J Comput Aided Mol Des (2007) 21:281–306
DOI 10.1007/s10822-007-9114-2

Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search

Ajay N. Jain

Received: 16 January 2007 / Accepted: 21 February 2007 / Published online: 27 March 2007

Abstract The Surflex flexible molecular docking method has been generalized and extended in two primary areas related to the search component of docking. First, incorporation of a small-molecule force-field extends the search into Cartesian coordinates constrained by internal ligand energetics. Whereas previous versions searched only the alignment and acyclic torsional space of the ligand, the new approach supports dynamic ring flexibility and all-atom optimization of docked ligand poses. Second, knowledge of well established molecular interactions between ligand fragments and a target protein can be directly exploited to guide the search process. This offers advantages in some cases over the search strategy where ligand alignment is guided solely by a ''protomol'' (a pre-computed molecular representation of an idealized ligand). Results are presented on both docking accuracy and screening utility using multiple publicly available benchmark data sets that place Surflex's performance in the context of other molecular docking methods. In terms of docking accuracy, Surflex-Dock 2.1 performs as well as the best available methods. In the area of screening utility, Surflex's performance is extremely robust, and it is clearly superior to other methods within the set of cases for which comparative data are available, with roughly double the screening enrichment performance.

Keywords Virtual screening · Enrichment · rmsd · Force field · Flexibility · Scoring function

A. N. Jain (✉)
UCSF Cancer Research Institute, Department of Biopharmaceutical Sciences, and Department of Laboratory Medicine, University of California, Box 0128, San Francisco, CA 94143-0128, USA
e-mail: ajain@jainlab.org

Introduction

Discovery of novel lead compounds through computational exploitation of experimentally determined protein structures, either derived from screening of databases or through focused design exercises, is well established [1], and methodological development within the docking field remains an active area of investigation for a large number of research groups. Many docking methods have been described, and they vary in their approaches to two components: scoring functions and search methods [2–15]. The searching and scoring problems are intimately tied together for two reasons. First, many search strategies make direct use of their scoring functions even deep within the search space in order to prune poor partial solutions. Second, the construction of a search strategy may make implicit assumptions about the space of ligand configurations. For example, a docker may choose to vary only the six translational and rotational parameters of alignment along with the dihedral angles of acyclic rotatable bonds (collectively called the *pose* parameters). Implicit in this treatment is the idea that the space covered by varying the pose parameters alone includes a high-scoring ligand pose that is close to correct. Earlier versions of Surflex-Dock were restricted purely to exploration of this pose parameter space [16, 17]. The focus of this work is in moving beyond the limitations of exploring the pure pose parameter search space and, when appropriate, in exploring it with the benefit of prior knowledge of a protein ligand interaction.

This paper presents three methodological enhancements to the search process:

1. Ligand energetic modeling: Limited to pure pose parameter variation, a docker may be unable to recover from poor initial ligand coordinates that result in high

strain. Also possible are cases where slight changes in a ligand that are achievable only through bending and flexing the structure are required to take a nearly correct but low-scoring ligand pose to a correct and high-scoring ligand configuration. Implementation of a small-molecule force field within Surflex-Dock 2.1 increases its robustness, particularly with respect to screening effectiveness.

2. Ring flexibility: With the internal implementation of a force field, a general method to dynamically search ring systems became straightforward. The method is both fast and general, and is not limited to a specific set of pre-computed ring structures. This feature is particularly useful in terms of enhancing docking accuracy, though it enhances screening utility in cases where the active ligands have flexible rings.
3. Knowledge-based docking: In practical use, it is very commonly the case that a ligand design exercise involves synthetic analogs of a compound whose bound structure is known. In such cases, a particular interaction may be well understood: for example a hinge interaction in a kinase, a metal chelation geometry in a metalloenzyme, or a P_1 specificity pocket interaction in a serine protease. Both for efficiency in workflow and for direct comparison of different analogs, specification of the position of a molecular subfragment can be advantageous. Surflex-Dock 2.1 implements this feature in a way that allows sensitive control of the use of the placed molecular fragment.

The practical appeal of these algorithmic enhancements will be generally evident to those who make use of docking tools since each of these algorithmic enhancements, on first-principles arguments, *should* result in improved performance. This paper establishes the quantitative results of applying these procedures with respect to both docking accuracy and screening utility. The former measures the proportion of cases where a docking algorithm is able to identify the correct bound configuration of a ligand from a starting pose unrelated to the bound one, either among the top set of returned poses or as the top scoring pose. The latter measures the ability of a docking algorithm, given a single protein structure, to correctly rank a set of known active ligands against a background of putative inactives (called decoys). A recent and valuable trend within the field has been the use of standard benchmarks on multiple methods [16–22]. For the work reported here, benchmarks were selected either for which previous versions of Surflex-Dock had been tested [16, 17, 19] in order to show the effects of new features, or where protein and ligand structures were publicly available along with performance of widely used methods [23, 24].

The four publicly available benchmarks addressed either screening utility or docking accuracy:

1. Screening utility, 27 proteins: This set consists of 27 protein structures, each with from 5 to 20 known ligands along with two different sets of decoy molecules. The set incorporates the widely used estrogen receptor and thymidine kinase test cases of Rognan’s group [13, 16–19]. This will be called the Pham set.
2. Docking accuracy, 81 complexes: This set consists of 81 protein/ligand complexes forming a subset of the 134 from the report of Jones et al. [11] and also used in the original report of Surflex-Dock [16]. This will be called the Jain set.
3. Screening utility, 4 proteins: Cummings et al. reported on the performance of four docking systems against four targets of pharmaceutical interest: HIV protease, thrombin (THR), protein-tyrosine-phosphatase 1b, and HDM2 (also called MDM2, a protein that binds to and inactivates the tumor suppressor p53) [24]. This set of four proteins, known actives, and a decoy set derived from the MDL Drug Data Report (MDDR) will be called the J&J set.
4. Docking accuracy, 100 complexes: Perola et al. [23] reported both docking accuracy and screening utility for three different docking systems, but the screening data were proprietary. Half of the docking accuracy data (100 complexes) were publicly available. This will be called the Vertex set.

Extensive validation experiments show that increasing Surflex-Dock’s exploration of the energetically accessible configurational space of ligands increases the robustness of the docking process. Very substantial enhancements in screening utility were made possible by overcoming issues with ligand strain that are common with widely used 3D structure generation techniques and which can arise from the docking process itself. More modest, but potentially important, enhancements in docking accuracy were seen with the explicit use of ring conformation exploration as an augmentation to the docking process.

Surflex-Dock 2.1, employing standard screening parameters on the Pham and J&J sets, obtained mean maximal enrichments of 100-fold and 43-fold, respectively, and mean enrichments at 1% coverage of 37-fold and 15-fold, respectively. Compared with other docking methods on the J&J benchmark, the Surflex screening protocol yielded average enrichments that were twice as good as each of Glide, GOLD, Dock, and DockVision at both the 2 and 5% coverage levels. Surflex-Dock 2.1, with standard parameters for geometric docking, identified excellent poses among the top 20 returned ~85–95% of the time based on the Jain and Vertex benchmarks. Variability was higher in identifying the correct pose as the top scoring, ranging from 50 to 75%. Docking accuracy of Surflex, Glide, and GOLD were comparable on the Vertex set. Surprisingly, small changes in protein proton positions were capable of producing large shifts in performance. This affected the comparison somewhat, but the observation also suggests that local protein optimization may be a viable avenue for improving pose ranking.

The software that implements the algorithms described here is available free of charge to academic researchers for non-commercial use (see http://www.jainlab.org for details on obtaining the software). Molecular data sets presented herein are also available.

Methods

The present study makes use of a number of publicly available data sets to demonstrate improvements, both tangible and operational, in the Surflex-Dock suite of algorithms. The following describes the molecular data sets, computational methods, detailed computational procedures, and quantification of performance.

Molecular data sets

The three primary criteria for evaluating docking strategies are geometric docking accuracy, screening utility, and scoring accuracy. Geometric docking accuracy depends both upon the scoring function of the docker having an extremum in the correct location of the energy landscape as well as the search strategy effectively exhausting the space of reasonable ligand configurations. Utility in screening requires that a docking method's scoring function will have a larger magnitude at (or near) its extremum for each true ligand that is large relative to the extrema for non-ligands. The search requirement is not as stringent; *some* pose whose score is near the extremum must be identified. Scoring accuracy sufficient to produce correct rankings *within* a set of true ligands requires greater accuracy in magnitude estimation. This can be very important in medicinal chemistry exercises, but it is not the focus of the present study.

A number of recent studies have made available public benchmarks for evaluation of docking methods. In this paper, four are used, two each to evaluate aspects of screening utility and docking accuracy. Figure 1 shows representative ligands for each of the four data sets:

J&J screening enrichment set

Cummings et al. [24] reported comparative performance of DockVision, Dock, GOLD, and Glide with respect to their ability to identify known ligands of HIV Protease (HVR), the human homolog of the mouse double minute 2 oncoprotein (HDM2), protein tyrosine phosphatase 1b (PTP1b), THR, and urokinase plasminogen activator (uPA). Of these, the ligands used for the first four were available from the authors. The protein structures correspond to PDB codes 1HVR, 1T4E, 1QBV, and 1C84, respectively. In all but the HDM2 case, the PDB structure was the same as in the original paper [24]. In the HDM2 case, the structure was proprietary at the time of publication, but PDB structure 1T4E was released by the group subsequently. The decoy set (1,000 presumed inactive molecules) was derived from Version 2002.1 of the MDL MDDR, and roughly matched the active ligands with respect to log P and numbers of acceptors and donors. One aspect of the original preparation of the ligand structures was unusual. Structures for the active molecules were generated using CORINA [25], since it preserved the specified chirality but eliminated any ''memory'' of the bound conformations. However, structures were generated by Concord for the 1,000 MDDR decoys.

In making use of this benchmark, with the exception of adding protons to the protein structures (required for Surflex, whereas other programs may require only polar hydrogens), all structures were used unmodified, in order to provide a fair comparison to the reported performance of the other methods.

Pham screening enrichment set

Pham and Jain [17] reported performance of Surflex-Dock Version 1.3 on 29 proteins, each with known ligands ranging in number from 5 to 20. Two decoy sets were used, one being a modified version of that made available by Rognan's group [18, 19], and the other being derived from the ZINC database. Three sources were used to generate the protein test cases. First, the two cases from the comparative paper of Bissantz et al. [18] were used, since they have become a common benchmark. These included protein structures for HSV-1 thymidine kinase (1KIM) and estrogen receptor alpha (3ERT), ten known ligands of TK in arbitrary initial poses, and ten known ligands of ERα in arbitrary initial poses. Second, inhibitors of PARP and PTP were taken from the results of a combination of both virtual and high-throughput screening [26, 27]. Third, the PDBbind database [28] was used to generate a large number of additional cases for testing screening utility. From the full 800 complex set, all proteins were identified that were represented with at least five different ligands. The final set included serine proteases, kinases, phosphatases, isomerases, aspartyl proteases, metalloproteases, nuclear hormone receptors, and a number of other protein types. Importantly, the range of ligand binding affinities was

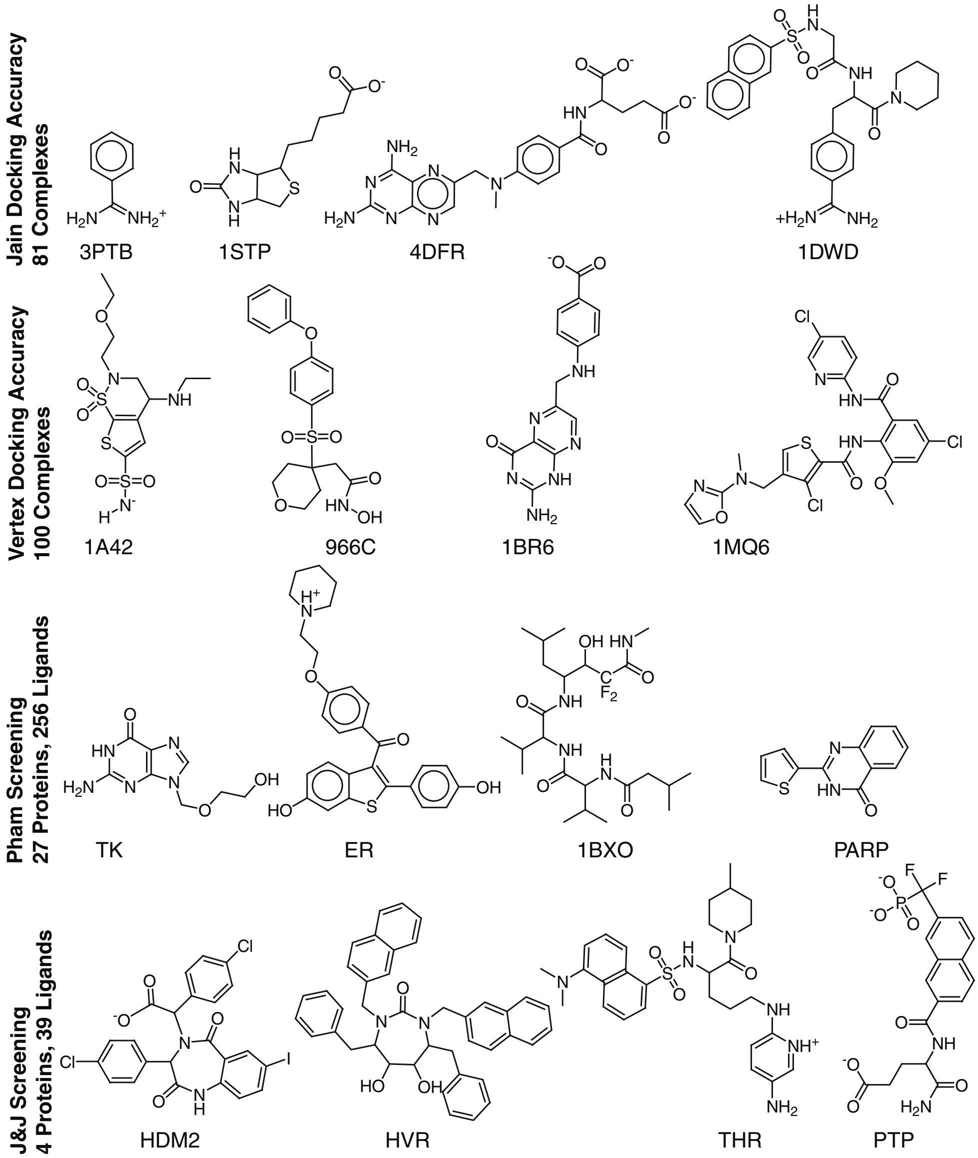

Fig. 1 Four representative ligands from each of the four benchmark data sets. The ligands span a broad range of physicochemical properties and binding affinities

large, with a substantial number of lower affinity ligands. Half of the ligands had pK_d <6.0 (micromolar or worse K_i or K_d), with just one fifth having pK_d >9.0 (subnanomolar or better).

In the present paper, the Rognan and ZINC decoy sets, as well as all active ligands, have been reprocessed by removing all protons, adding protons back using a fully automated procedure, and minimizing the resulting structures (see below for computational details). This was done to address structural inconsistencies in a portion of the molecules that had been introduced by an incorrect protonation procedure. Additionally, two of the 29 proteins have been dropped. The 2AMV case was dropped due to duplicate ligands and an annotation error regarding multiple binding sites. After corrections, there were too few cognate ligands to include this case. The PTP case was dropped since the bulk of the nominal true ligands appear not to be competitive reversible inhibitors of the enzyme. Note that these changes did not significantly affect the reported results from the Pham and Jain report [17] (discussed further below).

The Pham benchmark represents one of the largest sets for evaluation of screening efficiency that is publicly available. The present version (called revision 1) consists of 27 proteins, with 256 cognate ligands, and two decoy sets, the one derived from Rognan's work containing 861 molecules, and the one derived from drug-like ZINC screening molecules containing 1,000 structures.

Vertex docking accuracy set

Perola et al. [23] reported the docking accuracy of Glide, GOLD, and ICM on 200 complexes, 100 of which were available from the authors. This was a detailed study, in which the specific effects of scoring function application (including local optimization) after docking were examined. Complexes were selected where a binding constant was available, where the ligand and protein interaction was non-covalent, and where crystallographic resolution was <3.0 Å. Ligands were selected to have molecular weight between 200 and 600, 1–12 rotatable bonds, be drug/lead-like, and be structurally diverse. Proteins were selected from multiple classes to be relevant for drug discovery.

In making use of this benchmark for this study, all structures were used completely unmodified for direct comparison to the reported performance of the other methods. Further experiments that considered aspects of protein conformational optimization began from the original coordinates provided by the Perola et al. [23].

Jain docking accuracy set

Primarily for comparison with previous Surflex version, the 81 protein ligand complex set from the original Surflex-Dock paper was also used here [16]. It is a subset of the 134 complexes reported by the authors of GOLD [11], and it is described in more detail in the original publications. One change made for this work has been to report performance beginning with a single random pose, in contrast with the previous reports' use of ten random starting poses. This has been done for the sake of congruence with a converging standard in the field.

Computational methods

The core computational methods within Surflex-Dock have been reported in previous papers and will be described only briefly here. Those methods that represent modifications and enhancements will be presented in detail.

Scoring function

There are three approaches for addressing the scoring problem in molecular docking that are in wide use [29]. Two, typically termed physics-based and knowledge-based, share in common a direct grounding in physics. The former constructs functions-based directly on the theoretical physics that underlie molecular mechanics force-fields. The latter make use of knowledge of atomic contact preferences and are related to the statistical physics approach that employs potentials of mean force. Surflex employs one of the so-called empirical methods that take a different approach. The idea is to define a function composed of terms that are related to known physical processes that underlie the physics of protein ligand binding, and estimate the parameters of the function based on protein-ligand complexes of known affinities and structures. The scoring function used in Surflex (and in Hammerhead, which was Surflex's antecedent) borrowed heavily from the approach of Bohm [2, 3, 9, 30]. Bohm's approach had terms for hydrophobic contact, polar interactions, and entropic fixation costs for loss of torsional, translational, and rotational degrees of freedom. The family of empirically constructed scoring functions generally comprise this same basic set of terms, but the details of the underlying functional forms, the data used for parameterization, and the methods for optimization of parameters vary [4, 11, 12, 15, 31–35].

While sharing many aspects, the scoring function used in Hammerhead and Surflex makes a significant departure from other approaches in two important respects [2, 9, 30]. First, the function is composed of a sum of *non-linear* terms and it is continuous and first-order piecewise differentiable. Second, the parameter estimation regime for the function takes direct account of the problem of ligand pose variation. Very small changes in ligand pose can yield large differences in the nominal value of a scoring function. Rather than taking the precise pose from a crystal structure, the approach is to find the nearest local optimum and define the score at that optimum as the score for the ligand. This follows the approach developed for Compass, which established the conceptual framework for this approach, termed *multiple instance learning* within the computational machine learning field [36–39]. The scoring function was tuned to predict the binding affinities of 34 protein/ligand complexes (overlapping significantly with the Bohm training set), with its output being represented in units of -log(K_d) [2]. The range of ligand potencies in the training set ranged from 10^{-3} to 10^{-14} and represented a broad variety of functional classes.

The terms, in rough order of significance, are: hydrophobic complementarity, polar complementarity, entropic terms, and solvation terms. By far the most dominant terms are the hydrophobic contact term and a polar contact term. The polar term has a directional component and is scaled by formal charges on the protein and ligand atoms. These terms are parameterized based on distances between van der Waals surfaces, with negative values indicating interpenetration. Each atom on the protein and ligand is labeled as being non-polar (e.g., the H of a C–H,) or polar (e.g., the H of an N–H or the O of a C=O), and polar atoms are also assigned a formal charge, if present. The key terms that are parameterized by distance are as follows:

$$\mathrm{steric\_score} = l_1 \exp^{-(r+n_1)^2/n_2} + \frac{l_2}{1+\exp^{n_3(r+n_4)}} + l_3 \max\left(0, r+n_5\right)^2, \tag{1}$$

$$\text{polar\_score} = l_4 \exp^{-(r+n_6)^2/n_7} + \frac{l_5}{1+\exp^{n_3(r+n_8)}} + l_3 \max(0, r+n_9)^2. \quad (2)$$

The formal charge of atoms in a pairwise interaction scales the polar scoring term quadratically [2]. Each of the terms is composed of a Gaussian, a sigmoid, and a quadratic term that are dependent on r, which is defined as the atomic center-to-center distance of atom pairs less the sum of their radii (negative distances arise with nominally interpenetrating atomic radii). The first two terms of each equation were parameterized in the original report [2], and the latter was recently added to provide a formally estimated clashing term [30]. The parameters l_1 and l_4 are analogous to the lipophilic and hydrogen-bond weighting terms from other empirical functions. The polar term is also scaled by directionality and by formal charge, if present. Figure 2 shows plots of the hydrophobic term and the polar term for a hydrogen bond. The hydrophobic term (bottom curve, solid line) yields ~0.1 U of pK_d per ideal hydrophobic atom/atom contact. The top curve (dashed line) shows that a perfect hydrogen bond yields about 1.2 U of pK_d and has a peak corresponding to 1.97 Å from the center of a donor proton to the center of an acceptor oxygen.

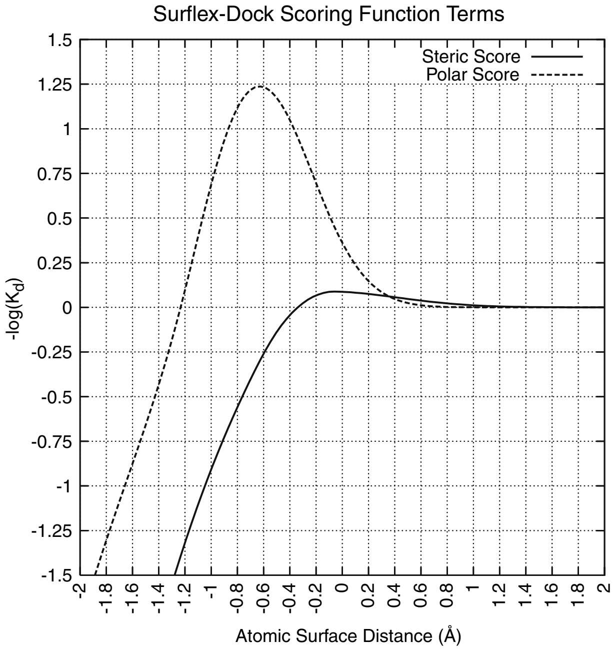

Fig. 2 The primary Surflex-Dock scoring function terms are pairwise atomic interactions between ligand and protein. The functional terms depicted are parameterized in terms of pK_d (*Y*-axis) and surface distances between protein and ligand atoms (*X*-axis). Surface distances are negative where the nominal atomic radii interpenetrate, as with a typical hydrogen-bond when using unscaled radii

As with Bohm's work, the Surflex scoring function includes an entropic penalty term that is linear in the number of rotatable bonds in the ligand, intended to model the entropic cost of fixation of these bonds, and a term that is linearly related to the log of the molecular weight of the ligand, intended to model the loss of translation and rotational entropy of the ligand. The solvation terms have little effect and are the subject of current investigation. For a more detailed discussion of the Surflex scoring function, please refer to the specific reports of the derivation and refinement of the function [2, 30].

Search strategy overview

A detailed account of the Surflex-Dock search algorithm can be found in the original paper [16]. Surflex employs an idealized active site ligand (called a protomol) as a target to generate putative poses of molecules or molecular fragments. Surflex's protomols utilize CH_4, C=O, and N–H molecular fragments. The molecular fragments are tessellated in the protein active site and optimized based on the Surflex scoring function. High-scoring fragments are retained, with redundant fragments being eliminated. The protomol is intended to mimic the ideal interactions made by a perfect ligand to the protein active site that will be the subject of docking.

Surflex utilizes the morphological similarity function and fast pose generation techniques described previously [40] to generate putative alignments of fragments of an input ligand to the protomol. Poses of the molecular fragments that tend to maximize similarity to a protomol are used as input to the scoring function and are subject to thresholds on protein interpenetration and local optimization. The partially optimized poses of the fragments form the basis for further elaboration of the optimal pose of the full input ligand. Since the scoring function is based on atom–atom pairwise interactions, it is possible to generate a score for any fragment of a docked pose. The procedure identifies high-scoring fragments that have compatible geometries to allow for merging in order to assemble a high-scoring pose of the full input ligand. The whole molecules resulting from the merging procedure are pruned based on docking score, and are subjected to further gradient-based score optimization. The procedure returns a fixed number of top scoring poses.

Local optimization: a quasi-newton method

In the original Surflex-Dock approach [16], local optimization of both partial and complete ligands was accomplished using a very simple implementation of gradient-descent, making use of numerically computed gradient information. The procedure was not optimized carefully, nor did it employ stopping criteria that could respond to magnitude changes in either the gradient or the stepwise improvement in ligand score. For Version 2.1, a quasi-Newton method (Broyden–Fletcher–Goldfarb–Shanno, BFGS) [41] was implemented, in addition to implementation of analytical derivatives for the Surflex-Dock scoring function. BFGS is a modified Newton's method for optimization that only requires direct computation of the first derivative of the function to be minimized. It works by computing an approximation to the inverse Hessian of the function, rather than requiring its explicit computation. The new method is both faster and more effective in terms of quality of optimization than was the pure gradient-descent approach of earlier versions.

The field of non-linear function optimization is well-developed, but the specific optimization protocol that is ideal for a particular function and specific landscape characteristics is highly problem-dependent and is also dependent upon the precise implementation of the optimizer. In comparison with the steepest descent approach that the BFGS approach has replaced, convergence is significantly faster, owing primarily to a reduction in the total number of function evaluations (both with and without gradient computation) that are required. Docking speed is proportional to an input ligand's number of rotatable bonds. Under default parameters, the median docking time per rotatable bond using Version 2.1 was 2 s compared with 3 s for Version 1.3 (2.8 GHz Intel Xeon processor under Windows). This represents an improvement of roughly 30%.

Covalent forces: access to the cartesian space

The original versions of Surflex-Dock varied molecular pose only by changing the six parameters of alignment (three translational and three rotational) and the dihedral angles of rotatable bonds. Neither flexibility in rings (except through independent dockings of multiple ring conformers) or more subtle bending and flexing of molecules were possible. Addressing such motions in the general case involves changing bond angles and bond lengths, which requires a force field in order to trade the internal ligand energetics against the interactions between the ligand and protein. The problem of parameterizing a force field to suit a wide variety of organic structures was addressed by Mayo et al. [42] with their DREIDING force field, which employed a limited number of atom types and yields broad coverage of small molecule structures. The tradeoff in employing such a force-field is that with the limitation on atom types comes a limitation on the specificity of force-field terms that can result in non-optimal energy estimation. However, the degree of structure normalization afforded by this approach is beneficial, as will become clear in the Results.

Surflex-Dock 2.1 implements the DREIDING force field's bond angle, bond length, torsional, and apolar non-bonded terms. Neither the inversion terms (to directly characterize planarity constraints) nor the polar non-bonded terms (for intramolecular hydrogen bonds and Coloumbic interactions) were implemented. The former were excluded for simplicity, and the latter since typical molecules for study in docking simulations tend to have few intramolecular electrostatic interactions compared with *inter*-molecular ones. This also avoided a requirement for computation of conventional partial charges. The implementation computes explicit analytical derivatives, and minimization of molecules makes use of the BFGS algorithm in the same manner as optimization for the Surflex-Dock scoring function. Input ligands may be minimized without docking (''surflex-dock **min** prot ligand_archive.mol2 min-file-prefix''), minimized prior to docking (''surflex-dock **-premin** dock_list...''), or have simultaneous optimization of internal energetics along with protein ligand interactions after docking (e.g., with ''surflex-dock **-remin** dock_list...'').

Ring flexibility

There are two particularly straightforward approaches to addressing ring flexibility in docking algorithms. The first is to employ a separate tool to pre-search each ligand to generate potential ring conformations and to then dock each ligand beginning from multiple starting configurations. The second is to pre-compute ring geometries for common ring systems, and then to use a substructure matching approach to generate the appropriate variants of the input ligand during conformational search. Both approaches are perfectly reasonable, but both have some limitations. In the former case, the quality of the results will depend on the interaction between the ring conformation generation process with the docking procedure employed, which may be unknown. In the latter case, it is very difficult to pre-compute ring geometries for all potentially interesting flexible ring systems. For example, in Fig. 1, the benzodiazepinedione (HDM2), the central ring of the Dupont Merck protease inhibitor (HVR), and the substituted sulfonamide-containing ring (1A42) may be well-known now, but may not have been well-enough known at the time of their modeling and synthesis to have been included in standard ring libraries.

Surflex-Dock 2.1 takes a different approach, made possible by the DREIDING implementation. For rings of five, six, and seven members, prototype conformations of cyclopentane, cyclohexane, and cycloheptane are used to produce variants of an input structure. These variants are then subjected to minimization and redundancy elimination. Figure 3 shows the central ring of the HVR inhibitor from Fig. 1 in the conformation that was generated by CORINA in the J&J screening benchmark. The arrows mark a bad clash that is impossible to alleviate through rotation of acyclic bonds but is alleviated using minimization. In the ring searching algorithm, minimization is carried out after making nitrogen inversions (if necessary) and prior to the process of identifying, matching, and instantiating the prototype conformations of flexible rings. The central depiction shows 28 new ring conformations of the input structure, each derived from the alignment of a single conformation of cycloheptane to one of its 14 possible matches to the seven-membered ring. Note that atom types do not participate in the ring matching, so any combination of atoms may be present in the input structure. Following the alignment of a prototype conformation, the relationship between the pendant functionality of the original structure to the new conformation is computed, and the proper transformation is applied such that the newly generated conformation *will not* have a bias to pull the new ring geometry back into the starting configuration. Note in the inset overlay that all of the ligand's pendant functionality waves along with the changes in the ring configuration. Of course, by using saturated hydrocarbons, both the bond lengths and angles of the original ligand are perturbed in this process, necessitating a final minimization. Not that while rare, this type of process can invert chiral centers. For this reason, chiral centers are checked against the original input structure after final minimization and conformations that contain any inversions are eliminated.

Ring searching is selected by the ‘‘**-ring**’’ option, and the behavior may be modified by changing the threshold beyond which high-energy ring conformations are deleted (the default being 5 kcal/mol). For the molecule shown in Fig. 3, the process of moving from the highly strained input conformation to the final set of five final conformations (one of which is shown) takes less than a second on standard desktop hardware. In cases where there are a number of flexible rings within a molecule, necessitating a combinatorial search of joint ring conformations, the process can be significantly longer. However, the dominant additional processing time in large libraries is the cost of docking multiple ring conformations rather than the generation process itself.

Knowledge in docking: placed substructural fragments

One of the key practical aspects in applying docking systems in drug design is in the situation where one is exploring a chemical structural space of analogs of some parent compound. In these cases, where it is reasonable to posit that a particular substructure will remain largely stationary in an active site (as with, for example, metal chelation moieties), making direct use of that knowledge to constrain the search space offers advantages in terms of workflow, speed, and direct comparison of different analogs. Surflex-Dock's fragment-based docking mode supports this type of analysis, and the ability to impose a real-valued constraint on the degree to which the placed fragment must ‘‘hold’’ the docked ligand offers the modeler a method to control pose variation.

The procedure is simple and is illustrated in Fig. 4. Panel A shows the Surflex-Dock protomol for HDM2 from

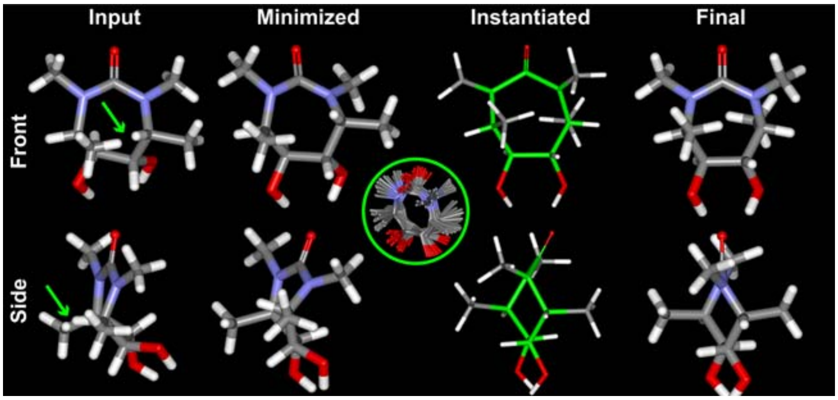

Fig. 3 The skeleton of the central ring system of the HVR inhibitor from Fig. 1 is shown progressing through the ring flexibility algorithm. The input conformation used in the J&J benchmark set had a significant clash (indicated by the *arrows*). The minimization step relieved the clash, but the ring conformation was still strained. The *circle* indicates a superimposition of the 28 new ring conformations obtained by instantiating two cycloheptane conformers with 14 possible graph-based matches each. The conformation highlighted in *green* led to the lowest energy ring conformation shown at *right*. Note that the bond lengths of what should be the C–N bonds were too long in the instantiated version since they came from cycloheptane. Also, the ring substituents moved based on changes in the ring atom positions, but they were not optimal. The final minimization addressed these considerations

the J&J set along with a placed fragment specified for a narrow hydrophobic cleft within the binding pocket. A placed fragment (in this case toluene, shown in purple) serves as a guide to the alignment of ligands that contain the fragment. Matches may be done either by ignoring all hydrogens (as in the example) or by explicit match of all atoms, and heuristics are employed to make intuitive choices with respect to hybridization of ligand atoms and fragment atoms. The ligand to be docked is aligned based on the fragment position, and its conformational space is enumerated depth-wise from the fragment outward. For each particular depth, if the conformational change to the parent pose affects the geometry of the atoms within the matched substructure, the alignment is recomputed. Then, the new conformation is scored with respect to its similarity to the protomol, using a rapid approximate computation. The best scoring of the new poses (shown in blue) are subjected to local optimization using the Surflex-Dock scoring function. This process may produce deviations from the initial alignment, as shown in the example. For molecules with a large number of rotatable bonds, the process iterates from the highest scoring poses after each round of local optimization. The process terminates when all rotatable bonds have been optimized. Panel D shows the relationship between the top scoring pose and the bound pose, a deviation of 1.0 Å.

Fragment-guided searching is selected by the ‘‘**-fmatch**’’ option, which requires a specified fragment molecule, and the behavior may be modified by imposing a real-valued penalty on the deviation from the placed fragment (‘‘**-cpen**’’ in units of $pK_d/\text{Å}^2$). For screening a library of molecules, the user may specify that molecules not containing the specific fragment be skipped (‘‘**-fskip**’’) and that hydrogens on the fragment should be matched explicitly (‘‘**-fhmatch**’’). Given an accurate location of a subfragment, the docking process can be sped up several-fold using the fragment-guided docking procedure compared with the de novo procedure required to produce results of similar quality.

Computational procedures

Computational procedures in studies such as this can have a remarkable impact on results, both with respect to the actual performance of algorithms but also as to the comparability of different methods that have been run on nominally the same benchmarks. Wherever possible, choices were made in the application of procedures to preserve meaningful comparisons across different studies.

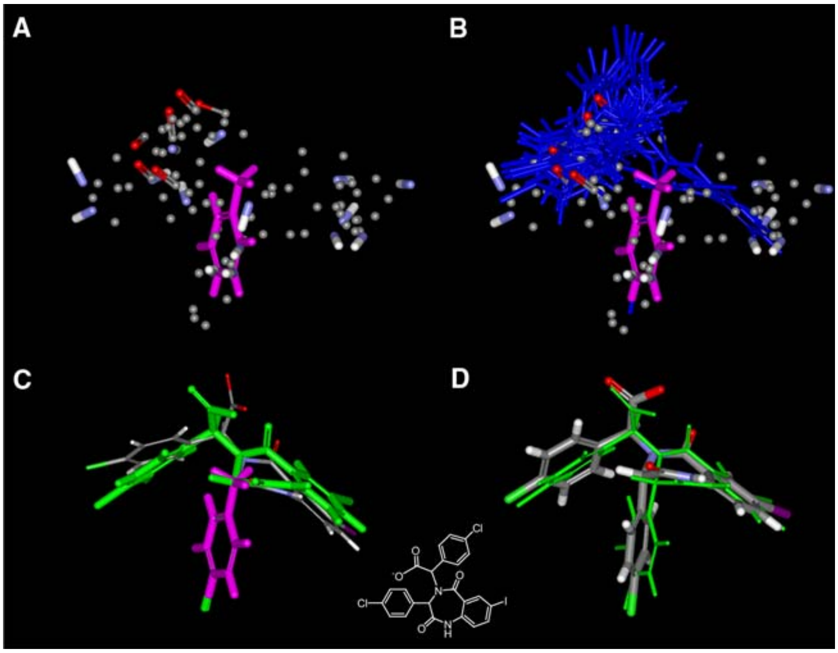

Fig. 4 *Panel A* shows the Surflex-Dock pocket characterization with a placed fragment specified for the HDM2 hydrophobic binding pocket. The protomol was composed of three types of fragments: methane (shown without hydrogens), N–H, and C=O. The fragment (toluene, in *purple*), matched the pendant chlorophenyls of the ligand (*white*) in two ways each, resulting in four different alignment matches. The alignment match that gave rise to the highest scoring final conformation is shown in *Panel B*. The poses depicted in *blue sticks* were the highest scoring conformations of the ligand-based solely on a fast similarity computation relative to the protomol (50 from a total of over 20,000). *Panel C* shows the particular pose that gave rise to the top scoring final conformation, following all-atom optimization within the HDM2 binding pocket. *Panel D* shows the relationship between the top scoring pose and the bound pose, a deviation of 1.0 Å. The docking took 11 s of real time on standard 2.8 GHz desktop hardware

In particular, this required either using public benchmarks in their unmodified state or showing the effects of modification. The publicly available data archive associated with this paper contains all protein, ligand, and protomol structures as well as example scripts for the primary experiments described. The following summarizes the procedures used.

Ligand preparation

For the experiments on the J&J and Vertex benchmark sets, all ligands were used in docking procedures completely unmodified as input to Surflex-Dock. For the J&J set, this involved ligands in SDF file formats, generated by either CORINA or Concord, as described above. For the Vertex set, this involved CORINA-generated ligands in SDF format.

For the experiments on the Pham set, as mentioned above, the 256 active ligands and the 1,861 decoys ligands were processed to remove all protons, re-add them using rules to yield appropriate protonation states at physiological pH, and to minimize the result. The command was ‘‘surflex-dock –fp –remin prot ligand_archive.mol2 prot_ligand_archive.’’ This was done to address a problem with some of the structures within the decoys data sets relating to fused aromatic ring systems as well as the protonation state of certain nitrogen atoms. Note that symmetric treatment of all actives and all decoys is an extremely important feature in systematic screening enrichment experiments, as will become apparent later. The automated procedures have been implemented in order to provide an objective method for molecule preparation to eliminate bias in benchmarking. Users are encouraged, however, to think carefully about their treatment of protons on both proteins and ligands and to use knowledge where appropriate.

For the experiments on the Jain set, instead of using ten random input conformations (the file ‘‘random10.mol2’’ in the archive from that study), a single random conformation was used (‘‘ran1.mol2’’ in the supplementary data for this paper). This was done to conform more closely with what has become the standard for testing docking accuracy: a single conformation and starting alignment of each test ligand that is unrelated to its crystallographic pose.

Protein preparation

Protein structures from the Vertex, Pham, and Jain sets were used unmodified from previous reports for all comparative experiments. The Vertex proteins were in MacroModel ‘‘.dat’’ format. The Pham and Jain sets employed SYBYL ‘‘.mol2’’ formats. Protein structures for the J&J set had hydrogens added where missing, beginning from PDB structures, and the resulting structures were saved as SYBYL ‘‘.mol2’’ files (using WebLabViewer). Protomols were generated using standard fully automated Surflex-Dock procedures [16]. Note that file formats do not materially affect the algorithms as long as atomic coordinates do not change and assignment of atom hybridization is unambiguous.

Screening experiments

Exploration of the effects of various algorithmic modifications on screening utility is a focus of this work. Direct comparisons to the Surflex-Dock Version 1.3 code were made using no new optional parameters (‘‘surflex-dock dock_list test_archive.mol2 p1-protomol.mol2 protein.mol2 log’’). In what follows, the effect of pre-docking ligand minimization is described as ‘‘preminimization’’ or abbreviated as ‘‘ligand pre-min’’ and corresponds to adding the ‘‘-premin’’ switch to the docking protocol (‘‘surflex-dock –premin dock_list test_archive.mol2 p1-protomol.mol2 protein.mol2 log’’). The effect of post-docking all-atom ligand minimization, combining the Surflex-Dock intermolecular scoring function with the DREIDING force-field, is described as ‘‘re-minimization’’ or as ‘‘all atom minimization’’ and corresponds to adding the ‘‘-remin’’ switch to the docking protocol (‘‘surflex-dock –premin –remin dock_list test_archive.mol2 p1-protomol.mol2 protein.mol2 log’’). This protocol is also implemented as a single parameter switch (‘‘-pscreen’’), reflecting its utility in producing robust screening performance. As discussed above, in the J&J HDM2 case, the fragment-based docking method was used with the toluene fragment shown in Fig. 4 (‘‘surflex-dock –fmatch frag.mol2 ...’’) with no other options. Consequently, all ligands in the test database for HDM2, both actives and inactives, that contained a pendant phenyl group were docked by matching to the placed toluene and all other ligands were docked in the normal fashion.

Docking accuracy experiments

Parameter choices for measuring docking accuracy were made to provide directly comparable data on the Vertex benchmark, for which the performance of Glide and GOLD were available. That study employed parameter choices that resulted in mean running times of 1–3 min per ligand for Glide and GOLD, with 20 poses returned at the end of docking. Surflex-Dock's geometric docking accuracy protocol is implemented as a single parameter (‘‘-pgeom’’) and resulted in a mean time per ligand on the Vertex set of

2 min. The search procedure is approximately fivefold slower than with the screening parameters (‘‘surflex-dock –pgeom dock_list test_corina_ligand.sdf p1-protomol.mol2 protein.dat log’’) and shares the pre-minimization and re-minimization features. The effect of exploring ring flexibility was also explored, and this corresponds to adding the ‘‘-ring’’ switch to the docking protocol (‘‘surflex-dock –pgeom –ring dock_list test_corina_ligand.sdf p1-protomol.mol2 protein.dat log’’). With the particular distribution of frequency of flexible rings in the Vertex set, this increased the mean docking time by a factor of 2. The same parameters were used for the Jain set.

Quantification of performance

Quantification of screening utility seeks to measure the enrichment of known ligands over decoys ligands based on a ranking generated by a virtual screening protocol (as seen in a number of recent reports of both docking and molecular similarity [13, 16, 18–20, 30, 43]). Quantification of the degree of separation between true positive ligands and false positives was done using enrichment plots in the case of the J&J data, where specific enrichments for other methods were available from the published data [24].

For the Pham set, following previous work, receiver operating characteristic (ROC) curves along with their corresponding areas were used. Given a set of scores for positives and negatives, the ROC curve plots the true positive proportion (*Y*-axis) with the corresponding false positive proportion (*X*-axis) at all possible choices of some threshold that would mark a binary distinction between a prediction of positive or negative class membership. The perfect ROC curve goes from [0,0] to [0,1] to [1,0] and results in an area of 1.0. Complete intermixing of positive and negative scores gives an area of 0.5, with areas <0.5 reflecting the case where true positives are ranked lower than false positives. In screening enrichment datasets, the number of positives (true ligands) is, of necessity, much smaller than the number of negatives (decoys). Consequently, it can be informative to compute confidence limits on the ROC areas, since perturbations in the ranks of a small number of positives can lead to very large changes in the computed ROC area. Multiple methods exist within statistics for confidence interval estimation in ROC analysis, but a particularly widely used method, called the bootstrap percentile, allows for computation of confidence intervals in a non-parametric fashion and is used here [44]. ROC analysis is employed in this study because it is a well-characterized statistical method, other types of performance measures are derivable from ROC curves (e.g., enrichment plots, maximal enrichment values, specific TP and FP rate tables), and ROC curves are insensitive to the effects of relative size of positive and decoy sets.

Table 1 shows the performance of Surflex-Dock Version 1.3 on the original Pham benchmark along with performance on revision 1 (i.e., with the reprocessed structures), which is the subject of the present study (see above). The last column of the table gives the 95% confidence interval for the ROC area on the revised data using the bootstrap percentile method. Values are highlighted with **bold underlining** for improvements and (parentheses) for degradations where the confidence limits exclude the value under the previous condition. In situations where the number of positive examples is small, the confidence limits will be wider than when the number of positive examples is

Table 1 Effects of decoy set revision for Rognan ACD database

Protein	N	Rognan decoys	Rognan decoys (revision1)	
		Version 1.3	Version 1.3	Version 1.3 95% CI
1AJQ	6	0.922	0.893	0.82–0.95
1B5J	16	1.000	1.000	1.00–1.00
1B7H	6	0.999	1.000	1.00–1.00
1BXO	5	0.746	0.875	0.66–0.99
1BZH	12	0.917	0.897	0.86–0.93
1C4V	20	0.876	0.879	0.81–0.94
1E66	6	0.764	0.804	0.64–0.92
1EIX	5	0.996	0.998	0.99–1.00
1F4G	10	0.693	(0.545)	0.43–0.67
1FH8	6	0.997	0.993	0.99–1.00
1FJS	6	0.980	0.962	0.88–1.00
1FMO	8	0.764	0.787	0.54–0.97
1GJ7	12	0.953	0.956	0.89–0.99
1PRO	20	0.862	0.829	0.71–0.92
1QBO	20	0.990	0.964	0.91–1.00
1QHC	6	0.791	0.830	0.69–0.96
1RNT	5	0.952	0.930	0.82–1.00
2QWG	7	0.965	0.954	0.89–0.99
2XIS	5	0.958	0.926	0.85–0.97
3PCJ	8	0.948	0.957	0.92–0.98
3STD	5	0.844	0.918	0.85–0.98
4TMN	13	0.828	0.808	0.67–0.93
7CPA	8	0.901	0.816	0.67–0.93
7TIM	6	0.966	0.978	0.96–0.99
ER	10	0.922	**0.998**	0.94–1.00
TK	10	0.963	0.971	0.95–0.99
PARP	15	0.846	0.860	0.77–0.93

For the 27 protein structures in the revised Pham/Jain benchmark, comparative ROC areas are shown along with the 95% confidence interval for the ROC area with the revised data set. One protein (1F4G) shows performance falling below the original data set version, indicated in parentheses. One protein (ER) shows performance exceeding the original version, indicated in bold and underlined. This convention will be used in subsequent tables to highlight significant performance changes moving from left to right by columns large (presuming identical variance of positive scores). Similarly, small variance among the positive scores leads to narrower confidence limits. In making the data revision, one case showed a significant decrease in performance (1F4G), but this was related to an inefficiency in local optimization. One case showed a significant increase in performance (ER), and was apparently related to an improvement in the geometry of one of the active ligands. Overall, performance was unaffected by the data revision.

Results

In what follows, the results of systematic application of Surflex-Dock's new computational procedures are presented on the four benchmark data sets described earlier (see Methods for details about the data sets, computational methods, and specific procedures).

Screening enrichment

Systematic screening experiments were carried out on the J&J and Pham screening benchmarks to test the effects and importance of searching beyond pure pose parameter space.

J&J benchmark

Benchmark data sets made available by pharma companies are of great interest, owing to the fact that the compositional choices tend to reflect the pharmaceutical relevance of the specific targets selected and also reflect the operational use of computational methods. One of the characteristics of large corporate structural databases is the extent to which, over time, different procedures may have been used, for example, to generate 3D structures from the primary 2D records. In the J&J benchmark, an aspect of this manifested in a difference between the preparation of the active ligands and the decoy ligands (see the Methods for details). While the intention of this choice was to provide greater fidelity in terms of chirality for the active ligands, there was an unintended consequence with respect to the strain energies of the different ligand sets.

Figure 5 shows cumulative histograms of the strain energy for the ligands of each of the four target proteins along with that of the decoy ligands (plot A, computed used Surflex's DREIDING force-field). The decoy ligands had a median strain of roughly 60 kcal/mol (light blue curve). The distribution of strain energies for the PTP ligands matched this closely (magenta curve). However, the ligands of HVR, HDM2, and THR showed markedly larger strains, with medians of roughly 140, 145, and 190 kcal/mol, respectively (green, red, and blue curves). Many docking algorithms, including Surflex-Dock in its default mode, alter the pose of their input ligands only through translation, rotation, and dihedral angle variation in rotatable bonds. Of

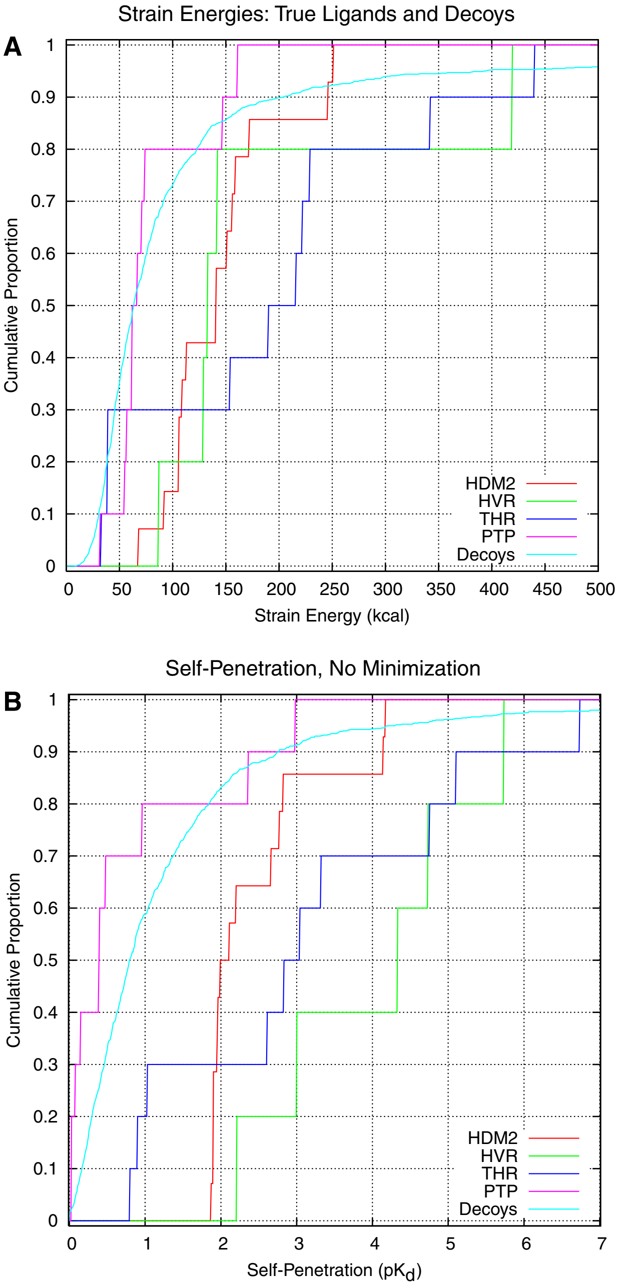

Fig. 5 *Panel A* shows the cumulative histograms of strain energies for the positive (*red*, *green*, *blue*, and *magenta curves*) and decoy ligands within the J&J benchmark set (*light blue curve*). The strain energies were high as a general feature, but the key issue was that a significant number of the cognate ligands had *higher* strain than the decoys. *Panel B* shows the self-penetration penalties in Surflex's output units of pK$_d$. For the PTP ligands (*magenta curve*), there was a relative advantage over the decoys, as with a small number of the THR ligands. For the remaining THR ligands and the full sets of HVR and HDM2 ligands, there was a strong and systematic bias against the cognate molecules

necessity, such algorithms must avoid introducing intramolecular clashes within the ligand; Surflex does this by imposing a restraint against self-clashing that is the same as the term that prevents intermolecular clashing with the protein. However, some input ligand geometries will result in internal clashes that are unresolvable by bond rotation (see the ring system in Fig. 3 for an example). In Fig. 5, plot B shows the magnitude of the *unresolvable* self-clashing penalties for the J&J ligands sets, which paralleled the strain energies. While the PTP ligands showed an advantage relative to the decoys (roughly 0.5 pK_d at the medians), the other ligand sets showed disadvantages of 1–3.5 U of pK_d. In assessing the results of a screening exercise, whether for methodological evaluation or for actual practical lead discovery, the difference between the distribution of docking scores between the actives and the inactives is all that is measured. In the case of methodological evaluation, an a priori bias of multiple log units against the known ligands will result in poor nominal performance. In the case of an actual lead discovery exercise, such a bias may result in missing the true ligands of the target protein altogether.

This observation should not be interpreted as a criticism of the J&J benchmark or of the study itself [24]. Rather, it highlights the need for virtual screening methods to offer options that are robust to these types of real-world variation in input structure preparation, since programs such as CORINA and Concord are widely used and well respected methods for database preparation. Note also that the energetic strain effects were chemical class specific, with CORINA generated PTP ligands having low strain but CORINA generated HVR ligands having very high strain. Figure 6 shows the effect of docking the HVR inhibitor from Fig. 1 under different parameter choices for Surflex-Dock, all beginning from the highly strained input structure (the central ring of which is depicted in Fig. 3). Panel A shows the result of docking under the assumption that the input conformation only requires changes in alignment and in its acyclic rotatable bonds. Not only was it clearly a poor result in terms of the geometric match to the bound pose, but it scored very poorly (1.2 pK_d). Panel B adds ligand minimization prior to docking and within the active site after docking, resulting in a vastly improved score (9.1) and a somewhat improved geometric quality. In this case, all atom optimization within the binding pocket allowed the hard intermolecular and intramolecular clashes to be eliminated while still allowing occupancy of the four hydrophobic pockets within the protease. Panel C adds ring flexibility, resulting in another improvement in score (pK_d of 13.6 for this subnanomolar compound) as well as an excellent geometric result (1.3 Å rmsd).

Figure 7 (plot A) shows the effect of ligand minimization on the distributions of self-penetration for the four active ligand sets and the decoys. While the HVR ligands still showed a slight disadvantage, it was <0.5 U, the distributions of the other ligand groups became quite homogeneous. Plot B shows the effect of adding pre-docking ligand minimization, post-docking all-atom minimization, and ring flexibility successively on screening enrichment with full ROC curves. Completely ignoring the strain issues of the input ligands resulted in docking performance that was worse than random (red curve). Adding ligand minimization prior to docking very significantly improved performance (green curve). Adding post-docking all atom optimization resulted in excellent performance (blue curve, ROC area 0.940), and adding ring flexibility added another marginal increase in performance (magenta curve). The effect of each successive change in protocol on screening performance reflected the degree to which each change affected the scores of the

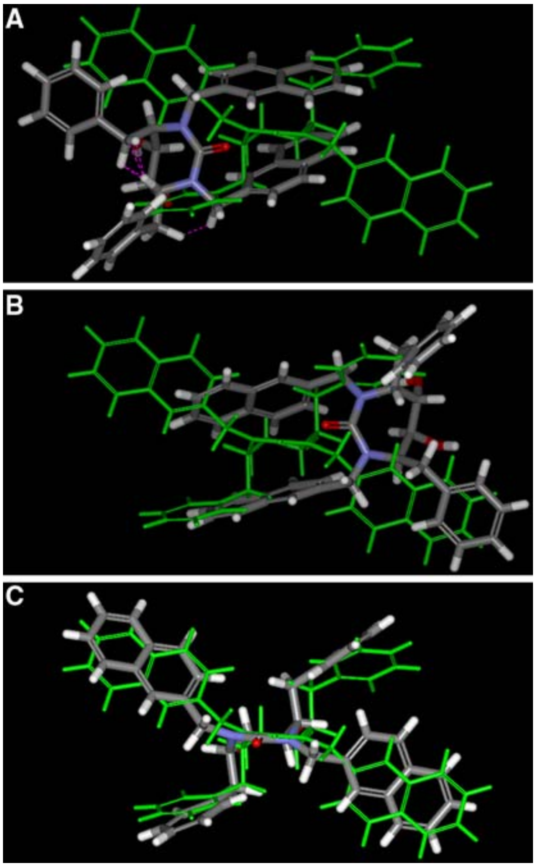

Fig. 6 *Panel A* shows the overlay of the unminimized docked HVR inhibitor from Fig. 1 with the crystallographic conformation (*green*). The *magenta lines* indicate self-clashes that were unresolvable through rotation of acyclic bonds. These were due to the highly strained ring conformation (see Fig. 3). *Panel B* shows the effect of the -premin and -remin Surflex-Dock parameters, which relieved all such interpenetrations and increased the score by ~8 U of pK_d. *Panel C* shows the effect of adding the -ring parameter, in which the conformational space of the seven-membered central ring is searched. The resulting docking was excellent (1.3 Å rmsd)

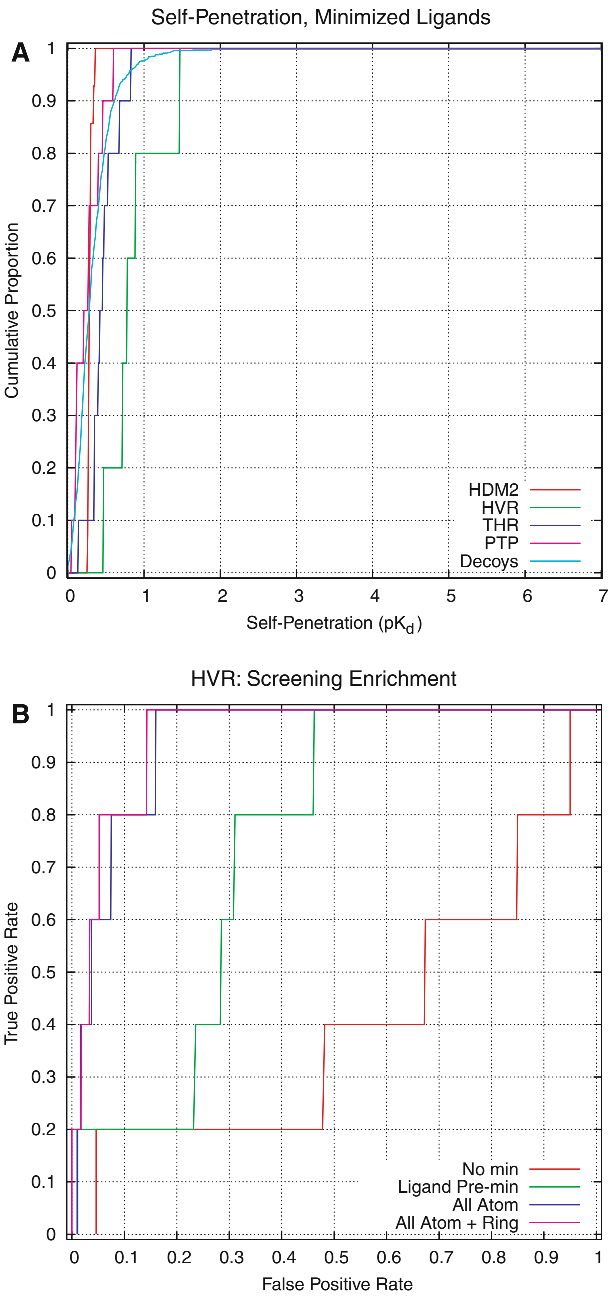

Fig. 7 Plot **a** shows the effect of ligand minimization on the computed Surflex-Dock self-penetration. In contrast to **a**, nearly all of the self-penetrations for all ligand sets were below 1 U of pK$_d$. Plot **b** shows the respective ROC curves for enrichment of HVR ligands against the J&J decoy set under different parameter settings. Without addressing the very high-ligand strain (*red curve*), only a single cognate ligand was identified at a false positive rate of 5% or lower. This increased to three cognate ligands (60%) with more aggressive search (*magenta and blue curves*). Pre-minimization made a significant improvement alone (*green*), but the addition of post-docking minimization (*blue*) yielded excellent performance. While adding ring flexing to the protocol improved the docking accuracy of the cognate ligands (see Fig. 6), it only marginally improved screening performance (*magenta*)

actives relative to the inactives. Based on Fig. 5, it would be possible to make a confident a priori prediction that pre-minimization would yield an improvement for HVR. However, the effects of the other protocol changes depended largely on whether a significant number of flexible ring systems existed in a set of actives compared with inactives. Given active ligands containing flexible ring systems, both post-docking all atom optimization and ring flexibility improved performance, but otherwise occasionally not.

Table 2 shows the effects of these protocol variations on the ROC areas of each of the four J&J protein test cases. In both the HVR and HDM2 cases, where flexible rings were important in the binding of the cognate ligands, ligand pre-minimization and post-docking all atom optimization significantly improved enrichment performance. Ring flexibility added only a marginal improvement. In the case of THR, ligand pre-minimization significantly improved performance, with marginal and insignificant decreases associated with the remaining two variations. Last, in the case of PTP, we see what might be considered a paradoxical result: each successive refinement in protocol lead to a marginal decrease in performance. However, recall from Fig. 5, PTP was the only case where the cognate ligands were *less* strained than the decoys. This advantage disappeared as the docking protocol took ligand strain out of the equation.

Given the overall tradeoffs in speed and screening performance, the recommended screening protocol is to

Table 2 Effects of Surflex-Dock procedural changes on screening enrichment in the J&J screening benchmark

	No minimization	Pre-minimization	Add re-minimization	Add ring flexibility
HDM2	0.303	**0.827**	**0.906**	0.911
HVR	0.400	**0.740**	**0.940**	0.951
THR	0.791	**0.949**	0.944	0.928
PTP	0.963	0.942	0.904	0.805

ROC areas are shown, with the same conventions as in Table 1. Due to the substantial bias in internal strain energies of active ligands versus decoys (see Figs. 1, 2), results without ligand minimization are acceptable only in the THR and PTP cases, where the bias is less unfavorable than for HDM2 and HVR. Adding ligand minimization prior to docking substantially improves performance in three cases, with a marginal change in the fourth. Adding post-docking all-atom optimization of the docked ligands in their bound poses substantially improves performance in the HDM2 and HVR cases, both of which have a significant number of ligands with flexible rings. Adding explicit ring searching yields only a marginal improvement in screening enrichment in those two cases, but increases docking time and generates marginal decreases in the THR and PTP cases which do not have significant flexible ring constraints in their cognate ligands employ ligand pre-minimization and post-docking all atom optimization (implemented in Surflex-Dock Version 2.1 as the ‘‘–pscreen’’ parameter). On the J&J set, against the large HVR site, the mean screening time per rotatable bond increased from 3.0 to 4.1 s (real user time) with the screening parameters compared with default parameters. Figure 8 shows a direct comparison of Surflex-Dock’s results with those reported in the original paper by Cummings et al. for DockVision, Dock, GOLD, and Glide. At the top end of the ranked database (2 and 5% levels), Surflex-Dock performed better than all other methods for HVR, HDM2, and THR. For PTP, Surflex’s performance, using the screening parameters, was slightly better than Glide’s but GOLD and DockVision performed better.

Overall performance of the methods across the different targets paralleled what one would expect. In the three cases with systematic bias against the actives, at least one of the dockers performed at or worse than random, with all but Surflex performing worse than random in at least one case. In the case of PTP, where the prior bias embedded in the ligands *favored* the actives over the decoys, no method performed worse than random, and both DockVision and GOLD performed very well. In this case, using default parameters (and thus taking advantage of the prior bias), Surflex’s performance was very similar to GOLD’s (data not shown).

The issue of the fairness of this comparison will be discussed later, but two points are worth elaborating here. The first is that among all of the methods, only Surflex and Glide employed all atom optimization of the docked ligands as part of the computational procedure. This would appear to explain the fact that Surflex performed well in the non-PTP cases, as did Glide in the HVR and THR cases, but both did relatively worse in the case of PTP. The second is that the advantage of addressing subtleties in ligand energetics is a genuine advantage in the real world of virtual screening. For Surflex, at a relatively low cost in terms of computational speed, substantial improvements in robustness were obtained in the face of variations in input.

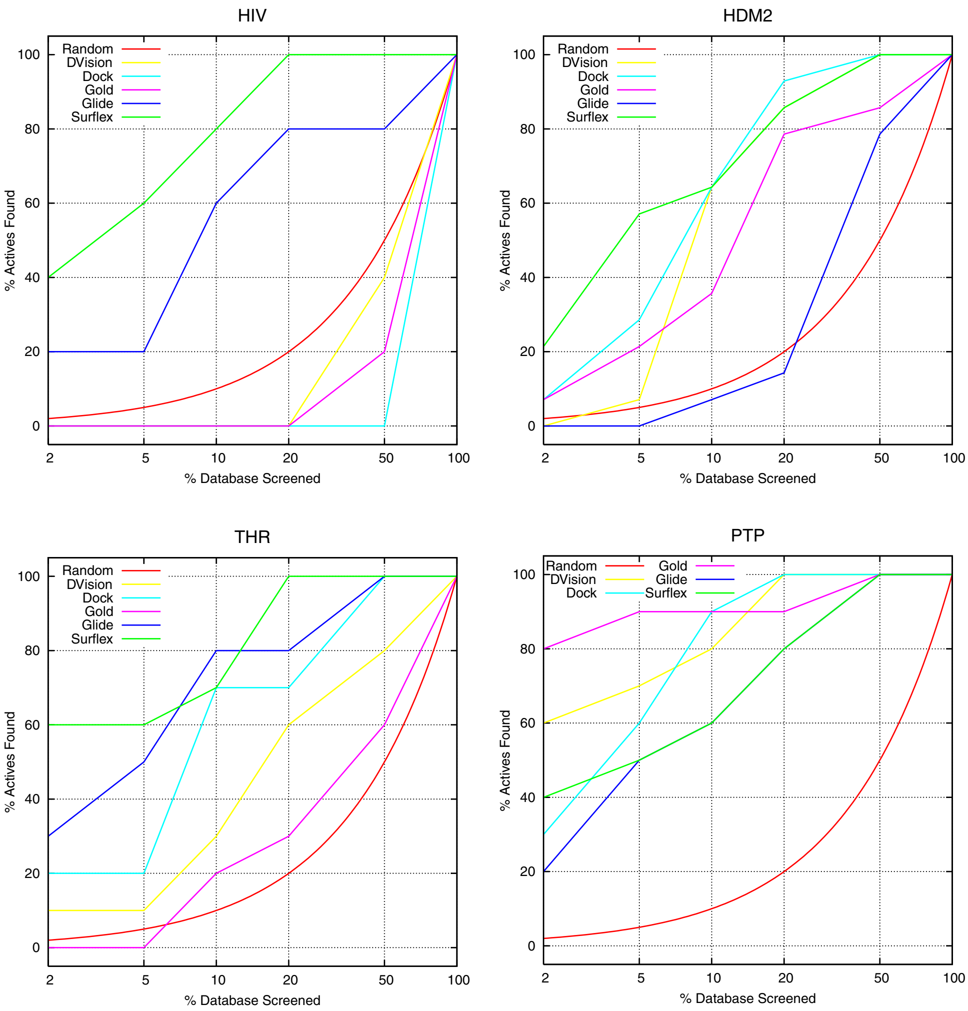

Fig. 8 The enrichment plots show the comparative performance of Surflex-Dock Version 2.1 with DockVision, Dock, Gold, and Glide. Data for all programs but Surflex-Dock were taken from Cummings et al.[24]. In all but the case of PTP, one of the programs performed worse than random, likely due to the differential energetic bias among the cognate ligands in the different cases (see Fig. 5). In all but the PTP case, Surflex’s performance at the early enrichment levels of 2 and 5% was best. All programs but Surflex performed worse than random in at least one case. Surflex’s performance is shown with standard screening parameters, and the HDM2 case employed the fragment-based docking protocol (see Fig. 4)

Pham benchmark

There are four primary differences between the J&J benchmark set and the Pham set for evaluating screening utility. The first and most obvious is that the Pham set contains 27 proteins, representing a broader set of target types. The second is that the Pham set contains a high proportion of ligands with relatively poor binding affinity ($pK_d < 6.0$ for 50% of the ligands). Third, for the ER and TK targets, adapted from Bissantz et al. [18], there are a number of comparative performance reports in the literature. Last, the Pham set contains two different decoy sets, one derived from Bissantz et al. (termed the Rognan decoys) and one derived from the ZINC database.

Table 3 shows ROC areas for a number of conditions, comparing both the old Version (1.3) with the new Surflex-Dock Version 2.1, the effects of different decoys, and the effects of making use of the recommended screening parameters, which correspond to adding pre-docking ligand minimization and post-docking all atom ligand optimization to the default case. For the Rognan decoy set, the change from Version 1.3 to 2.1 yielded significant improvements in four cases (1BXO, 1F4G, 1FJS, and 1QHC), with no cases exhibiting a significant performance decrease. The primary difference was the improvement in local optimization using the BFGS algorithm (see Methods). Coupled with analytical gradients for the Surflex scoring function, the new version, in addition to showing quantitatively better performance, was ~30% faster. As observed in the original report, changing from the Rognan decoy set to the ZINC decoy set made little difference, with just two cases showing marginally significant decreases in performance.

Use of the recommended screening parameters made no statistically significant change overall within the 27 protein set, with a single case exhibiting a minor difference beyond the 95% confidence interval of ROC area. The mean ROC area for Version 2.1 was 0.91 (Rognan decoys, default parameters), 0.90 (ZINC decoys, default parameters), and 0.88 (ZINC decoys, screening parameters). In terms of population differences, none of these changes were statistically significant. This reflected the fact that the Pham actives and decoys were all pre-processed including a minimization step, which reduced any impact of the screening parameters. Also contributing to the level performance, the Pham actives, as a group, did not have a high proportion of centrally located flexible rings. Recall from Table 2, the mean performance of Surflex Version 2.1 on the J&J set changed from 0.61 (default parameters), to 0.86 (ligand pre-minimization), to 0.92 (screening parameters), to 0.90 (adding ring searching). Ligand minimization significantly improved performance in three of four cases, and post-docking all atom optimization significantly improved performance in two of four cases.

Table 3 Effects of decoy set and Surflex-Dock version on screening utility for the Pham benchmark

Protein	N	Rognan decoys (rev1)		Zinc decoys (rev1)	
		Version 1.3	Version 2.1	Version 2.1	Version 2.1 - pscreen
1AJQ	6	0.893	0.900	(0.828)	0.726
1B5J	16	1.000	1.000	1.000	1.000
1B7H	6	1.000	1.000	1.000	1.000
1BXO	5	0.875	**0.958**	0.965	0.986
1BZH	12	0.897	0.913	0.924	(0.861)
1C4V	20	0.879	0.938	0.927	0.926
1E66	6	0.804	0.765	0.736	0.619
1EIX	5	0.998	0.999	1.000	1.000
1F4G	10	0.545	**0.702**	0.643	0.605
1FH8	6	0.993	0.936	0.924	0.947
1FJS	6	0.962	**0.986**	0.988	0.997
1FMO	8	0.787	0.747	0.738	0.681
1GJ7	12	0.956	0.958	0.951	0.948
1PRO	20	0.829	0.896	0.888	0.952
1QBO	20	0.964	0.968	0.956	0.949
1QHC	6	0.830	**0.927**	0.917	0.894
1RNT	5	0.930	0.922	0.919	0.919
2QWG	7	0.954	0.945	0.938	0.939
2XIS	5	0.926	0.930	0.946	0.923
3PCJ	8	0.957	0.914	(0.868)	0.806
3STD	5	0.918	0.879	0.836	0.802
4TMN	13	0.808	0.845	0.849	0.881
7CPA	8	0.816	0.862	0.787	0.757
7TIM	6	0.978	0.962	0.961	0.927
ER	10	0.998	0.977	0.989	0.999
TK	10	0.971	0.961	0.960	0.955
PARP	15	0.860	0.846	0.821	0.699

Comparative ROC areas are shown, with the convention from Table 1 regarding significant differences. Four cases showed significant improvement in moving from Surflex Version 1.3–2.1, with the primary change being improved local optimization using the BFGS optimization scheme with analytical gradients. In sharp contrast to other reports, the ZINC decoy set did not substantially decrease performance, with only a two cases registering minor, but statistically meaningful, decreases. Making use of the preferred screening parameters, which are intended as a robust safeguard against biases in ligand preparation, did not significantly alter performance. None of the column pairs have significantly different populations of ROC areas based on multiple statistics

In terms of comparing results from the J&J set and the Pham set, the ‘‘default’’ parameter case is not sensibly comparable due to the asymmetric ligand energetics in the J&J set discussed earlier. However, the comparison between the ligand pre-minimization protocol on the J&J set and the Pham set using drug-like Zinc decoys is fair, with mean ROC areas of 0.86 and 0.90, respectively. Similarly, the comparison using screening parameters is reasonable, with mean ROC areas of 0.92 and 0.88 for the J&J and Pham sets, respectively. It is striking that the comparable results from two very different benchmarks yielded such similar performance levels.

Summary: screening enrichment

The practical significance of what has been reported on the combination of the J&J and Pham data sets is that by employing a uniform procedure with Surflex-Dock's ''-pscreen'' option, one can obtain robust and reliable performance across a wide variety of targets with varying compositions of actives and inactives.

The results presented above focused on ROC areas, since they support direct statistical comparisons. A somewhat more intuitive measure of screening efficiency is enrichment, which corresponds to the ratio between the actual number of active ligands identified versus the number expected by chance given some proportion of a ranked database. Maximal enrichment values are bounded above by the ratio of the total number of compounds in a library to the number of actives (up to 200-fold in the cases discussed above) and are computed by considering all possible coverage levels of the ranked database. Enrichment values at fixed percentages are bounded above by the reciprocal of the proportion of the library selected (e.g., 100-fold for 1% coverage). Using Surflex's screening parameters, on the Pham set with the ZINC decoys, the mean maximal enrichment was 100-fold, the mean enrichment at 1% coverage was 37-fold, and at 5%, it was 12-fold. For the J&J set, the comparable numbers were 43-fold, 15-fold, and 11-fold. With respect to the comparison with other docking methods on the four targets of the J&J benchmark, the Surflex screening protocol yielded average enrichments that were twice as good as each of Glide, GOLD, Dock, and DockVision at both the 2% and 5% coverage levels. This corresponds well to the comparisons of Surflex-Dock to other methods on the ER and TK benchmarks of Rognan's group from other studies [13, 16, 19, 21].

Note that comparisons of docking methods to *ligand-based* methods for screening are complicated to interpret, since there are subtle issues of inductive bias present in any ligand-based approach. Direct knowledge of the actual structures of some true positives are used to retrieve other ligands. This is discussed extensively in a previous paper on the Surflex-Sim methodology [43]. A recent paper by Hawkins et al. makes a direct comparison between ROCS and Surflex-Dock using the Pham set, showing comparable performance of both methods [45]. However, in 50% of the Pham set retrieval cases, a trivial 2D molecular similarity method is able to perform nearly perfectly (>0.95 ROC area) and in an additional 30%, the trivial method is able to perform very well (>0.80 ROC area). Meaningful comparisons of ligand-based and protein structure-based methods for virtual screening are an area of current investigation.

Docking accuracy

One of the key applications of docking methodology is in the modeling of synthetic analogs of compounds for which a bound structure is known of a parent compound. In these cases, a docker's prediction of bound pose can have a significant impact on synthetic choices, with the decisions guided both by the extent to which a proposed compound fits its new moiety in a desired place and possibly also by the nominal predicted scores of the proposed analogs. In such cases, it is frequently the case that reliable knowledge exists about the placement of a particular molecular substructure. By providing that information to a docking system, a modeler can directly explore the question of interest (''Where do the new pieces end up?'') quickly and while factoring out some aspects of variance in docking procedures. Figure 9 shows an example from the Vertex benchmark that is typical of a number of metal catalyzed enzymes in the set. In Panel A, the result of docking the inhibitor using the constraint of the hydroxamic acid fragment (shown in magenta) is shown overlayed with the true bound pose of the ligand (green). In this docking, the inhibitor placement is *guided* by the fragment but is not *constrained* by it. Surflex-Dock allows the user to enforce a quantitative constraint on the deviation from the placed fragment (see Methods). With this constraint set at a high value, the restraining force on the ligand restricts its movement, with the result being perfect overlap with the known ideal metal chelation geometry. This type of experiment is common in the practical use of docking, but it is not informative in assessing docking accuracy for comparative purposes.

In evaluations of docking accuracy, it is typical to employ a somewhat artificial test, one in which a docker seeks to recover the true binding mode of a ligand given the structure of a protein in the state in which it is bound to the *same* ligand. This type of test can be helpful in establishing an upper bound on the expected performance of a docker performing in a de novo docking mode with respect to the two criteria of best returned pose by rmsd and rmsd of the top scoring returned pose. The intention is that such characterizations will effectively differentiate between algorithms that will be reliable in practice.

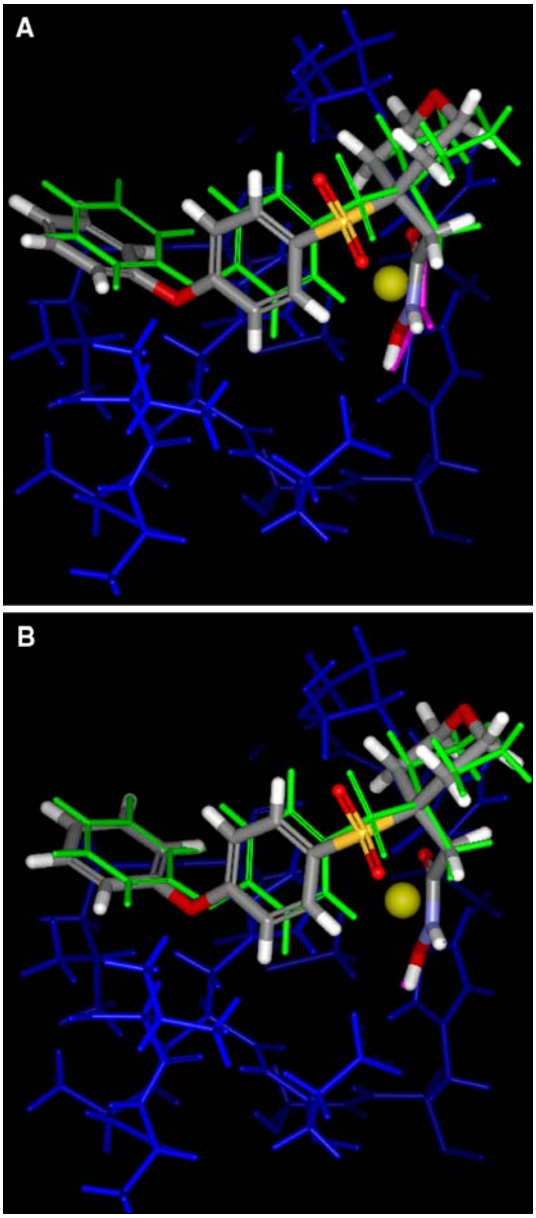

Fig. 9 The two panels depict the effect of adding a constraint to hold a ligand's match to a specified placed molecular fragment. *Panel A* shows the result with no constraint, with the fragment in *magenta*, the bound conformation of the ligand in *green*, and the top scoring pose in atom color (0.73 Å rmsd). *Panel B* shows the result with a constraint (-cpen 100), resulting in an improved pose (0.50 Å rmsd) at the expense of a slightly lower score. Note that the fragment deviation was reduced in this case, with the docked pose exactly overlapping the placed fragment. Constraint of metal chelation geometries, as with this hydroxamic acid interaction with an active site zinc of collagenase (PDB code 966C from the Vertex set), can be beneficial both in terms of docking speed as well as (obviously) for geometric accuracy

Surflex performance: the effect of ring search

Figure 10 shows Surflex's performance on both the Jain 81 complex set and the Vertex 100 complex set, with comparative performance for both Glide and GOLD from the report of Perola et al. [23]. The plots are cumulative histograms of rmsd, with results shown for Surflex using the geometric optimization parameters (''-pgeom'') either alone or with ring flexibility enabled. For the Jain set, the proportion of cases in which the best pose of the 20 returned poses has rmsd <2.0 Å was ~90% under both parameter settings. For the top *scoring* pose, ring flexibility improved performance by 7% points (from 67 to 74%). For the Vertex set, performance was nearly as good with respect to the *best* pose (84 and 83% with and without ring flexibility, respectively). However, recognition of low-rmsd poses as being the highest scoring was significantly worse, with 54 and 50% success, respectively, at the 2.0 Å accuracy level. Inclusion of ring flexibility was a clear benefit, particularly in terms of improving choice of top scoring pose, and it was also a benefit in increasing the proportion of very low-rmsd solutions within the Vertex set. There was a slight enhancement in performance over Surflex Version 1.3 due to improvements in local optimization (data not shown).

Comparison to glide and GOLD: the effect of protein optimization

In the foregoing results, all protein atoms (both hydrogens and heavy atoms) were kept precisely in their original positions. In this section alone, movement of protein atom positions will be explored to investigate the effects on nominal docking accuracy. In Fig. 10, plots C and D, comparative data are shown for both Glide and GOLD. The comparative data came from docking on *exactly* the same protein structures (including the protonation states, tautomer choices, and proton positions) as well as exactly the same ligand starting coordinates. With respect to the best respect to best pose of the top 20 returned for Glide, GOLD, and Surflex with no ring search, performance is essentially indistinguishable. Surflex with the addition of ring search showed a significant improvement at low rmsd (~10% points over Glide), with a smaller advantage at higher rmsd. However, rmsd of top scoring pose was somewhat different, with Glide showing an advantage in success rates at rmsd from 2.0 to 5.0 Å. In the original study [23], additional experiments were carried out, showing that post-docking optimization of the final poses using the OPLS-AA force field improved GOLD's performance to equal that of Glide with respect to the quality of top scoring poses. The same procedure *did not* affect Glide's performance, since Glide employed OPLS-AA in the final stages of docking.

The protein preparation procedure used by Perola et al. included a step where protons on *both* the bound ligand and protein were optimized using the OPLS-AA force field. It

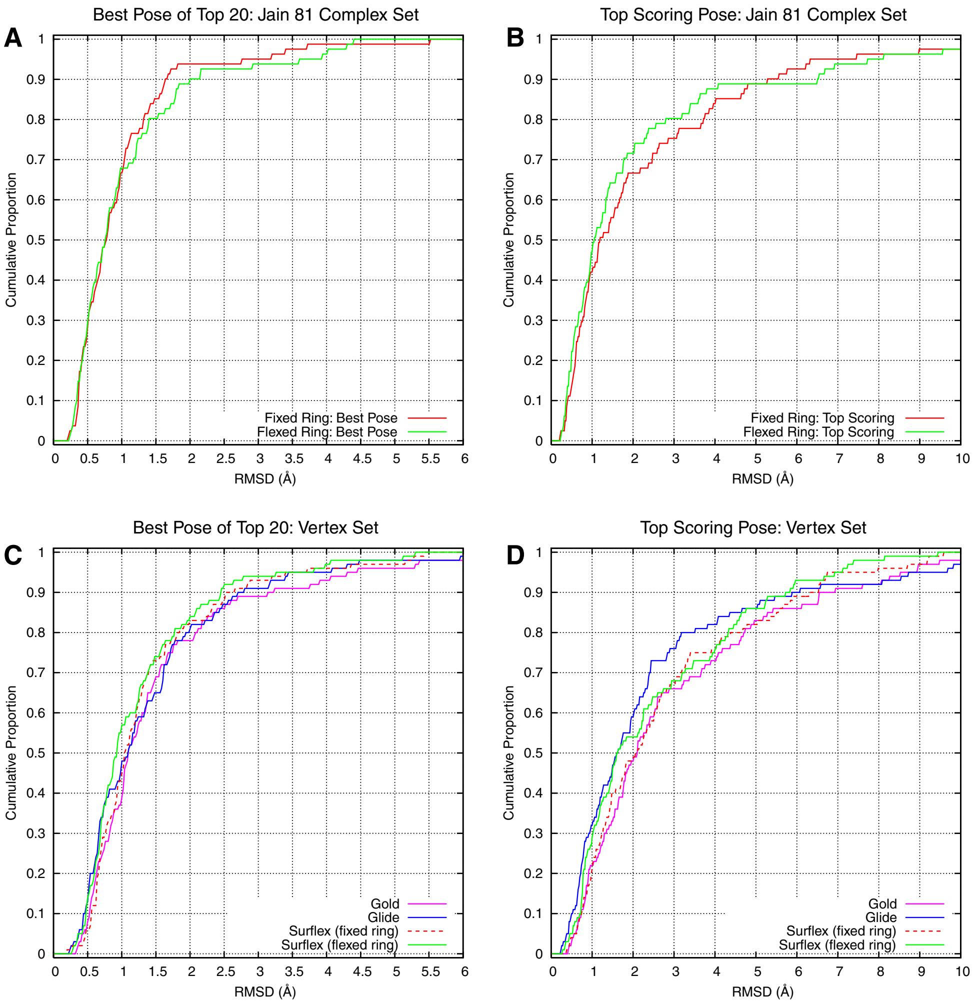

Fig. 10 Plots **A** and **B** show Surflex docking accuracy for the Jain 81 complex benchmark (best pose of the top 20 and top scoring pose, respectively). Plots **C** and **D** show comparative docking accuracy on the Vertex set. In both cases, Surflex's ring searching improved docking performance with respect to the top scoring pose, and in the Vertex set also appeared to improve performance with respect to the best pose by rmsd. All three programs performed similarly in terms of identifying a good pose within the top 20 returned. Glide and Surflex (with ring flexibility) appeared to show an advantage in terms of the proportion of highly accurate poses (rmsd < 1.5 Å). Glide appeared to exhibit a slight advantage in the moderate accuracy range (2–4 Å rmsd), with Surflex showing a modest advantage in terms of the proportion of poor dockings (>7 Å rmsd) chosen as the top scoring. The differences among programs were small within a sample size of 100 and may be explained best by aspects of sampling bias and protein preparation bias

is possible, then, that the reason that OPLS-AA optimization after docking improved the rmsd of the top scoring pose for GOLD was because the optimization procedure imparted a ‘‘memory’’ within the protein of the correct bound geometry of the ligand with respect to the *specific choice* of OPLS-AA. The degree of distortion of protein sites was modeled using optimized bound ligand strain as a surrogate value. Figure 11 shows the cumulative histograms of top scoring pose rmsd for all three dockers, separated by low strain and high-strain cases. In both the Surflex and GOLD cases, there was a roughly 30% point gap in performance at 2.0 Å rmsd, but for Glide this value was <10% points. In the low-strain cases, Surflex and GOLD yielded 70 and 64% correct at 2.0 Å, with Glide yielding 64% as well. The nominal Glide advantage was in the high-strain cases.

It is possible that the high-strain cases were simply more difficult and that OPLS-AA is a genuinely better solution in such cases. However, Fig. 12 shows the effects of re-optimization of the proteins' protons using a combined force field that used Surflex's scoring function for intermolecular contacts and the DREIDING force field for the protein's bonded and internal non-bonded terms. The new proton positions (denoted ‘‘Surflex protons'') did not deviate in magnitude any more than the original OPLS-AA protons did from the standard coordinates achieved with no knowledge of the bound ligand. The results of docking to the original proteins were rescored and re-ranked using the modified protein proton positions with local optimization. This procedure was analogous to the post-docking OPLS-AA optimization in the Perola et al. study [23]. In both the high- and low-strain cases (plots A and B), Surflex's top scoring pose performance increased by as much as 15% points. The net result (plot C) was indistinguishable performance between Surflex and Glide up to 4.0 Å rmsd, with slightly better performance for Surflex at higher deviations. Note that even larger increases in nominal performance for Surflex were obtained by allowing larger excursions of the protons (data not shown).

Surflex's performance in terms of docking accuracy appears to be at least as good as other methods for which

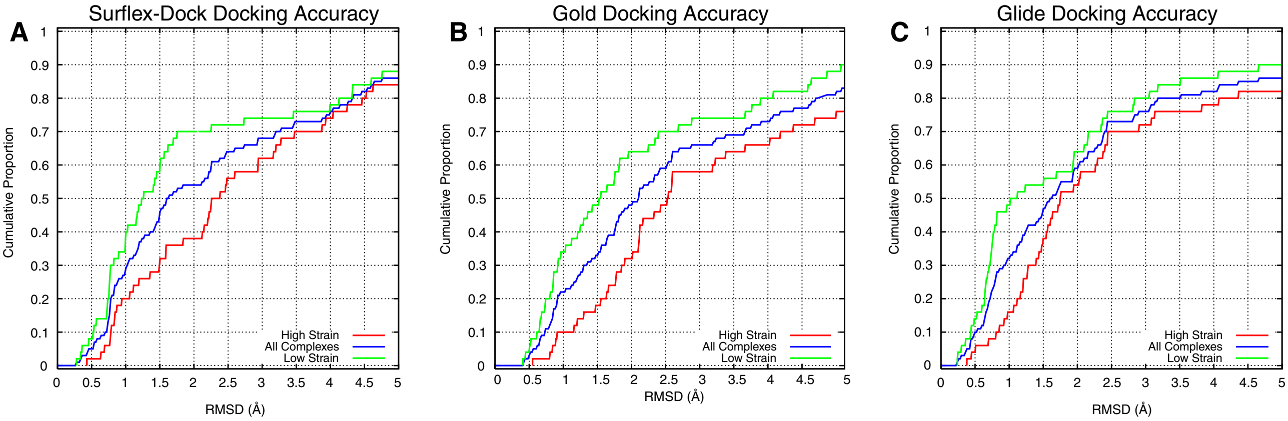

Fig. 11 The relationship between docking accuracy with respect to top scoring poses was strongly dependent on the strain energy of the bound ligand. In the Vertex set, the ligand and protein protons were optimized using the OPLS-AA force-field. Plots **A–C** show the cumulative histograms of top pose rmsd for Surflex, Gold, and Glide, respectively. In each plot, the *blue* curve depicts the cumulative histogram for all complexes, the *green* for complexes with relatively low-ligand strain (50 complexes with <70 kcal/mol strain), and the *red* for high-strain complexes (50 complexes with strain greater of 70–400 kcal/mol). Glide's results showed markedly less pronounced differences between high and low-strain cases. Both Surflex and Gold showed stronger effects. The results for Glide employed OPLS-AA for final scoring of docked poses, possibly biasing results since the proteins' hydrogens had a ''memory'' of the ideal proton positions that were compatible with the bound ligand with its protons optimized

comparative data exists, both with respect to identifying accurate poses within the top set returned as well as in recognizing which among them should score best. The addition of ring flexibility appears to give Surflex a slight advantage over other methods, but the proportional gains are small in the context of test sets of order 100 structures. The effects of inadvertent bias in protein preparation can be difficult to detect, but they can have significant effects that can modify conclusions about relative performance. This issue will be discussed further below.

Summary: docking accuracy

Generally, Surflex-Dock, Glide, and GOLD performed similarly when controlling for aspects of protein preparation. This is in agreement with the Kellenberger et al. study [19], in which those three methods were tested along with DOCK, FlexX, Fred, Slide, and QXP. In that study, Surflex-Dock, Glide, and GOLD all yielded success rates at roughly 55% in returning the top scoring pose within 2.0 Å rmsd, with FlexX somewhat lower, and a number of pro-

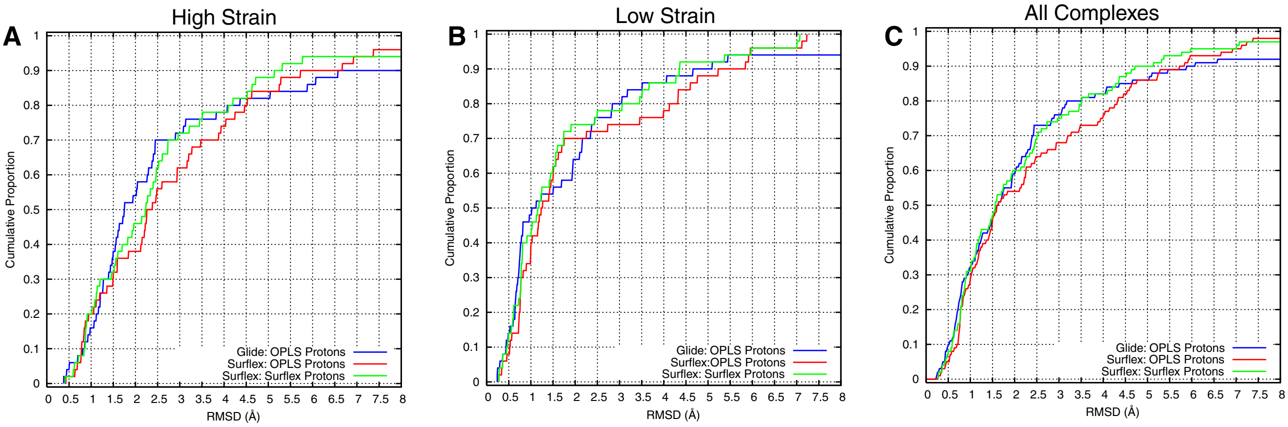

Fig. 12 Optimization of the protein protons using a force-field that combined DREIDING-type covalent forces with Surflex's scoring function (''Surflex Protons'') for intermolecular contacts significantly improved Surflex's performance for docking accuracy in terms of choosing a correct pose from among the top 20 returned. The final poses from the Surflex docking protocol employing the original version of the proteins were rescored with local optimization using the modified protein coordinates. Performance on both the high- and low-strain complex sets improved significantly, with slightly more improvement in the former set. Overall Surflex performance with the modified proteins was statistically indistinguishable from Glide's performance with the OPLS-AA protons at low rmsd, with slightly better performance at the high end

grams still lower. The docking protocols employed in that study were selected to yield roughly equivalent timings for the different dockers; the effects of that choice on the different docking systems was variable, and for programs that exhibited poor performance this may have been a dominating factor.

Not surprisingly, ring conformational search within Surflex yielded benefits for accurate docking, producing gains particularly with respect to the proportion of very low-rmsd solutions among the set of top scoring poses returned. The surprising finding here has been that very subtle changes in protein structure can have substantial effects on the detailed ranking of returned poses. Surflex-Dock can identify excellent poses ~85–95% of the time based on the Jain and Vertex benchmarks, but variability in correctly identifying the correct pose as top scoring ranged from 50 to 75%. Small changes in protein proton positions were capable of producing 10 percentage point shifts in performance.

Discussion

The most important finding of this study is that more effective exploration of the reasonable space of energetically accessible conformations of ligands leads to better performance of a docking system both with respect to screening experiments and with respect to geometric docking accuracy. This should not be surprising, but the practical significance of accessing that space with computational approaches that are not burdensome is very important. With the computational speedups afforded by highly efficient local optimization, Surflex-Dock Version 2.1, employing procedures for pre-docking minimization *and* post-docking all-atom in-pocket optimization, is just as fast as the previous version *without* those steps. A general procedure for ring conformation exploration provides the user with a fast and robust method that does not rely on pre-computed ring libraries. Expansion of the ligand search space using these tools led directly to improvements in both screening utility and docking accuracy. These enhancements, coupled with added functionality for exploiting knowledge of protein ligand interactions, should collectively yield improvements in application of Surflex-Dock in real-world situations.

Docking for screening

The data on screening performance here agrees in an interesting qualitative way with the observations of other researchers [22, 24] in that it does not appear that highly accurate docking in the sense of correctly identified binding modes is required for effective screening performance. For example, in the HVR case (see Fig. 6) while ring flexibility is important to achieving correct dockings, it is yields only a marginal improvement in screening efficiency (Table 2). It is possible that this effect does not have a physically meaningful interpretation, but it might. If so, a reasonable interpretation is that a significant proportion of true ligands of a binding site have multiple binding modes that can have an impact on the propensity of binding. In a kinetic sense, this would provide an energetic basin from which a partially organized binding event could transition into its final form. In the HVR case depicted in Fig. 6, it is easy to imagine that once the ‘‘arms’’ of the ligand have reached partially into the protease active site (Panel B), the central ring could shift into its proper conformation, accompanied by a rotation of the ligand, as in the crystallographically determined structure.

If this is a real effect, it should be the case that true ligands of a binding site will have a greater number of different high-scoring binding modes than non-ligands. In the J&J set, considering the *mean* score of the ten returned poses (instead of the maximum) for each docked molecule improved screening enrichment in three of four cases. In the Pham set, enrichment was marginally better in 50% more cases than it was marginally worse. Neither of these effects was large enough to make a statistical argument, but the area deserves some additional investigation. The docking protocol for screening has been optimized to identify the maximal scoring pose of an input ligand, not a diverse set of high-scoring poses, which would be required to properly consider whether this could be an advantageous strategy.

The construction of decoy sets and screening benchmarks has become controversial as the use of benchmarking has become widespread. Rognan’s ACD decoy set [18], which was among the first widely used, has been characterized by others as being more hydrophobic than one would expect of drug-like screening molecules [46]. Groups from pharmaceutical companies have made use of proprietary internal libraries for the construction of decoy sets [22, 23], and academic groups such as Miteva et al. [20] have made use of large and carefully selected decoy sets of ACD compounds. The recent paper by Huang et al. [46] introduced a new screening benchmark. Specific comparisons were made of the ZINC decoys used extensively here, the Rognan decoys, a new set of MDDR decoys, and a new set of decoys (called DUD) specifically constructed to match the chemical properties (but not the topological ones) of the specific true ligands of 40 different target proteins. Their results with DOCK Version 3.5.54 showed a marked difference in performance depending on which decoy set was used, with the least challenging being the Rognan set and the MDDR set. The ZINC decoy set used here and the DUD decoy set had properties most similar to the true ligands in terms of molecular weight, log P, number of donors and acceptors, and number of rotatable bonds. The DUD set was more challenging in the reported experiments with DOCK, with the ZINC decoy set being the next most challenging [46]. The study of Miteva et al. [20] employed a hybrid protocol using FRED with Surflex-Dock with tuned parameters on a large screen. The results presented here may reduce the need for parameter tuning, since the new Surflex-Dock procedures normalize some aspects of variation in screening.

The results of Huang et al. [46] stand in contrast with what has been reported here. Whereas their experiments showed a strong dependence on decoy set choice, our results here comparing performance of Surflex-Dock using either the Rognan or ZINC sets were very similar on the Pham benchmark, and results on the J&J benchmark with MDDR decoys were quantitatively similar as well. There are very significant differences in the DOCK approach compared with the Surflex-Dock approach, especially with respect to scoring. In particular, treatment of ligand desolvation energies and polar interactions are very different, and these might explain differential sensitivity to decoy set composition with respect to hydrophobicity and number of polar features per ligand. Further tests are planned for Surflex-Dock, making use of different decoys and different targets, but the results presented here offer some reason for optimism regarding robustness across a diverse set of conditions.

Docking to inform rational design

As mentioned above, one of the key practical aspects in applying docking systems in drug design is in the situation where one is exploring a chemical structural space of analogs. In these cases, where it is reasonable to posit that a particular substructure will remain largely stationary in an active site, making direct use of that knowledge to constrain the search space offers advantages in terms of workflow, speed, and direct comparison of different analogs. For example, in a case like methotrexate docking to DHFR, placement of the diaminopyrimidine improves the workflow in terms of both speed and direct comparability of the scores of analogs. In considering diaminopyrimidine analogs, while docking each de novo may seem like a more ‘‘correct’’ and unbiased approach, making use of reliable knowledge is more sensible. It is a safe bet that the diaminopyrimidine will bind in the conventional manner, and factoring that moiety out of any influence on scoring by restraining it may lead to improved rankings of synthetic analogs.

Frequently, however, the task at hand may involve a very different ligand structure from anything for which a co-crystal structure has been solved. It is this case for which the standard docking accuracy benchmark is used to document the expected behavior of a particular approach. The problem is that the standard construction of such benchmarks makes use of co-crystal structures themselves. Both the Jain and Vertex benchmarks employed in this study shared this feature. For proteins with the potential for significant active site atomic motions, performance on such a test is likely to be predictive only of a docker's performance on analogs chemically very similar to the ligand that influenced the shape of the protein binding site in the experimental determination of structure. Such benchmarks represent an artificially easy case for dockers; it is therefore a mistake to further bias the protein structure using knowledge of the bound ligand. Choosing a protonation state or a tautomer is probably reasonable, with rotamer choices for hydroxyls and thiols probably less so. However, influencing a bond length, bond angle, or bending constraint in the protein is unwise and can bias a test in specific ways toward a particular method.

Recall the discussion of Fig. 12. Very small movements of protons alone that were guided by the scoring function to be tested were able to increase performance of that particular method. Surflex-Dock using protein active site protons that were optimized for Surflex's scoring function performed equivalently to Glide using protein active site protons that were optimized using Glide's scoring approach (OPLS-AA). However, neither result informs a user of how either program will function on the problem that is important, since the optimized protein structure will not have the same effect on a *new* ligand. Figure 13 illustrates this problem further, by not only optimizing the protons but also the active site heavy atoms of the protein. The plot shows a systematic increase in the proportion of correct top scoring poses for Surflex across the full range of rmsd, now at 65% success for 2 Å (compared with 54% with the original Vertex structures). The atomic motions within the protein were small and reasonable (the example shown was typical), but they required knowledge of the experimentally correct answer. This is not available in a standard modeling experiment. Consequently, this procedure is not recommended in either practical applications of docking or in investigations of algorithm performance.

While the procedure itself is flawed from the perspective of evaluation, this line of investigation does suggest a way forward. Within the top 20 poses returned by Surflex-Dock, the expectation from the Jain and Vertex sets (with no bias in protein preparation) is that roughly 85% of the time a good pose will exist (rmsd < 2 Å). It should be computationally feasible to optimize the local protein structure for each returned ligand pose for roughly the same cost as for optimizing the ligand itself. The key is that only a limited number of bonded and non-bonded protein terms need to be computed. The problem then becomes one of

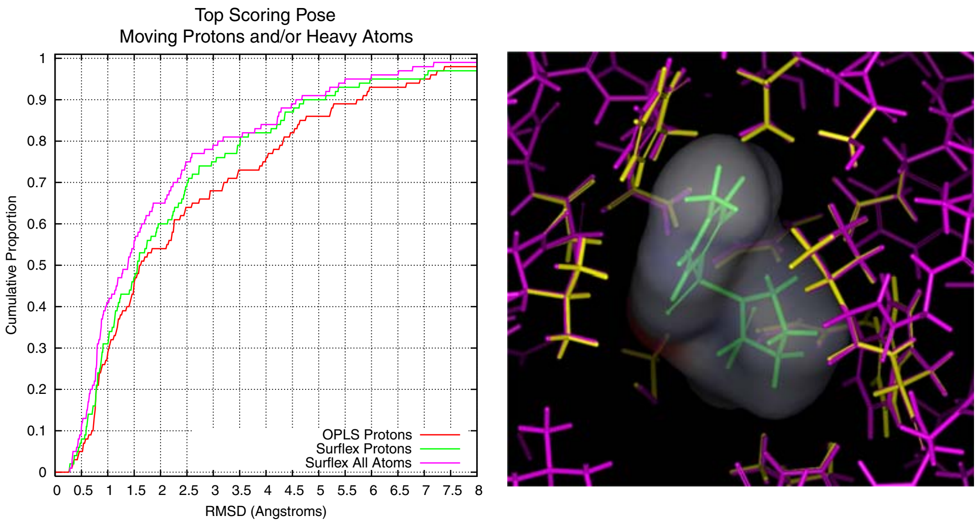

Fig. 13 Optimization of all protein active site atoms (‘‘Surflex All Atoms’’) significantly improved Surflex’s performance over optimizing just the hydrogen atoms (‘‘Surflex Protons). As with Fig. 12, the final poses from the previous docking protocol were rescored with local optimization using modified protein coordinates. While this procedure is clearly biased and incorrect from the perspective of evaluating docking accuracy (see discussion), it motivates the idea of making use of protein atom movement in cases where detailed scoring differences are important. At right is a representative example of the amount of movement (PDB complex 1F4E, thymidylate synthase complexed with tosyl-d-proline). Most of the protein atoms did not move from the original coordinates (*purple*), but the indole ring shifted ~0.7 Å (*yellow*) to fully accommodate the proline and other constraints of the binding pocket

understanding the quantitative relationship between the changes in energy observed within the ligand, within the protein, and between the two in a manner that *improves* the detailed scoring of the best ligand poses. This is an area of current research and is challenging. Small changes in atomic positions within a protein lead to moderate changes in the magnitude of internal protein energies (large numbers), and small differences between these large numbers determine the ‘‘winning’’ pose.

Fairness of comparisons

In the comparison among different versions of Surflex-Dock and different parameter switches, the relative performance changes are due solely to uniform application of standardized fully automatic deterministic computations. The conclusions discussed above with respect to the utility of ligand minimization, ring search, and post-docking all-atom minimization should translate to practical improvements. In the comparisons with other methods, great care has been taken to make fair comparisons and to be explicit about parameter choices and computational protocols. In the J&J and Vertex sets, the precise ligand structures used in the original papers were used unmodified in all Surflex experiments. For the J&J proteins, hydrogens were added but were not subjected to force-field optimization. For the Vertex proteins, except where discussion was explicit about protein optimization, the protein structures were used completely unmodified.

Screening utility tests involve a single docking invocation to test many ligands against one protein structure, and these are therefore relatively difficult to bias toward or against particular methods. Improvements in performance will tend to derive from algorithmic differences among different methods tested. For the J&J screening utility assessments, both Surflex-Dock and Glide had an advantage with respect to ligand optimization beyond internal coordinates compared with the other methods, and their performance reflected that advantage. Surflex’s superior performance in the direct comparisons with Glide parallels the observations of other reports [18, 19], and by the methods developers themselves on common benchmarks [13, 16, 21, 29]. However, the number of distinct protein cases for which there are directly comparable performance data is small. Unfortunately, protein and ligand structural data for screening utility evaluation from reports of Glide’s development are not available from the authors [13, 47].

In contrast to screening utility, it is easy to introduce bias into docking accuracy tests, since they consist of many pairs of protein/ligand complexes, so each application of a docker may be influenced by some bias in its input (one docking invocation per input complex). Figure 14 shows the direct comparison of Surflex’s performance in two cases. The first made use of the Vertex protein structures for docking and used the crystallographic ligand coordinates to measure rmsd. The second made use of ‘‘cooked’’ proteins and ligands. The ligand and protein active site

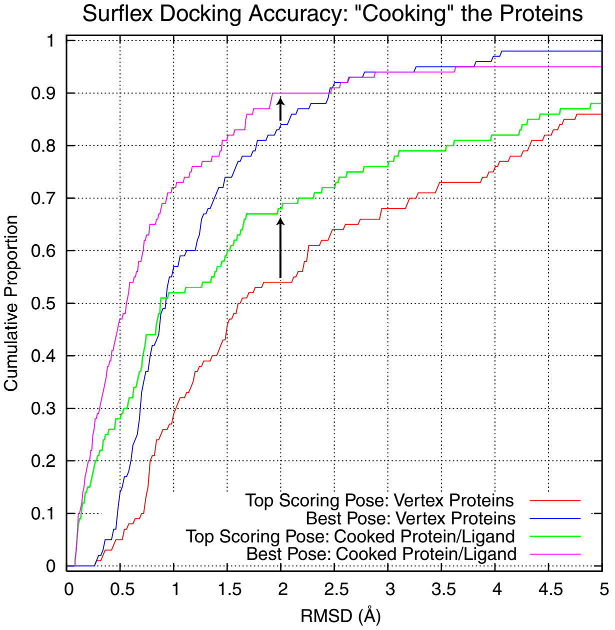

Fig. 14 By optimizing the protein and ligand atoms of each complex, including both the protons and heavy atoms close to the active site, it is possible to greatly improve nominal docking accuracy, using precisley the same input ligand structures and docking procedures. Here, the modified proteins were used as targets of docking, and the modified crystallographic ligand positions were used as the gold standard from which to measure rms deviation for top scoring pose and best pose of the top 20 returned. Performance increases were particularly large in terms of reported success at very low rmsd. ‘‘Cooked’’ docking success rates at the standard threshold of 2.0 Å rmsd increased from 84 to 90% for best returned pose and from 54 to 69% for top scoring pose (*black arrows*). Average rmsd values were also biased

atoms were optimized using Surflex’s scoring function combined with the DREIDING force field (as in Fig. 13). Following a full re-docking (not just rescoring as in the experiments above), the optimized bound ligand coordinates were used from which to measure rmsd. While the protein atom motions were reasonable, and while the changes in reference ligand coordinates were small, the performance of Surflex increased markedly. The numerical increases are remarkable in magnitude, and this is a fair comparison for the purpose of quantifying the effects of optimizing the complexes and of measuring rmsd from optimized ligand coordinates. However, it would *not* be fair or informative to compare Surflex’s ‘‘cooked’’ results with another docker’s using ‘‘uncooked’’ proteins and reference ligands.

The optimization procedure imparts a memory within the protein active site of the joint arrangement of protein and ligand *specifically* at an optimum according to the preferred scoring function. Within the optimized complex, any movement of an atom on the ligand or protein results in a less optimal score. So, presuming that a docking procedure is able to generate *some* solution that is close to correct, the protein’s configuration guarantees a basin of attraction to the right answer. The optimized bound reference ligand coordinates share the memory of the joint optimum. By redefining the location of the right answer to the coordinates of the optimized bound ligand, the procedure artificially reduces the measured rmsd of the specific configuration that is *preferred* by the scoring function under study. Essentially arbitrary levels of nominal improvement are possible by making use of such a protein optimization procedure coupled with a redefinition of the position from which one measures ‘‘correctness.’’

This somewhat subtle intrinsic bias is not universally understood. Glide’s developers, in their paper introducing the method, used precisely this type of comparison to make a case for improvements in docking accuracy over multiple competing methods [32]. Protein atom positions (both protons and heavy atoms) were optimized *and* results were computed relative to the coordinates of the optimized reference ligand. The paper reported an improvements in mean rmsd over competing methods of 0.5–1.6 Å and also reported advantages in success rates at the 1.0 Å rmsd threshold of more than 10% points. In the controlled experiment with Surflex-Dock that yielded the results in Fig. 14, the improvement due *solely* to the effects of biased benchmark construction were 0.4 Å in mean rmsd, with a 22 percentage point increase in proportion of successes at the 1.0 Å rmsd threshold. It is not clear whether the degree of protein and ligand perturbation that gave rise to the results shown here was comparable to that in the Glide report, but the data sets were not available from the authors to establish the comparison directly with the proteins used in their study [32].

These results suggest that protein optimization prior to docking a cognate ligand serve primarily to make an already artificial experiment less challenging. The recent paper by Hartshorn et al. makes this point as well as part of a broader discussion of the proper selection of complexes on which to carry out docking algorithm testing [48]. They have constructed a set in which all structures are very high-quality. They made use of crystallographic structure factors to assess the experimental binding mode of the ligands and to verify the degree to which the electron density accounted for all parts of the ligands. They excluded complexes in which ligands made hard clashes with the protein or contacted physiologically irrelevant crystallographically related subunits. In such cases, structures will have little uncertainty in the positions of the heavy atoms of either the protein or the ligand, and the question of whether a docker can recover the correct pose is relevant, made more so by limiting the use of knowledge of the bound ligand state in preparing the protein structure. Further optimization of protein atom positions *with knowledge of the bound ligands* leads only to bias in the experimental assessment of accuracy. They list three cases within the Vertex benchmark (1cet, 1nhu, and 1nhv) where they were not able to find good support for the positioning of the ligands. In each of these cases, Surflex-Dock was unable to identify a solution that matched the nominal experimental ligand coordinates to within 2.0 Å rmsd, but this probably reflects limitations in the structures more than limitations in the methodology.

If the goal is to improve docking performance for the real-world application of identifying the configuration of a *new* ligand bound to a pre-existing protein structure, the most sensible strategy is to work first on a set of high-quality complexes without any a priori bias from knowledge of the bound ligand. Given that one shows solid performance without protein atom movement, convincingly demonstrating that protein atom optimization at the end of the docking process can improve the ranking of the top scoring poses would be an important step forward that would likely yield real-world improvements in the practical application of docking methods.

Conclusions

The studies reported here have shown that a generalization of search purely within ligand pose space (defined narrowly as translational, rotational, and torsional parameters) to allow access to the broader Cartesian space of accessible ligand configurations makes a substantial impact on the practical effectiveness of docking. Exploration of the effects of such generalization to the atoms of the protein active site gives hope that significant improvements, particularly in docking accuracy, should be possible and should not necessarily require combinatorial exploration of protein configurational space simultaneously with ligand configurational space. It may be possible to employ local optimization of protein active site atoms, after docking, to obtain these benefits without incurring a burdensome computational cost. This is an area of current investigation.

Acknowledgments The author gratefully acknowledges NIH for partial funding of the work (grant GM070481). Dr. Jain is indebted to Max Cummings for sharing the J&J set, Emanuele Perola for sharing the Vertex Set and performance data, to Bob Clark and Essam Metwally for identifying structural inconsistencies within the original Pham set, and to Ann Cleves for comments on the manuscript. Dr. Jain has a financial interest in BioPharmics LLC, a biotechnology company whose main focus is in the development of methods for computational modeling in drug discovery. Tripos Inc., has exclusive commercial distribution rights for Surflex-Dock, licensed from BioPharmics LLC.

References

1. Walters PW, Stahl MT, Murcko MA (1998) Drug Discov Today 3:160–178
2. Jain AN (1996) J Comput Aided Mol Des 10:427–440
3. Bohm HJ (1994) J Comput Aided Mol Des 8:243–256
4. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) J Comput Aided Mol Des 11:425–445
5. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V (1999) J Med Chem 42:4650–4658
6. Wang R, Liu L, Lai L, Tang Y (1998) J Mol Model 4:379–384
7. Muegge I, Martin YC (1999) J Med Chem 42:791–804
8. Gohlke H, Hendlich M, Klebe G (2000) J Mol Biol 295:337–356
9. Welch W, Ruppert J, Jain AN (1996) Chem Biol 3:449–462
10. Goodsell DS, Morris GM, Olson AJ (1996) J Mol Recognit 9:1–5
11. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727–748
12. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261:470–489
13. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) J Med Chem 47:1750–1759
14. Schulz-Gasch T, Stahl M (2003) J Mol Model (Online) 9:47–57
15. Zavodszky MI, Sanschagrin PC, Korde RS, Kuhn LA (2002) J Comput Aided Mol Des 16:883–902
16. Jain AN (2003) J Med Chem 46:499–511
17. Pham TA, Jain AN (2006) J Med Chem 49:5856–5868
18. Bissantz C, Folkers G, Rognan D (2000) J Med Chem 43:4759–4767
19. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Proteins 57:225–242
20. Miteva MA, Lee WH, Montes MO, Villoutreix BO (2005) J Med Chem 48:6012–6022
21. Jain AN (2004) Curr Opin Drug Discov Devel 7:396–403
22. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912–5931
23. Perola E, Walters WP, Charifson PS (2004) Proteins 56:235–249
24. Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP (2005) J Med Chem 48:962–976
25. Gasteiger J, Sadowski J, Schur J, Selzer P, Steinhauer L, Steinhauer V (1996) J Chem Inf Comput Sci 36:1030–1037
26. Perkins E, Sun D, Nguyen A, Tulac S, Francesco M, Tavana H, Nguyen H, Tugendreich S, Barthmaier P, Couto J, Yeh E, Thode S, Jarnagin K, Jain AN, Morgans D, Melese T (2001) Cancer Res 61:4175–4183
27. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK (2002) J Med Chem 45:2213–2221
28. Wang R, Fang X, Lu Y, Wang S (2004) J Med Chem 47:2977–2980
29. Jain AN (2006) Curr Protein Pept Sci 7:407–420
30. Pham TA, Jain AN (2006) J Med Chem 49:5856–5868
31. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Proteins 52:609–623
32. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) J Med Chem 47:1739–1749
33. Wang R, Lai L, Wang S (2002) J Comput Aided Mol Des 16:11–26
34. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW (2000) J Comput Aided Mol Des 14:731–751
35. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) J Comput Chem 19:1639–1662

36. Jain AN, Dietterich TG, Lathrop RH, Chapman D, Critchlow RE Jr, Bauer BE, Webster TA, Lozano-Perez T (1994) J Comput Aided Mol Des 8:635–652
37. Jain AN, Koile K, Chapman D (1994) J Med Chem 37:2315–2327
38. Jain AN, Harris NL, Park JY (1995) J Med Chem 38:1295–1308
39. Dietterich TG, Lathrop RH, Lozano-Perez T (1997) Artif Intell 89:31–71
40. Jain AN (2000) J Comput Aided Mol Des 14:199–213
41. Fletcher R (1987) Practical methods of optimization, 2nd edn. Wiley-Interscience, New York
42. Mayo SL, Olafson BD, Goddard WA (1990) J Phys Chem 94:8897–8909
43. Jain AN (2004) J Med Chem 47:947–961
44. Lei S, Smith MR (2003) IEEE Trans Softw Eng 29:996–1004
45. Hawkins PC, Skillman AG, Nicholls A (2007) J Med Chem 50:74–82
46. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789–6801
47. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT (2006) J Med Chem 49:6177–6196
48. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW (2007) J Med Chem 50:726–741