations described below. An advantage of the current protocol is that the test set can easily be setup for molecular dynamics simulations (MD). Future work is planned to use MD results in an effort to derive more accurate DOCK scoring functions.

Binding Site Preparation Steps. A three-step procedure (see Figure 2) was used to prepare receptor binding sites for DOCK calculations as previously described.[3,4] Briefly, a molecular surface of the receptor (without hydrogen atoms) is computed using the program DMS (step one).[41] The program SPHGEN is then used to generate a set of "spheres", located at regions of high inner curvature on the molecular surface, where ligand atoms could potentially interact favorably with the receptor (step two).[42] Spheres are used to guide ligand placement during RGD and FLX docking. Finally, the accessory program GRID is used to precompute van der Waals and Coulombic energy grids, storing at each grid point the intermolecular energies between a dummy probe placed at each point and all receptor atoms (step three).[43] Docking grids are primarily used to speed up the calculations. Input files for the DMS, SPHGEN, and GRID preparatory steps are provided as part of the testset distribution.

To define each binding site, spheres were retained within ca. 8 Å of ligand heavy atoms in the crystallographic pose, up to a maximum of 75 (closest first), which provided a reasonable number of points for RGD and FLX orienting routines. As a baseline, energy grids employed a 8 Å margin size, a 0.3 Å grid spacing, a 4*r* distance-dependent dielectric constant, and 6−9 van der Waals exponents. The protocols described above yield on average 62 docking spheres. To probe the effects of using smaller numbers of spheres, test calculations using FLX protocols were also performed in which the sphere cutoff criteria was varied from 2 Å (24.3 average spheres) to 8 Å (65.2 average spheres) in 1 Å increments. Interestingly, under these conditions overall success varied by only about 2%, which is an indication of robust sampling. We elected to retain a larger sphere protocol, which in principle should be important for screening applications in which greater coverage of a targeted binding site would be desirable. For comparison, prior DOCK studies have employed sphere sets ranging from 100 to 130 spheres.[4,5] Additional testing to determine optimal protocols is in progress, including examination of parameters that affect docking orienting routines and ligand growth such as grid spacing, grid van der Waals exponents, and partial atomic charge.

Figure 3 graphically depicts setups for HIV protease (1HVR) and HIV reverse transcriptase (1VRT), which are representative, and lists average, minimum, and maximum grid sizes and number of spheres for the entire testset. It is important to note that the 8 Å margin parameter is computed in relationship to all retained docking spheres, this results in relatively large grids (average size 34 × 34 × 34 Å^3) that generously encompass each binding site. In addition, grid generation protocols use essentially an infinite cutoff (999 Å); thus, all protein atoms are included at every grid point and each docked pose interacts with the entire receptor.

DOCK Codebase and Input Parameters. All docking calculations in this paper employed a modified version of DOCK available to registered users as DOCK6.4 through the official UCSF Web site (http://dock.compbio.ucsf.edu). Briefly, major changes include (1) code restructuring and modifications to save ligand growth trees, (2) an enhanced ligand internal energy function, and (3) implementation of an rmsd-based harmonic restraint for minimization. The growth tree feature allows the user to visualize all stages of ligand growth, before and after each sampling/minimization step, from the initial anchor placement to the final scored pose. The modified internal energy functions include a repulsive-only term for minimization, pruning, and clustering that significantly improves sampling, reduces runtime, and most importantly effectively eliminates ligand intermolecular clashes that could occur under certain circumstances with earlier DOCK versions. The rmsd-based restraint allows the user to perform energy minimizations in which a pose can be tethered to a reference. A paper describing code improvements is in preparation.

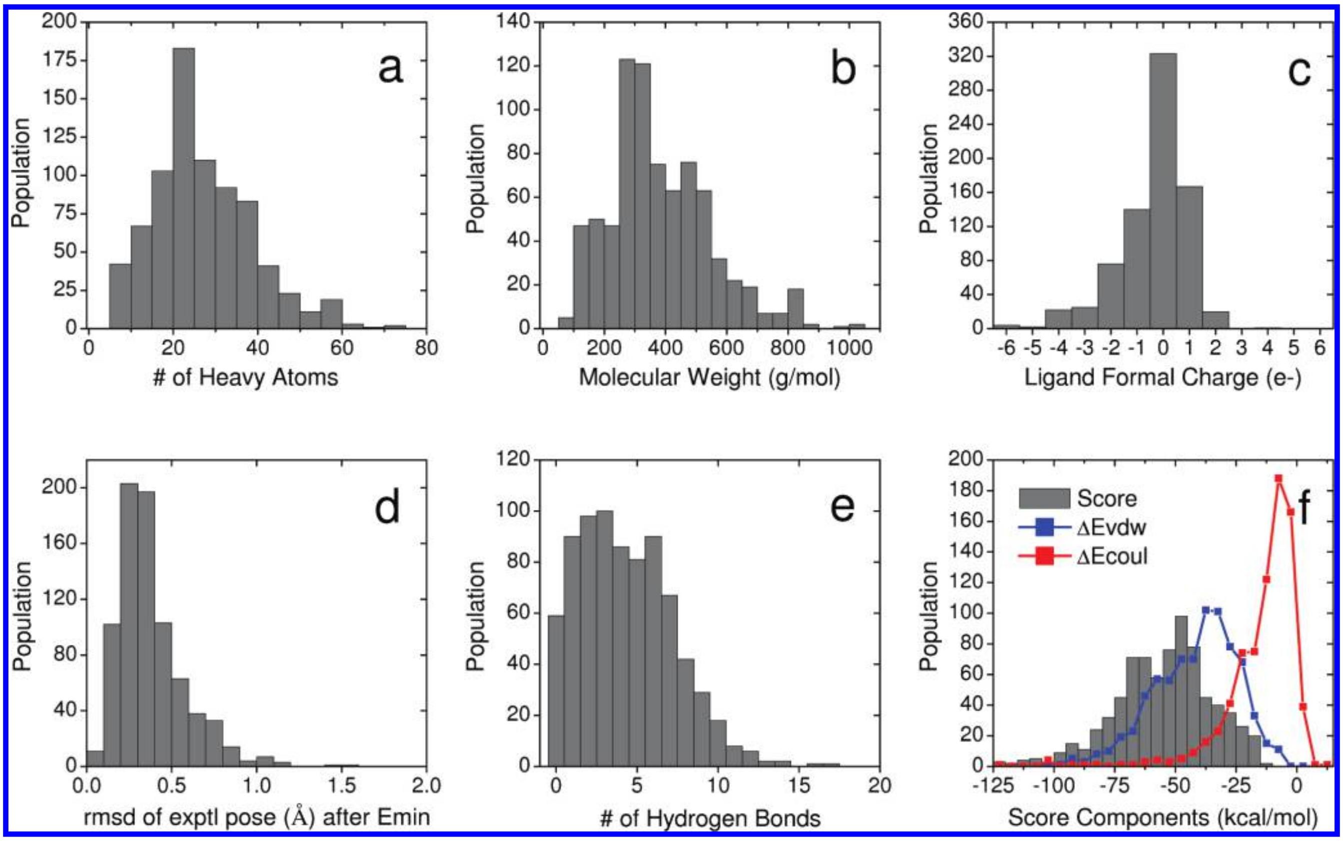


Figure 4. Properties of the SB2010 docking testset ($N = 780$): (a) number of ligand heavy atoms, (b) ligand molecular weight, (c) ligand formal charge, (d) rmsd of ligand experimental pose after energy minimization, (e) number of intermolecular hydrogen bonds after minimization, (f) DOCK intermolecular energy score (gray), van der Waals energy components (ΔE_{vdw}, blue), and Coulombic energy components (ΔE_{coul}, red) after minimization.

Complete input files for the RGD, FAD, and FLX protocols used in this work are provided as part of the testset distribution. Key sampling parameters include 1000 orientations (max_orientations) for RGD and FLX protocols. During simplex minimization, 1000 iterations were used for RGD, while 500 iterations each for anchor orienting (simplex_anchor_max_iterations) and ligand segment growth (simplex_grow_max_iterations) were used for FAD and FLX. Multiple anchor fragments were enabled, with a minimum anchor size of five heavy atoms and a maximum of 1000 anchor orients (pruning_max_orients). For pruning partially grown conformers at every step during anchor and grow, an energy cutoff of +100.0 kcal/mol (pruning_conformer_score_cutoff), with a target population of 100 conformers (pruning_clustering_cutoff), was used. A van der Waals repulsive-only ligand internal energy function with an exponent of 12 was used in all calculations. Final pose ensembles are composed of representative favorably scored members clustered by similarity (2 Å rmsd) and ranked by energy score.

Cross-Docking Setup and Methods. Structural alignment of SB2010 entries was additionally performed to assess docking sensitivity for placing all the ligands into all the receptors in common for a given protein family (termed cross-docking). Members were aligned to a master protein by minimizing differences in C-α positions using the matchmaker tool in the program Chimera.[44] Care was taken to ensure alignments yielded reasonably low pairwise rmsds to the master reference frame (typically 0.2−0.9 Å depending on the family) and that a high number of backbone α carbon atoms were used in the match. Entries with less than ~60% C-α matches or large rmsds to the reference were excluded from the cross-docking families. For each aligned protein, the same transformation matrix was applied to the ligand, which yielded a set of bound complexes in the same coordinate frame. The same binding site preparation steps for standard docking consisting of molecular surface generation (DMS), spheres generation (SPHGEN), and docking grids (GRID) were also applied to cross-docking. Select family alignments are provided as part of the testset distribution.

RESULTS AND DISCUSSION

SB2010 Testset Properties. Figure 4 plots properties of interest, including the number of ligand heavy atoms, ligand molecular weight, ligand formal charge, ligand root-mean-square-deviation (rmsd) after energy minimization, number of intermolecular hydrogen bonds, and DOCK intermolecular score, which is the sum of van der Waals (ΔE_{vdw}) and Coulombic (ΔE_{coul}) energy score components. Most molecules in Figure 4 have 15−25 heavy atoms, Lipinski-like molecular weights (MW ≤ 500 g/mol, $N = 608$), and formal charges in the range −1 to +1 ($N = 631$). Although the majority of ligands have 7-or-less rotatable bonds ($N = 423$), a substantial number of ligands with medium (#RB = 8-to-15, $N = 268$) and high (#RB = 15-plus, $N = 89$) flexibility are also included (Figure 1a). While this may reduce overall success rates, it results in a challenging testset more useful as an evaluation tool. Importantly, no systems were removed from SB2010 because ligands failed to dock correctly. Instead, the set was deliberately populated with entries where docking was observed to be problematic.

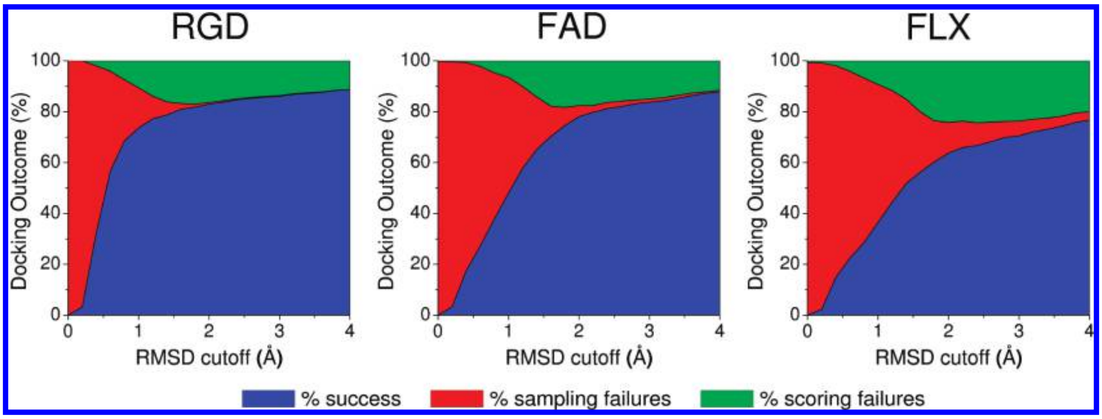


Figure 5. Docking outcomes using SB2010 ($N = 780$) as a function of rmsd from 0 to 4.0 Å for rigid (RGD, left), fixed anchor (FAD, middle), and flexible (FLX, right) ligand docking. The outcomes are docking success rates (blue), sampling failures (red), and scoring failures (green); the vertical width of each solid color defines the percentage of each outcome at a given rmsd value. See Computational Methods for protocol definitions.

Intermolecular properties in Figure 4 include rmsd of the crystallographic ligand pose after an energy minimization, number of hydrogen bonds (#HB), overall DOCK energy score, and energy component breakdowns. Low rmsds after an energy minimization (Figure 4d) help to confirm that both the experimental structures and subsequent computational models with added hydrogen atoms and force field parameters are reasonable starting points for docking. Minimizations using a tether-based restraint of 10 kcal/mol Å^2 that resulted in larger than expected movement (>2 Å rmsd) or unfavorable (positive) scores indicate potential experimental or computational setup errors and were not included. The majority of ligands make one to six intermolecular H-bonds (Figure 4e); however, there is a wide range from 0 to 18. Most DOCK energy scores are in the range from −40 to −50 kcal/mol (Figure 4f), which reflects the large number of neutral small organic molecules in the testset. Neutral ligand scores tend to be dominated by van der Waals (ΔE_{vdw}) energy given that the Coulombic components (ΔE_{coul}) are scaled by a distance dependent dielectric constant ($\varepsilon = 4r$) to crudely mimic solvent screening. For charged molecules, however, scores can be dominated by ΔE_{coul}.

Global Docking Success. Overall docking results, as a function of reproducing experimental poses (from 0 to 4 Å rmsd), are plotted in Figure 5, which shows success in blue, sampling failures in red, and scoring failures in green for the three different RGD, FAD, and FLX protocols (see methods for definitions). For each individual system, success rates are computed using the best-scoring pose found, while sampling and scoring failures statistics are derived using the ensemble of docked cluster heads. The sum of the three metrics at any defined rmsd value equals 100%, which provides a convenient way to assess success in relationship to the underlying cause of docking failures. Sampling and scoring failures should ideally be near 0%. Success (Figure 5, blue) is probably the most useful day-to-day metric, it implicitly includes all sampling and scoring failures, and deviation from 100% provides a quantitative way for users to measure potential docking accuracy.

Changing the rmsd criteria from perfect pose overlap (0 Å) to more realistic values (2 Å) shows, in particular for RGD and FAD protocols, a relatively steep initial increase in overall success (Figure 5, blue) followed by an inflection region after which successes begin to plateau. Compared with the other protocols, RGD yields a particularly sharp knee point at around 1.0 Å. In contrast, the FLX success curve rises much less steeply and the inflection region is less sharp. As expected, all protocols yield initially high sampling failures (Figure 5, red) at very tight rmsds; however, these quickly drop as tolerance for perfectly generating the crystal pose is loosened. By about 1.4 Å for RGD, 2.0 Å for FAD, and 3.0 Å for FLX, the failures in sampling are minimal (4−5%). Failures in scoring show a different behavior. At very low rmsds, scoring failures are essentially nonexistent (Figure 5, green), suggesting that the standard DOCK scoring function is highly accurate, provided sampling algorithms are able to generate a pose that overlaps closely with the experimental pose (presumed global minimum). Conversely, as the rmsd cutoffs increase, the conformational space available for generating acceptably correct poses also increases, which makes ranking more difficult due to potentially larger, more diverse ensembles. Interestingly, maximum errors in scoring appear at ca. 1.5−2.0 Å rmsd (Figure 5, green) for all three protocols. Moreover, the eventual plateau of scoring failure curves beginning in this region highlights the greater need for improved scoring functions relative to sampling across the entire range. Overall, the general shape for the intersection of the three curves (Figure 5, green, red, blue) suggests that a 2.0 Å rmsd definition for docking success is reasonable and, unless otherwise stated, this criteria is used throughout this work.

Alternative Ligand Geometries. To probe possible starting condition effects, additional docking calculations were performed using energy-minimized ligand geometries, optimized without the protein, using one of six methods as shown in Table 1. Optimizations employed the General Amber Force Field (GAFF)[45] as implemented in AMBER8 using 100, 1000, or 10,000 steps of minimization ($\varepsilon = 4r$ dielectric), the Merck Force Field (MMFF94x)[46,47] as implemented in MOE using default protocols, and AM1[48] and PM3MM[49] semiempirical methods as implemented in Gaussian98[50] using max 10,000 cycles of minimization. Ligand geometries optimized using MMFF94x with MOE are provided as part of the testset distribution.

As expected, docking success rates decrease when using energy-minimized ligand geometries as input versus the crystallographic structures (Table 1). The most likely reason involves the fact that on-the-fly DOCK algorithms do not

Table 1. Effect of Initial Ligand Geometry on Rigid (RGD), Fixed Anchor (FAD), and Flexible (FLX) Docking

	ligand rmsd: Emin vs crystal[b]					
starting geometry[a]	avg (Å)	min (Å)	max (Å)	RGD success (%)[c]	FAD success (%)	FLX success (%)
Crystal				83.9	79.0	64.9
GAFF 100	0.23	0.01	2.05	78.9	76.0	62.5
GAFF 1000	0.64	0.01	2.11	73.6	74.3	58.2
GAFF 10,000	0.79	0.01	4.45	69.0	72.0	57.9
MMFF94x	0.78	0.02	3.85	66.5	71.5	57.1
AM1	1.09	0.02	6.36	54.8	70.0	55.8
PM3MM	1.01	0.03	6.36	57.8	71.0	55.5

[a] Crystallographic ligand geometries versus those from energy minimizations using the General Amber Force Field (GAFF) for 100, 1000, or 10,000 steps, the Merck Force Field (MMFF94x), or AM1 and PM3MM semiempirical methods. [b] Average, minimum, and maximum root-mean-square-deviations (Å) for energy minimized ligands relative to original crystal structures before docking. [c] Success statistics are for the intersection of systems ($N = 697$) for which force field parameters could be assigned automatically and energy minimizations were completed successfully for all of the methods.

currently sample dihedral angles within ring systems, bond angles, or bond lengths. Thus, as average rmsds and min/max ranges for the ligands used as docking input increase relative to the original crystallographic geometries, docking success rates generally decrease (Table 1). On the other hand, even short optimizations will yield modified bond lengths, bond angles, and dihedral angles; thus, the small 2−5% decrease using input geometries obtained from the GAFF 100 step protocol is notable. Unsurprisingly for RGD protocols, given that dihedral angles are not sampled during docking, the results in Table 1 show the widest range of variation depending on which method was used to generate the input (5−30% decrease). In contrast, FAD (3−9% decrease) and FLX (2−9% decrease) protocols appear much less sensitive, as expected. An interesting observation is that GAFF (10, 000 steps), MMFF94x, AM1, and PM3MM results are all very similar for FAD (70.0−72.0%) and FLX (55.5−57.9%), despite the fact that four different models (GAFF, MMFF94x, AM1, PM3MM) and three different modeling programs (Amber8, MOE, Gaussian98) were used to generate alternative structures. While these results might be indicative of typical FAD and FLX success rates for DOCK6.4, regardless of the source of ligand geometry, additional testing to assess the role of initial starting conditions is required. For this study, given the potential bias in choosing any single procedure to generate alternative input structures, the original crystallographic geometries were employed in all analysis and discussion that follow.

Results by Ligand Flexibility. Table 2 and Figure 6 show a detailed breakdown of the global results from Figure 5 for subsets of ligands with 7-or-less (low), 8-to-15 (middle), or 15-plus (high) number of rotatable bonds (#RB). Table 2 also contains raw numerical values (no. of molecules) and the average docking run times (minutes/molecule). As a result of sampling and optimization of torsions during growth, compared to RGD docking times which are linear, FAD and FLX timings exhibit the expected exponential increase (Figure 6b) going from rigid to more flexible subsets. Overall, for virtual screening on this particular platform, a maximum of up to 15 rotatable bonds per ligand represents a reasonable compromise between docking speed (Figure 6b) and accuracy (Figure 6a) using current FLX protocols.

For RGD, nearly constant success values of 83.2%, 83.9%, and 78.7% are obtained for low, middle, and high flexibility subsets, respectively (Table 2, column E). For all 780 systems, RGD success equals 82.9%. Notably, RGD failures in sampling (0.4−1.1%) are negligible (Table 2, column G), which indicates excellent orienting. Scoring failures for RGD across the subsets yield 16.1%, 15.7%, and 20.2% (Table 2, column I). Overall, RGD docking is the protocol evaluated for which success, sampling failures, and scoring failures remains approximately constant and independent of ligand flexibility (Table 2).

For FAD, as ligand flexibility increases, success rates drop with near perfect linear behavior (Figure 6a) going from 7-or-less (87.5%), 8-to-15 (71.6%), and 15-plus (52.8%) subsets (Table 2, column E). Sampling failures steadily increase across the three subsets (0.9%, 5.2%, 18.0%), although a low overall failure of 4.4% ($N = 780$) indicates conformer growth by torsion sampling is generally successful if the initial scaffold (anchor) is placed accurately (Table 2, column G). Nevertheless, reparameterization of dihedral angle torsion drive files could potentially be used to correct some FAD sampling failures and is currently under investigation.

Table 2. Success Rates for Rigid (RGD), Fixed-Anchor (FAD, and Flexible (FLX) Ligand Docking

			total success[a]		sampling failures[a]		scoring failures[a]		
docking method (A)	#RB (B)	subset size (C)	no. (D)	% (E)	no. (F)	% (G)	no. (H)	% (I)	avg time (min/mol) (J)
RGD	7-or-less	423	352	83.2	3	0.7	68	16.1	2.0
RGD	8-to-15	268	225	83.9	1	0.4	42	15.7	5.8
RGD	15-plus	89	70	78.7	1	1.1	18	20.2	13.6
RGD	all	780	647	82.9	5	0.6	128	16.4	4.6
FAD	7-or-less	423	370	87.5	4	0.9	49	11.6	0.9
FAD	8-to-15	268	192	71.6	14	5.2	62	23.1	11.2
FAD	15-plus	89	47	52.8	16	18.0	26	29.2	67.1
FAD	all	780	609	78.1	34	4.4	137	17.6	11.9
FLX	7-or-less	423	315	74.5	18	4.3	90	21.3	3.5
FLX	8-to-15	268	148	55.2	41	15.3	79	29.5	17.3
FLX	15-plus	89	35	39.3	35	39.3	19	21.3	74.4
FLX	all	780	498	63.8	94	12.1	188	24.1	16.6

[a] Computed using a 2.0 Å rmsd definition of success using experimentally observed ligand poses as reference.

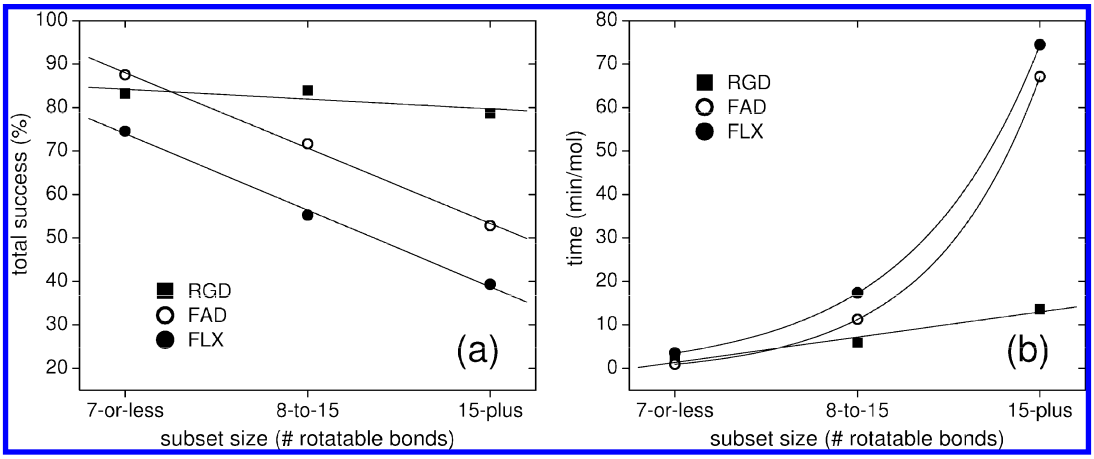


Figure 6. Results by subset flexibility using rigid (RGD), fixed anchor (FAD), and flexible (FLX) protocols for (a) total docking success (%) and (b) average docking time (min/mol) on 3.2 GHz Intel Xeon processors.

Failures due to scoring for FAD (Table 2, column I) increase in going from low (11.6%), to middle (23.1%), to high (29.2%) numbers of rotatable bonds. Interestingly, despite the complexities associated with flexible growth, scoring failures for the FAD 7-or-less subset are less than those for RGD docking by ca. 5%. In general, the high average success rate of 78.1% ($N = 780$) indicates good overall FAD outcomes. This is important, as FAD protocols are expected to be useful for placement of functional groups when the position of a ligand scaffold is already known (i.e., generation of structures for subsequent SAR studies).

For FLX, success performance also degrades linearly (Figure 6a) in going from low to high (74.5%, 55.2%, 39.3%) flexibility subsets, respectively (Table 2, column E). Despite the challenges associated with anchor orientation and on-the-fly growth for larger flexible ligands, the average success rate of 63.8% ($N = 780$) is very encouraging. However, although sampling failures are low for the rigid 7-or-less group (4.3%), poorer statistics for the flexible 8-to-15 (15.3%) and 15-plus (39.3%) groups indicate substantial room for improvement (Table 2, column G). In particular, higher sampling failures for FLX compared with FAD indicate that alternative anchor orientation and pruning protocols should be explored. Although FLX protocols for the 15-plus group yield sampling failures (39.3%) that outweigh scoring (21.3%), for the more rigid 7-or-less and 8-to-15 groups, in general, scoring failures are always dominant. This trend is in general agreement with the global rmsd spectrum plots shown in Figure 5. To identify which systems have particularly poor sampling or scoring and low or high success rates, as described further below, SB2010 members were alternatively grouped into families consisting of related protein entries.

Family-Based Analysis. As shown in Figure 7 and Table 3, a large percentage (73%) of the testset could be grouped into 25 different protein families containing seven or more members ($N = 566$) and is referred to here as "large families". The remaining systems ($N = 214$), which include 25 smaller families containing from two to six members as well as all single unique proteins, were combined into a group termed "small families". Both groups, together with "all systems" ($N = 780$), represent baseline results. A true protein diversity subset of SB2010 would contain 166 complexes. Results in Figure 7 employ the three-color scheme defined earlier (successes blue, sampling failures red, scoring failures green). For FLX results, the differences between sampling and scoring failures are also shown (Figure 7d, orange). The results in Figure 7 and Table 3 are sorted on the basis of decreasing FLX success rates (blue).

Family-Based Successes. Although some variation is observed, in general, family-based successes follow the order reflecting the relative difficulty of the various experiments (RGD > FAD > FLX). Notably, results for the three baseline groups cluster together (large families, small families, all systems). Specific families with particularly high docking success (>80% across all protocols) are sialidase, OMP decarboxylase, HIV RT, neuraminidase, factor Xa, T4 lysozyme, and estrogen receptor. Although all methods generally yield good results, there are some significant outliers (Table 3, columns D, G, J, and Figure 7a−c, blue bars). For example, RGD protocols yield <50% success for egg lysozyme (42.9%). For FAD, < 50% success is obtained for three systems: egg lysozyme (28.6%), carbonic anhydrase (44.8%), and phospholipase A2 (46.7%). For FLX, nine families have <50% success: HIV protease (43.3%), thermolysin (42.3%), acetylcholinesterase (42.1%), matrix metalloprotease (35.7%), carbonic anhydrase (31.0%), egg lysozyme (28.6%), phospholipase A2 (26.7%), carboxypeptidase A (25.0%), and thymidylate synthase (16.7%). Interestingly, in three of the nine families with poor FLX results, failures are dominated by poor sampling (Figure 7d, orange). Further, four of the nine also contain Zn^{2+} in the binding site suggesting a potentially systematic problem with zinc. This latter result is consistent with findings in an earlier DOCK study by Moustakas et al.[4]

Family-Based Sampling. In general, the three protocols yield excellent-to-reasonable sampling across the families (Figure 7, Table 3). RGD sampling failures are trivial, and for FAD, only two families show ≥15%. For FLX, eight families show ≥15% sampling failures. Although greater ligand flexibility can lead to increases in FLX sampling failures (Figure 7c red bars), interestingly, there is not a well-defined or systematic correlation between sampling failure and average number of ligand rotatable bonds (#RB). For example, matrix metalloprotease (#RB = 7.5) and carboxypeptidase A (#RB = 8.1) ligands are more rigid than for many of the systems studied, yet both families stand out as having the highest number of FLX sampling failures (35.7% and 62.5%, respectively). In contrast, the more flexible sialidase (#RB = 10.5) and HMG COA reductase (#RB =

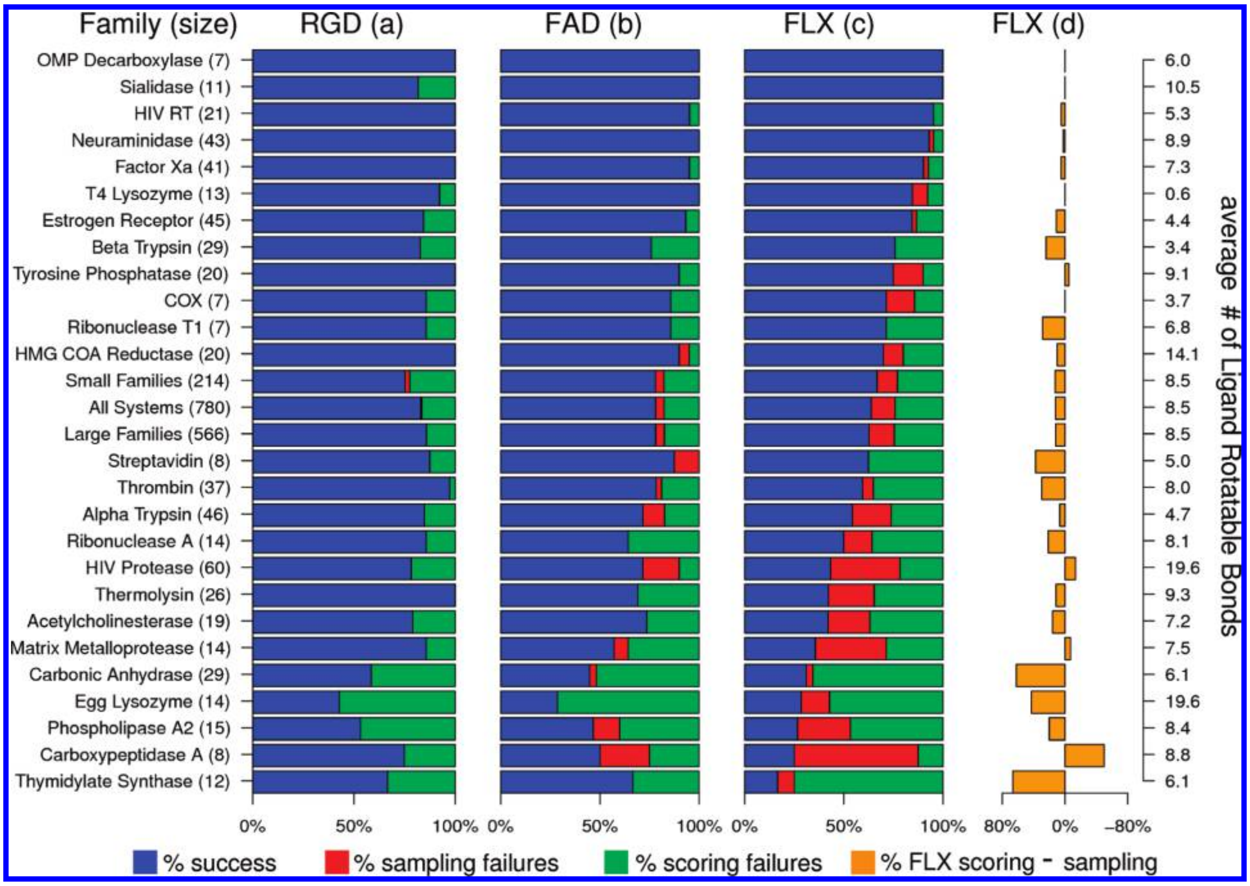


Figure 7. Family-based (a) rigid (RGD), (b) fixed anchor (FAD), and (c) flexible docking (FLX). For each system, the sum of the widths of success (blue), sampling failures (red), and scoring failures (green) is equal to 100%. The left *y*-axis shows protein family and size and the right *y*-axis shows the average number of ligand rotatable bonds (#RB). (d) Orange bars show the difference in scoring − sampling failures for FLX.

Table 3. Family-Based Success Rates for Rigid (RGD), Fixed-Anchor (FAD, and Flexible (FLX) Docking

			RGD (%)			FAD (%)			FLX (%)		
				failures			failures			failures	
protein family	size	avg #RB	success total	sample	score	success total	sample	score	success[a] total	sample	score
(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)
OMP decarboxylase	7	6.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
sialidase	11	10.5	81.8	0.0	18.2	100.0	0.0	0.0	100.0	0.0	0.0
HIV RT	21	5.3	100.0	0.0	0.0	95.2	0.0	4.8	95.2	0.0	4.8
neuraminidase	43	8.9	100.0	0.0	0.0	100.0	0.0	0.0	93.0	2.3	4.7
factor Xa	41	7.3	100.0	0.0	0.0	95.1	0.0	4.9	90.2	2.4	7.3
T4 lysozyme	13	0.6	92.3	0.0	7.7	100.0	0.0	0.0	84.6	7.7	7.7
estrogen receptor	45	4.4	84.4	0.0	15.6	93.3	0.0	6.7	84.4	2.2	13.3
β- trypsin	29	3.4	82.8	0.0	17.2	75.9	0.0	24.1	75.7	0.0	24.1
tyrosine phosphatase	20	9.1	100.0	0.0	0.0	90.0	0.0	10.0	75.0	15.0	10.0
COX	7	3.7	85.7	0.0	14.3	85.7	0.0	14.3	71.4	14.3	14.3
ribonuclease T1	7	6.8	85.7	0.0	14.3	85.7	0.0	14.3	71.4	0.0	28.6
HMG COA reductase	20	14.1	100.0	0.0	0.0	90.0	5.0	5.0	70.0	10.0	20.0
small families	214	8.5	75.2	2.3	22.4	78.0	4.2	17.8	66.8	10.3	22.9
all systems	780	8.5	82.9	0.6	16.4	78.1	4.4	17.8	63.8	12.1	24.1
large families	566	8.5	85.9	0.0	14.1	78.1	4.4	17.5	62.7	12.7	24.6
streptavidin	8	5.0	87.5	0.0	12.5	87.5	12.5	0.0	62.5	0.0	37.5
thrombin	37	8.0	97.3	0.0	2.7	78.4	2.7	18.9	59.5	5.4	35.1
α- trypsin	46	4.7	84.8	0.0	15.2	71.7	10.9	17.4	54.3	19.6	26.1
ribonuclease A	14	8.1	85.7	0.0	14.3	64.3	0.0	35.7	50.0	14.3	35.7
HIV protease	60	19.6	78.3	0.0	21.7	71.7	18.3	10.0	43.3	35.0	21.7
thermolysin[b]	26	9.3	100.0	0.0	0.0	69.2	0.0	30.8	42.3	23.1	34.6
acetylcholinesterase	19	7.2	78.9	0.0	21.1	73.7	0.0	26.3	42.1	21.1	36.8
matrix metalloprotease[b]	14	7.5	85.7	0.0	14.3	57.1	7.1	35.7	35.7	35.7	28.6
carbonic anhydrase[b]	29	6.1	58.6	0.0	41.4	44.8	3.4	51.7	31.0	3.4	65.5
egg lysozyme	14	19.6	42.9	0.0	57.1	28.6	0.0	71.4	28.6	14.3	57.1
phospholipase A2	15	8.4	53.3	0.0	46.7	46.7	13.3	40.0	26.7	26.7	46.7
carboxypeptidase A[b]	8	8.8	75.0	0.0	25.0	50.0	25.0	25.0	25.0	62.5	12.5
thymidylate synthase	12	6.1	66.7	0.0	33.3	66.7	0.0	33.3	16.7	8.3	75.0

[a] Table sorted in decreasing order of FLX success rate. [b] Zinc ion in binding pocket.

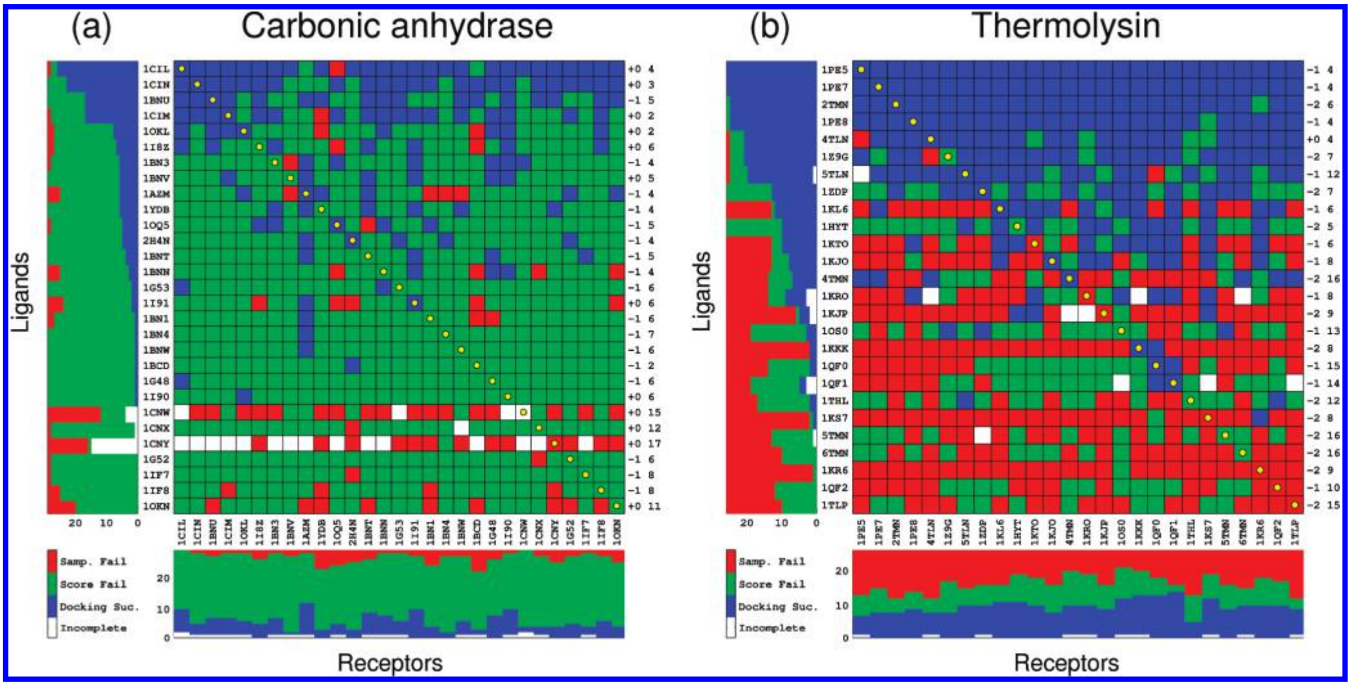


Figure 8. Family-based carbonic anhydrase (a) and thermolysin (b) FLX cross-docking results. Matrix rows and columns correspond to a given ligand or receptor and are identified by PDB codes. Diagonal entries indicate cognate docking (yellow circles). Docking outcomes are classified as sampling failure (red), scoring failure (green), docking success (blue), or incomplete growth (white). Ligand formal charge and number of rotatable bonds are listed on the right *y*-axis. Bottom stacked bar plots indicate outcomes for all ligands with a given receptor. Left stacked bar plots indicate outcomes for any given ligand with all receptors. Matrix elements are sorted in order of decreasing success for each ligand with all receptors (left stacked bar plots).

14.1) groups show low sampling failures of 0% and 10%. Particularly encouraging sampling results are obtained for the highly flexible egg lysozyme family (#RB = 19.6, sampling failures = 14.3%). On the other hand, the HIV protease family with the same average number of rotatable bonds shows 2-fold worse sampling (#RB = 19.6, sampling failures = 35.0%). For the latter case however, a subgroup (N = 11) of HIV protease ligands (1HVR, 1AJV, 1DMP, 1G35, 1HVH, 1HWR, 1MER, 1MES, 1MET, 1PRO, 1QBS) based on a cyclic urea scaffold (avg #RB = 12.5) yield much lower sampling failures (2/11 = 18.2%), although, interestingly, the success is ca. the same (5/11 = 45.4%). To explore if including the well-known flap water in HIV protease would improved sampling, additional FLX calculations were run for the data set excluding the group of cyclic urea-based inhibitors designed to displace the flap water. However, unlike the carbonic anhydrase systems, for which including crystallographic waters did lead to enhanced results (see cross-docking section below), interestingly, no improvement was found here for HIV protease. Finally, an extreme sampling case is for the highly flexible carbohydrate-based ligands contained in the smaller (N = 6) hevamine family (#RB = 26.5). Notably, 0% sampling failures are obtained across all docking protocols for hevamine, and for the FLX results, four of the six members yield correctly docked lowest-energy poses (success = 66.6%) with low rmsd: 1KQZ (0.63 Å), 1KR0 (0.55 Å), 1KR1 (0.65 Å), and 1LLO (0.86 Å). In total, the sampling results overall (Table 2 and Figure 7) indicate that for most families the modified anchor-and-grow sampling routines in DOCK are performing well.

Family-Based Scoring. In agreement with the rmsd spectrum results presented earlier for the testset as a whole (Figure 5), family-based RGD, FAD, and FLX results (Figure 7a−c, green) show that scoring failures overwhelmingly dominate sampling failures. For the FLX results, only four systems show the opposite trend (Figure 7d, orange). In terms of magnitude, the three families with the most significant FLX scoring failures (>50% Figure 7c, green) include thymidylate synthase (75.0%), carbonic anhydrase (65.5%), and egg lyzozyme (57.1%). Interestingly, side-by-side comparisons of FAD vs FLX scoring failures for all the families reveal that in all but the above three cases, fewer failures are obtained using FAD. In a global context this may be partly explained by the fact that FAD protocols sample poses in a smaller region of conformational space closer to the native pose, which leads to more accurate scoring/ranking. This hypothesis is consistent with the conclusions derived from the global rmsd spectrum plots in Figure 5, which revealed that at very low rmsds scoring failures are negligible.

Cross-Docking Matrices (Heatmaps). A natural extension of the family-based analysis involves cross-docking, as illustrated for two representative families (carbonic anhydrase and thermolysin) in Figure 8. In cross-docking, all members of a given family are aligned to a common reference structure and all ligands are docked into all receptors in the common frame. The structural alignments provide a crystallographic pose reference to assess docking success for all possible receptor−ligand combinations. An underlying assumption is that all the ligands theoretically could be accommodated into all the receptors. While for many systems this is reasonable, larger ligands, for example, might not sterically fit into some binding sites. Alternatively, receptor point mutations could preclude some ligands from binding or adopting the same binding pose as the hypothetical reference. Thus, while the diagonal elements in cross-docking matrices (yellow circles) represent experimentally observed cognate ligand−receptor pairs, off diagonal elements represent virtual results and thus should be interpreted with care. Despite these cautions, cross-

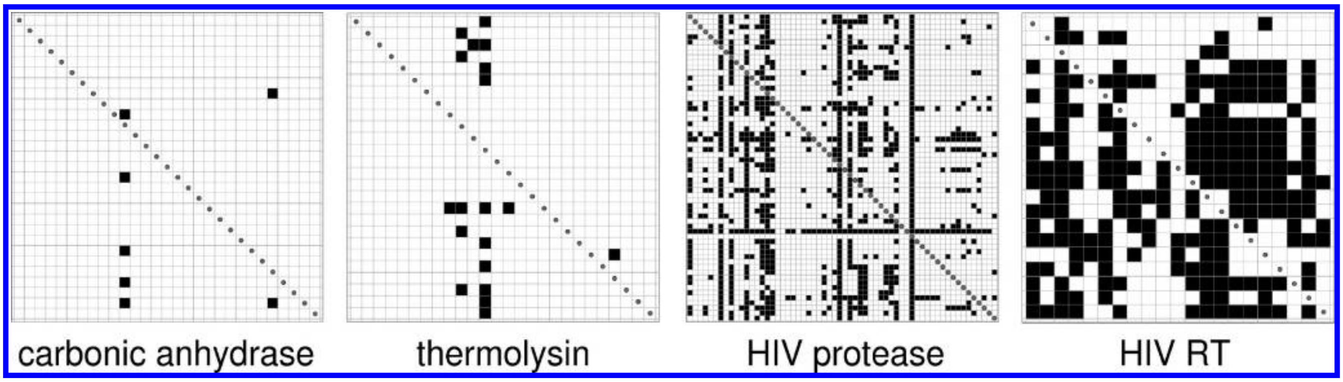


Figure 9. DOCK score matrices from energy minimizations of experimental ligand poses in each receptor color coded by favorable (white) and unfavorable (black) intermolecular energies for carbonic anhydrase, thermolysin, HIV protease (HIV PR), and HIV reverse transcriptase (HIV RT) systems.

docking matrices allow for rapid identification of particularly good or bad docking trends, which will likely be of interest to researchers making specific receptor choices for virtual screening. Matrix colors show success in blue, scoring failures in green, and sampling failures in red.

Figure 8a clearly reveals that carbonic anhydrase has a relatively low matrix success rate (number of blue squares in matrix = 17.7%), despite the fact that nearly complete sampling is achieved (number of red squares = 8.2%). Thus, this family would be a good test system to evaluate alternative procedures that affect scoring (see discussion below). On the other hand, thermolysin shows greater matrix success (36.9%); however, sampling failures in this system are also significantly higher (37.4%). Compared to carbonic anhydrase, thermolysin would provide a good test system for evaluation of alternative docking procedures that would primarily affect sampling.

An alternative way of assessing potential compatibility between ligands and receptors is the evaluation of baseline reference scores after a short energy minimization of each experimental ligand pose in each receptor. Figure 9 shows prototypical matrices for four structurally aligned families in which the elements colored white represent favorable DOCK scores compared with unfavorable scores in black. The calculations here employed a 10 kcal/mol Å^2 harmonic tether to help reduce rmsd differences between minimized and unminimized ligand poses. The central idea is that compatible partners will show favorable scores while complexes with intermolecular clashes will show unfavorable scores. As expected, on the basis of the good sampling for carbonic anhydrase seen in Figure 8 (green elements) the energy-minimized structures for carbonic anhydrase reveal only a few combinations with intermolecular clashes (Figure 9, black elements). Interestingly for thermolysin, there are also relatively few intermolecular clashes, which is surprising given the large number of cross-docking sampling failures seen for this family (Figure 8, large numbers of red elements). Here, an undersampling of torsions, specific to ligands in the thermolysin family, could be involved, and this hypothesis is being investigated. In sharp contrast, results for HIV protease and HIV reverse transcriptase both show a significant number of matrix elements with unfavorable (positive) scores. The presence of unfavorable matrix elements in the structurally aligned reference systems strongly suggests that significant structural rearrangements may be required to achieve compatibility, which, in some instances, will be difficult to achieve in the context of a rigid receptor. Importantly, the protocols outlined here provide a framework for identifying potentially problematic systems and for evaluating the utility of alternative sampling methods which aim to improve cross-docking.

Focusing on carbonic anhydrase, two ligands show systematic poor sampling in all receptors (Figure 8, rows 1CNW, 1CNY) and in some cases no viable poses were generated (Figure 8, white matrix entries). However, in comparison with ligands without failures, these entries are much more flexible (15 and 17 rotatable bonds, Figure 8a, right *y*-axis); thus, the results have a physical explanation. Examination of diagonal results (Figure 10a, black lines), in comparison with the experimental structures (Figure 10, molecular surface), reveals a substantial number of docked poses that occupy space normally containing crystallographic waters (Figure 10b, red spheres). And waters which interact with nearby H64, Y31, and zinc (Figure 10) have been shown to be particularly important for enzyme function.[51] Given the remarkably low number of sampling failures in this system (Figure 8a), a reasonable hypothesis is that the absence of waters contributes to high scoring failures. Specifically, waters likely provide steric hindrance and hydrogen-bonding capability, without which artificially good scores may result due to favorable interactions with H64 and Y31 (Figure 10a). In general agreement with this hypothesis, calculations redone including waters within 10 Å of the binding site zinc ion (Figure 10b) show docked poses overlay more completely with the envelope defined by experiment (Figure 10a vs 10b). Further, both cognate (17% increase) and matrix (14% increase) success rates improved. However, as a cognate success rate of <50% is less than optimal, other scoring options should be explored. Moustakas et al.[4] recently showed that results for zinc-containing systems could be improved by using different Lennard-Jones parameters depending on local coordination states. However, examination of the docked vs experimental poses in Figure 10 shows zinc to be correctly coordinated; thus, other factors (e.g., polarization, desolvation) are more likely to be involved.

As another example of the utility of cross-docking, Table 4 and Figure 11 shows results for the anti-influenza[52] target neuraminidase (43 × 43 matrix). All ligands in this family are based on the same sialic acid scaffold and include the FDA approved compounds oseltamivir (Tamiflu) and zanamivir (Relenza) and the related drug candidate peramivir (emergency use authorization).[53] Several PDB codes among

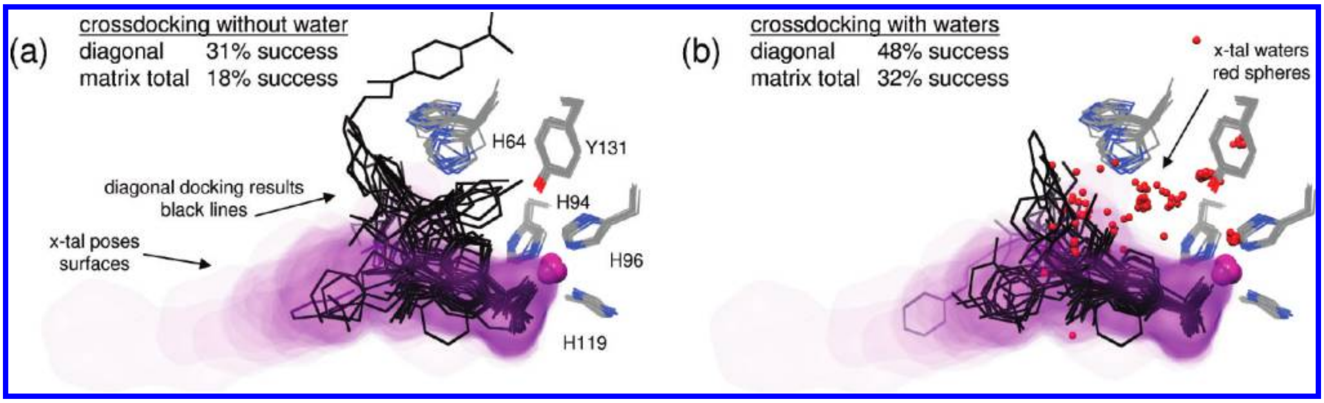
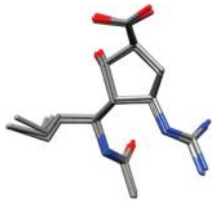


Figure 10. Carbonic anhydrase cross-docking results ($N = 29$ diagonal/cognate structures) from calculations (a) without or (b) with crystallographic waters within 10 Å of zinc ion. Representative side chains are shown as CPK-colored lines, docked poses as black lines, experimental poses as molecular surfaces, and waters as red spheres.

Table 4. Ligand-Based Crossdocking Analysis for Neuraminidase

(A) DANA exptl poses			(B) oseltamivir exptl poses			(C) zanamivir exptl poses			(D) peramivir exptl poses		
											
	No. failures[b]			No. failures			No. failures			No. failures	
lig[a]	samp	score	lig[a]	samp	score	lig[a]	samp	score	lig[a]	samp	score
1NSD	0	1	3CL0	0	0	1NNC	0	2	2HTU	0	0
2HTR	1	0	2HT8	0	1	2HTQ	0	2	1L7G	0	1
2QWC	1	2	2HU0	0	1	3CKZ	1	1	1L7F	0	2
1NNB	0	4	2HU4	0	1	1A4G	1	2	1L7H	0	2
1F8B	1	5	2QWH	0	1						
1IVF	1	6	2HT7	0	2						
2HTW	0	8	2QWK	0	2						
30 failures / 301 possible[c] (N = 7 rows)[d]			8 failures / 301 possible (N = 7 rows)			9 failures / 172 possible (N = 4 rows)			5 failures / 172 possible (N = 4 rows)		

[a] Different crystallographic PDB codes containing the same ligand (DANA = 7, oseltamivir = 7, zanamivir = 4, or peramivir = 4). [b] Number of row-based failures for cross-docking each inhibitor from a given PDB code to all 43 neuraminidase receptors. [c] Sum of all failures out of the total number possible (N rows × 43 receptors).

the group contain the same ligand (i.e., seven entries were cocrystallized with oseltamivir) which permits additional examination of the effects of different ligand starting conditions on docking outcomes. Ideally, all oseltamivir ligands should yield identical results. Table 4 lists row-based failures derived from docking multiple copies of the inhibitors DANA ($N = 7$), oseltamivir ($N = 7$), zanamivir ($N = 4$), and peramivir ($N = 4$) to all 43 neuraminidase receptors (columns A−D). Graphics depicting the experimental poses for each group are also shown. The small structural variation among the references highlights both good receptor alignments (based on C-α atoms) and well-defined experimental poses.

Importantly, results for sampling and scoring failures show minor differences (zero to two failures for each row of 43 receptors) for most entries. Only the DANA group (Table 4, column A) shows larger than expected variation. And in all cases the larger differences arise from scoring and not sampling. Interestingly, the largest changes observed for DANA occur using ligands 1IVF and 2HTW, both of which have the central ligand carboxylate group rotated by ca. 90° in comparison to other crystal poses (Table 4, column A). Although subtle, small internal coordinate differences (bond lengths, angles, dihedrals, ring pucker, etc.) ultimately influence all aspects of the calculations, including partial charge assignments, anchor orientation, functional group sampling, energy minimization, and final rankings. Nevertheless, the remarkably consistent results in Table 4 obtained using multiple ligand copies for neuraminidase are highly encouraging. Studies to characterize docking noise among other families are ongoing.

In comparison with carbonic anhydrase and thermolysin families (Figure 8), neuraminidase shows striking overall matrix success (Figure 11a, blue squares total 87.6%). Birch

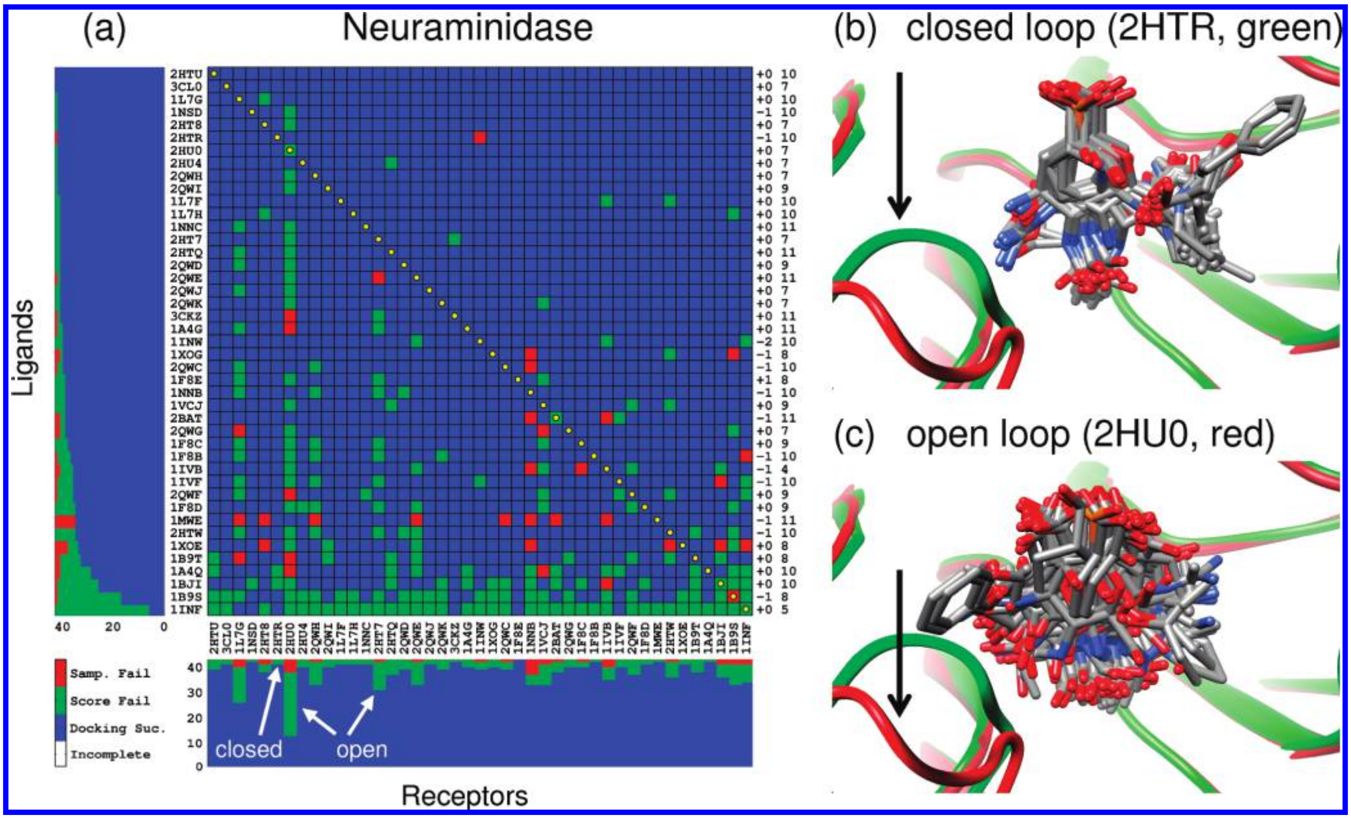


Figure 11. (a) Family-based FLX cross-docking results for neuraminidase. Legend description as in Figure 8. (b) Column-based results for docking all ligands with receptor 2HTR. (c) Column-based results for docking all ligands with receptor 2HU0. Ligands show as CPK colored licorice, neuraminidase shown as green (closed form) or red (open form) colored tubes.

et al.[54] similarly found good results for the neuraminidase systems in a comprehensive cross-docking study using the program GOLD. Against the mostly blue heatmap in Figure 11, several distinct vertical (receptor-based) failure patterns are visible, in particular for receptors 2HU0 (30 failures), 1L7G (17 failures), and 2HT7 (12 failures). Interestingly, 1L7G is the only system studied here that contains the deleterious E119G mutation. For many neuraminidase inhibitors, strong electrostatic interactions between the glutamic acid side chain at position E119 and the positively charged functionality on the ligand lead to enhanced binding.[52,55] Thus, the reduction in accuracy using receptor 1L7G appears to have a physical explanation. Interestingly, the other two receptors with the poorest overall results represent the recently reported open loop[56] form of neuraminidase (2HU0 subtype N1, and 2HT7 subtype N8). For 2HU0, E119 adopts a rotameric state that points away from the ligands; thus, the poor receptor-based results in Figure 11a for 2HU0 (open loop) and 1L7G (E119G mutant) likely share a common origin. In support of this argument, results using 2HT7 show fewer failures, despite having an open loop conformation, likely because the E119 rotamer more closely resembles that seen in closed loop forms.

Finally, Figure 11b,c shows a comparison of results for docking all 43 ligands with 2HU0 (open loop) vs the adjacent entry 2HTR (closed loop), which were both originally cocrystallized with oseltamivir.[56] Structurally, all poses docked into 2HTR show strikingly tight grouping in comparison with the 2HU0 open loop form (Figure 11b vs 11c). Although ligand poses in 2HU0 are mostly ranked incorrectly when compared with the reference poses, given that the sampling failures are in fact negligible, the greater structural variation seen in Figure 11b suggests that the open loop form provides an alternative yet favorable binding environment. This observation is important, as the open loop form of neuraminidase has been proposed as an attractive drug target.[56,57] In total, 26 out of 43 neuraminidase receptors show perfect sampling with all ligands tested, and the best overall docking success is for receptor codes 1NSD (type B), 3CKZ (type N1), 1F8E (type N9), 1F8B (type N9), 1MWE (type N9), and 1A4Q (type B). These results are generally consistent with the study by Birch et al.,[54] in which enhanced results were obtained using 1MWE (100%) and, to a lesser extent, 1F8B (93%), 1F8E (93%), and 1A4Q (93%) structures using a success definition of <2.0 Å. As a general rule, cross-docking matrices should be useful to identify receptors with the lowest number of sampling failures and highest overall success rates for docking known ligands. Depending on the target of interest, specific receptors with particularly good properties across a variety of ligand chemotypes could be considered as well-behaved for the purposes of virtual screening.

CONCLUSION

This work has resulted in construction and refinement of a database called SB2010, consisting of 780 complexes (Figures 1a and 3 and Table S1, Supporting Information) derived from the Protein Data Bank, which can be used to assess accuracy for ligand pose prediction in comparison with experiment. The primary goal is to aid development of robust, improved docking procedures for structure-based drug design and virtual screening. Strengths of the testset include (1) publically available files downloadable in a ready-to-dock format, (2) larger than other comparable hand-curated databases, (3) large numbers of ligands with significant flexibility, (4) ligand protonation states visually compared against available crystallographic references, (5) family-based and ligand flexibility subsets, and (6) family-based cross-docking protocols.

Using three distinct computational experiments, representing rigid (RGD), fixed anchor (FAD), and flexible (FLX) ligand docking, database statistics as a whole were evaluated in terms of success, scoring failures, and sampling failures (Figure 5) with increasing root-mean-square-deviation (rmsd). As rmsd criteria changes from perfect pose overlap to higher values, all three docking protocols show steep increases for overall success (Figure 5, blue) with concomitant decreases for failures due to sampling (Figure 5, red) and increases for failures due to scoring (Figure 5, green). At the commonly employed 2.0 Å rmsd definition (Table 2), docking success tracks with the relative difficulty of the calculation with RGD (82.3%) > FAD (78.1%) > FLX (63.8%). In general, failures in scoring overwhelmingly outweigh failures due to ligand sampling.

To evaluate the effects of ligand flexibility, the database was arranged into three groups consisting of 7-or-less ($N = 423$), 8−15 ($N = 268$), and 15-plus ($N = 89$) number of rotatable bonds. Docking successes (Table 2) degrade linearly across the increasingly flexible subsets using FAD (87.5 > 71.6 > 52.8%) or FLX (74.5 > 55.2 > 39.3%) protocols in contrast to RGD results (83.2−83.9−78.7%), which remain relatively constant. Average docking times (minutes/molecule) show analogous behavior with RGD protocols exhibiting linear behavior compared with FAD and FLX, which show exponential growth (Figure 6).

To identify which systems show particularly high or low docking success rates, SB2010 members were arranged into families ($N = 25$) consisting of related protein entries provided groups contained seven or greater members (Figure 7, Table 3). Family-based successes also generally follow the RGD > FAD > FLX trend. Families with particularly high docking success (>80% across all protocols) include sialidase, OMP decarboxylase, HIV RT, neuraminidase, factor Xa, T4 lysozyme, and estrogen receptor. Families with particularly low docking success (>50% across FLX, FAD protocols) include egg lysozyme, carbonic anhydrase, and phospholipase A2. Three of the nine families with the poorest FLX results show failures are dominated by poor sampling (Figure 7d, orange); however, in general there is no well-defined or systematic correlation between the average number of ligand rotatable bonds (#RB) and sampling failure. Four of the nine contain Zn^{2+} in the binding site, suggesting a potentially systematic problem.

As further examples of the potential utility of the SB2010 database, cross-docking experiments were performed (Figures 8, 10, and 11 and Table 4), for three representative protein families (carbonic anhydrase, thermolysin, neuraminidase), in which all ligands of a given family are docked into all receptors. The resultant carbonic anhydrase heatmap shows relatively low matrix success, despite the fact that nearly complete sampling is achieved (Figure 8a, green). Additional calculations suggest high scoring failures in this system are due in part to the absence of key binding site waters (Figure 10). In contrast, results for thermolysin show significantly more failures as a result of sampling (Figure 8b, red). For neuraminidase, a remarkably high matrix success is observed (Figure 11a, blue). In general agreement with earlier studies,[54] the identification of specific heatmap patterns (Figure 11a) for neuraminidase receptors containing either an open loop form or an E119G mutant demonstrate how cross-docking matrices can be used to gauge which receptor(s) might be most appropriate for virtual screening. The similarity in cross-docking results obtained using multiple copies for four different neuraminidase inhibitors (Table 4) suggests that small variations in starting conditions may minimally impact final docking outcomes. In general support of this observation, experiments using a large subset of the database (Table 1, $N = 697$) revealed a 2−9% decrease in FAD and FLX success rates which was dependent on which energy-minimization protocol and/or force field was used to generate alternative sets of input geometries.

In conclusion, the composition of SB2010 provides a versatile resource for users to address performance across a wide range of docking experiments, including global success/failure, ligand flexibility, family-based analysis, and cross-docking. In conjunction with the database itself, the well-tested protocols will likely be useful for researchers performing a wide variety of real-world docking projects as well as additional methodological studies to characterize the effects of various starting conditions on final docking outcomes.

ACKNOWLEDGMENT

Gratitude is expressed to Dimir Abrar, Ricky Chachra, Rashi Goyal, Patrick Holden, Yulin Huang, Amee Patel, Navamita Ray, and Chetan Rupakheti for assistance with database construction/testing, to Demetri Moustakas and Terry Lang for helpful discussions, and the entire DOCK development team. This research utilized resources at the New York Center for Computational Sciences at Stony Brook University/Brookhaven National Laboratory, which is supported by the U.S. Department of Energy under Contract No. DE-AC02-98CH10886 and by the State of New York. This work was funded in part by the Stony Brook University Office of the Vice President for Research, the New York State Office of Science Technology and Academic Research (NYSTAR), and NIH grants R01GM083669 (to R.C.R.) and F31CA134201 (to T.E.B.).

Supporting Information Available: List of PDB codes used in the construction of the testset (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule−ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(2) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: Applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1–5.

(3) 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.

(4) Moustakas, D. T.; Therese Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Broojimans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601–619.

(5) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. DOCK 6: Combining techniques to model RNA−small molecule complexes. *RNA* **2009**, *15*, 1219–1230.

(6) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228–241.

(7) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76–90.

(8) 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.

(9) 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.

(10) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein−ligand docking using GOLD. *Proteins: Struct. Funct. Genet.* **2003**, *52*, 609–623.

(11) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078–1082.

(12) Hardy, L. W.; Malikayil, A. The impact of structure-guided drug design on clinical agents. *Curr. Drug Discov.* **2003**, 15–20.

(13) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.

(14) Klebe, G. Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580–594.

(15) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(16) Jain, A. Bias, reporting, and sharing: Computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201–212.

(17) Irwin, J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.

(18) Jain, A.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.

(19) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(20) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.

(21) Irwin, J. J.; Shoichet, B. K.; Mysinger, M. M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Automated docking screens: A feasibility study. *J. Med. Chem.* **2009**, *52*, 5712–5720.

(22) Hu, L. G.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins: Struct. Funct. Bioinf.* **2005**, *60*, 333–340.

(23) 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.

(24) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S.; Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(25) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(26) Roche, O.; Kiyama, R.; Brooks, C. L. Ligand−protein database: Linking protein−ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592–3598.

(27) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424–440.

(28) Sandak, B.; Wolfson, H. J.; Nussinov, R. Flexible docking allowing induced fit in proteins: Insights from an open to closed conformational isomers. *Proteins: Struct. Funct. Genet.* **1998**, *32*, 159–174.

(29) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209–225.

(30) 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.

(31) Moitessier, N.; Therrien, E.; Hanessian, S. A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic beta-secretase (BACE 1) inhibitors. *J. Med. Chem.* **2006**, *49*, 5885–5894.

(32) Amaro, R. E.; Baron, R.; McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 693–705.

(33) Ferrari, A. M.; Wei, B. Q. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076–5084.

(34) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: A way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 565–582.

(35) *MOE*; Chemical Computing Group: Montreal, Canada, 2008.

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

(37) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

(38) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(39) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R., Jr.; M. K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(40) *Amber8*; University of California: San Francisco, CA, 2004.

(41) *DMS*; UCSF Computer Graphics Laboratory: San Francisco, CA, http://www.cgl.ucsf.edu/Overview/software.html (accessed May 04th2010).

(42) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **1988**, *31*, 722–729.

(43) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505–524.

(44) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera−A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

(45) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(46) Halgren, T. A. Merck molecular force field 0.1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(47) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.* **1999**, *20*, 720–729.

(48) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The development and use of quantum-mechanical molecular-models. 76. Am1−A new general-purpose quantum-mechanical molecular-model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(49) Stewart, J. J. P. Optimization of parameters for semiempirical methods. 1. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.

(50) *Gaussian 98*, revision A.9; Gaussian Inc.: Pittsburgh PA, 1998.

(51) Hakansson, K.; Carlsson, M.; Svensson, L. A.; Liljas, A. Structure of native and apo carbonic anhydrase II and structure of some of its anion−ligand complexes. *J. Mol. Biol.* **1992**, *227*, 1192–1204.

(52) von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W.; et al. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418–423.

(53) Drugs@FDA website. http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm (accessed May 18, 2010).

(54) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 855–869.

(55) Chachra, R.; Rizzo, R. C. Origins of resistance conferred by the R292K neuraminidase mutation via molecular dynamics and free energy calculations. *J. Chem. Theory Comput.* **2008**, *4*, 1526–1540.

(56) Russell, R. J.; Haire, L. F.; Stevens, D. J.; Collins, P. J.; Lin, Y. P.; Blackburn, G. M.; Hay, A. J.; Gamblin, S. J.; Skehel, J. J. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **2006**, *443*, 45–49.

(57) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878–3894.

CI1001982