JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

# Improving Docking Results via Reranking of Ensembles of Ligand Poses in Multiple X-ray Protein Conformations with MM-GBSA

P. A. Greenidge,*[,†] C. Kramer,[‡] J.-C. Mozziconacci,[§] and W. Sherman[∥]

[†]Novartis Institutes for Biomedical Research, Novartis Pharma AG, Forum 1, Novartis Campus, CH 4056 Basel, Basel-Stadt, Switzerland

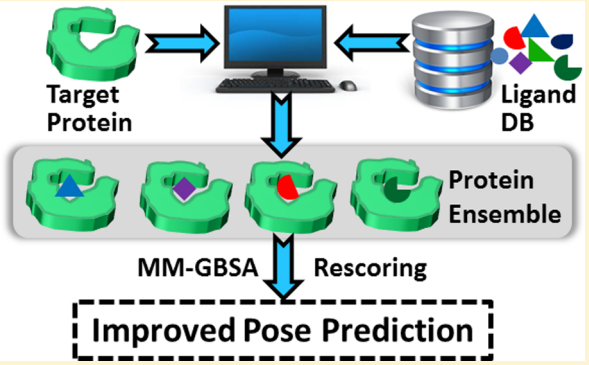
[‡]Center for Chemistry and Biomedicine, Institute for General, Inorganic and Theoretical Chemistry, University of Innsbruck, Innrain 82, 6020 Innsbruck, Tyrol, Austria

[§]Schrödinger GmbH, Dynamostrasse 13, 68165 Mannheim, Germany

[∥]Schrödinger Inc., 120 West 45th Street, New York, New York 10036, United States

Supporting Information

**ABSTRACT:** There is a tendency in the literature to be critical of scoring functions when docking programs perform poorly. The assumption is that existing scoring functions need to be enhanced or new ones developed in order to improve the performance of docking programs for tasks such as pose prediction and virtual screening. However, failures can result from either sampling or scoring (or a combination of the two), although less emphasis tends to be given to the former. In this work, we use the programs GOLD and Glide on a high-quality data set to explore whether failures in pose prediction and binding affinity estimation can be attributable more to sampling or scoring. We show that identification of the correct pose (docking power) can be improved by incorporating ligand strain into the scoring function or rescoring an ensemble of diverse docking poses with MM-GBSA in a postprocessing step. We explore the use of nondefault docking settings and find that enhancing ligand sampling also improves docking power, again suggesting that sampling is more limiting than scoring for the docking programs investigated in this work. In cross-docking calculations (docking a ligand to a noncognate receptor structure) we observe a significant reduction in the accuracy of pose ranking, as expected and has been reported by others; however, we demonstrate that these alternate poses may in fact be more complementary between the ligand and the rigid receptor conformation, emphasizing that treating the receptor rigidly is an artificial constraint on the docking problem. We simulate protein flexibility by the use of multiple crystallographic conformations of a protein and demonstrate that docking results can be improved with this level of protein sampling. This work indicates the need for better sampling in docking programs, especially for the receptor. This study also highlights the variable descriptive value of RMSD as the sole arbiter of pose replication quality. It is shown that ligand poses within 2 Å of the crystallographic one can show dramatic differences in calculated relative protein−ligand energies. MM-GBSA rescoring of distinct poses overcomes some of the sensitivities of pose ranking experienced by the docking scoring functions due to protein preparation and binding site definition.




## INTRODUCTION

Docking programs aim to correctly position ligands (conformation and orientation) into a protein binding site;[1,2] scoring functions aim to correctly predict the complementarity and/or biological activity of the ligand to the binding site via the evaluation of molecular interactions.[2,3] For a given ligand, the best pose is typically determined by the best score, although some docking programs use different scoring functions for pose selection and binding energy estimation. Given these two primary objectives (pose prediction and scoring), a practical question arises: what is the main limitation in current docking programs, finding the correct pose or predicting the correct score. In the work presented here, we address this question through a carefully curated subset of protein−ligand complexes from the PDB that contains high-resolution X-ray diffraction data for the ligands. As such, we are able to decouple the sampling and scoring problems, assuming the experimental crystal structure of the protein−ligand complex can be considered as "accurate" for the pose. Since the docking/scoring field has been intensively explored in the past, we first give an overview of the field and review the literature for other studies that have addressed aspects of the key question considered here.

In virtual screening, the docking score is important because it determines which compounds to keep as putative hits and

Received: April 11, 2014

   dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

which to reject as inactives. As such, for virtual screening applications one could argue that predicting the correct score is more important than having the correct pose. However, it is not clear whether one can consistently obtain accurate scores from incorrect poses.[4] In contrast to virtual screening applications, structure-based drug design (SBDD) requires a correct pose in order to understand the key molecular recognition motifs and to investigate the consequences of structural modifications. In this case, one may not be interested so much in the score but rather in the pose. Again, it is not clear whether can one consistently generate correct poses from an inaccurate scoring function. Finally, in real-world applications we do not know the exact right conformation of the receptor to bind the ligand of interest, so finding the correct pose can depend on both ligand and protein sampling. Of course, an accurate description of protein−ligand binding must correctly predict both the pose and the score, which should be the ultimate aim of a docking and scoring workflow.

A central objective in our field is to develop docking programs that can address both aspects of drug discovery (virtual screening and SBDD), and there are tests to assess the degree to which different methods can achieve this goal. In virtual screening, one can screen a database of known binders (actives) and nonbinders (decoys) to see how they improve the fraction of actives relative to decoys in the early part of the hit list (enrichment).[5−7] In the quest for accurate poses for SBDD, X-ray structures of known binders can be used to analyze how the docking program handles the pose by comparing to the experimental pose using a metric like RMSD of the heavy atoms or real space R-factor (RSR) to account for the reproduction of the ligand electron density.[8]

In recent years there has been a trend in the docking community toward "rescoring", i.e., using a more computationally intensive and assumed to be more accurate scoring function on a subset of the poses that have been generated with faster, less accurate docking methods.[9−13] If the rescoring is done without alteration to the coordinates (e.g., local minimization or conformational sampling), the new score will inevitably reflect the original pose and not necessarily the quality of the underlying scoring function.[14] An inherent danger in this approach is that actives could be missed even with a good scoring function because an accurate pose is not retained for the rescoring process. Eventually, the sensitivity of the scoring functions to the coordinates is of essential importance.

Poor results in docking are often blamed on inadequate scoring functions,[2,15] but apart from obvious problems like how best to account for water-mediated interactions[16] and poor parameters for specific metal interactions, the scoring functions should be able to account for many of the important aspects of the molecular recognition (vdW, electrostatics, H-bonds, ligand strain, etc.). As such, it could be that many of the existing scoring functions would produce much better results if the underlying sampling (for the ligand, protein, and waters) were more extensive. The literature has many studies that compare the suitability of different docking programs for reproducing poses of cocrystallized ligands,[17−20] cross-docking (docking of a ligand in a binding site other than the one in which it was cocrystallized),[21] or virtual screening enrichments.[22−24] However, as discussed by several authors, "comparing protein-ligand docking programs is difficult" for several reasons.[25,26] Velec et al.[27] have, like many others, discussed the interconnectedness between pose generation and scoring. They concluded that ligand sampling is indeed the major challenge facing docking programs, but they appear to be in the minority with their opinion. They find that the widely accepted RMSD of 2 Å[28] between the heavy atoms of the native and docked pose as defining a good pose to be too generous for accurate scoring, as has been suggested by others as well.[29,30] Jain[31] and Corbeil and Moitessier[32] have found that enhanced conformational sampling, especially with respect to saturated rings, improves pose prediciton. The different ways for treating saturated rings has been an issue in comparisons between the performance of different docking programs.[25]

Kitchen et al.[2] have voiced an opinion more typical of the majority of our field when stating that robust and accurate docking algorithms are hindered by the imperfections of available scoring functions typically used for docking, particularly as they tend to neglect entropic and solvation effects.[5,6,26,33] A benchmark study involving 195 diverse protein−ligand complexes and 16 scoring functions has previously been performed by Cheng et al.,[34] where they found that no single scoring function consistently outperforms others but a consensus scoring approach could be applied to improve the results over single scoring functions. Similarly, Plewczynski et al.[35] used a data set of 1300 complexes to perform self-docking with seven docking programs. In addition, a critical analysis of 10 docking programs with respect to cross-docking has been carried out by Warren et al.,[15] where the working group consisted of experts with respect to a particular protein target and software. The authors of these studies concluded that scoring functions are unable to consistently identify the pose closest to the crystal structure conformation.

In the work presented here, we focus on the use of the docking programs GOLD[16,36,37] and Glide[22,38] and the available scoring functions within those programs. Based on previous studies, the different scoring functions contained in both GOLD and Glide have been determined to be among the best scoring functions, thereby making them suitable for this study. Finally, the aim of this study is not to compare docking programs *per se* but to carefully dissect the origin of docking failures and propose possible improvements. In accordance with Li et al.[39] and Cheng et al.[34] we consider two criteria to determine the success of a docking program. "Docking power" is the ability of the scoring function to recognize the pose closest to the experimental one by giving it the best score, while "scoring power" is the ability of the scoring function to correlate with experimental binding affinities. Previous studies concluded that there is not a single scoring function that is universally applicable for all types of molecules and protein families. Plewczynski et al.[35] found that the ability of scoring functions to rank-order binding modes was poor. The poses within 2 Å of the experimental ones (self-docking) were given the best scores in only 60% and 50% of cases for Goldscore and Glide SP, respectively, suggesting a need for improved pose ranking. Some efforts to improve docking power have focused on a consensus approach;[40−43] however, it is unpredictable which combination of scoring functions will give the best results for a given problem and does not directly address the possibility that scoring is being performed on incorrect poses.

In this work we compare the scoring power of five scoring functions on a data set of 855 protein−ligand complexes. The data set consists of multiple protein families and diverse ligands, with a broad representation of size, polarity, and functionality. Our data set[44] is sufficiently large to give a statistically significant interpretable comparison between the scoring functions with experimental binding data. For example, the


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

95% confidence ranges of a correlation coefficient of $R^2 = 0.63$ (that being observed in the previous work) based on 20, 100, and 855 samples are 0.29−0.84, 0.50−0.73, and 0.59−0.67, respectively. As such, to obtain confidence in the $R^2$ value to better than 0.1 units requires well over 100 samples.

Next, we investigate the docking power of the scoring functions on a high-resolution subset of these complexes in the context of both self- and cross-docking. What underlies our study is the belief that if the correct poses for multiple ligands can be consistently identified, then a superior scoring function in a postscoring step can be used to obtain a better correlation with binding affinity. As such, docking power and pose ranking are the primary aim of this study (not scoring power). In this part of the study we select a subset of 20 complexes from the full set of 855 (Table 1) that corresponds to the intersection between the 855 complexes described above and the "Iridium" data set,[45] which comprises a set of carefully curated protein−ligand complexes designed to validate and improve docking programs. This data set has been divided into three categories of trustworthiness based on various structural assessment criteria[45] — the categories are "highly" (HT), "mildly" (MT), and "nontrustworthy" (NT). For instance Iridium-HT membership requires the ligand to have complete electron density and an occupancy of 1 with no alternate conformations of the side chains of active site residues in contact with the ligand. These requirements are relaxed for the other two membership categories (MT and NT).

**Table 1. 20 Complexes with Associated Information about the Quality of the Ligand Density[45] Included in the Previous Greenidge et al. Study[44a]**

| PDB ID | trust category | target protein | exptl $pK_d$ | $\Delta pK_d$ (MM-GBSA) |
|---|---|---|---|---|
| 1ai5 | high | penicillin amidohydrolase | 3.72 | +0.6 |
| 1aj7 | nontrustworthy | FAB antibody | 3.87 | +1.7 |
| 1br6 | high | ricin | 3.22 | +1.5 |
| 1ezq | high | factor Xa | 9.05 | −1.2 |
| 1f0u | high | bovine trypsin | 7.16 | −0.3 |
| 1fcx | high | human retinoic acid receptor gamma-1 (hRAR) | 7.19 | +2.2 |
| 1fcz | high | human retinoic acid receptor gamma-1 (hRAR) | 9.22 | +0.2 |
| 1fh8 | high | beta-1,4-xylanase | 6.89 | −1.0 |
| 1fh9 | high | beta-1,4-xylanase | 6.43 | +0.1 |
| 1fhd | high | beta-1,4-xylanase | 6.82 | −0.6 |
| 1fl3 | high | blue fluorescent antibody | 6.80 | +0.1 |
| 1h1p | high | CDK2 kinase | 4.92 | +0.5 |
| 1hls | high | CDK2 kinase | 8.22 | −1.9 |
| 1l2s | high | beta-lactamase | 4.59 | +0.6 |
| 1lpz | high | factor Xa | 7.60 | +0.6 |
| 1mq6 | high | factor Xa | 11.15 | −2.3 |
| 1n2v | high | queunine tRNA-ribosyltransferase | 4.08 | +1.1 |
| 1n46 | high | thyroid hormone receptor | 10.52 | −2.1 |
| 1v48 | medium | purine nucleoside phosphorylase | 7.80 | −1.5 |
| 2tpi | medium | trypsinogen | 4.31 | +1.5 |

[a] $\Delta pK_d$ refers to the difference between experimental and calculated binding affinities (MM-GBSA with the VSGB 2.0 energy model[48]). $pK_d$ = −log of experimental binding affinity ($K_d$ or $K_i$). $+\Delta pK$ and $-\Delta pK$ stand here for overestimated and underestimated $pK_d$ values, respectively.

Unlike the other studies in the literature, we do not generate and store large numbers of ligand conformations in order to overcome potential ligand sampling issues.[34,35,39] One major finding from the CSAR Docking Benchmarking Exercise of 2011−2012[46] was that, as compared to ligand conformations generated on the fly, pregenerated conformations gave lower RMSD values relative to the crystal structure pose. This is suggestive of ligand sampling problems within the search engines of docking software. Here, we focus on using nondefault settings to improve conformational sampling within the software itself. Instead of a consensus approach for improving pose ranking, we focus on optimal software settings and reranking of multiple binding modes with MM-GBSA. Interestingly, Broccatelli and Brown[47] have recently reported that suboptimal ligand conformational sampling affects docking success rate more than the use of a suboptimal protein conformation. Lastly, we consider an additional data set chosen to simulate protein flexibility. Rather than explicitly sampling protein conformations during docking we use multiple protein structures and perform ensemble docking. While multiple relevant crystal structures might not be available in all structure-based projects, this study is aimed at assessing the impact of having adequate protein structures for ligand docking and scoring. Indeed, this is the theme throughout the work presented here — our aim is to investigate the docking power of scoring functions through a carefully designed set of experiments rather than to compare docking programs or directly address the protein sampling problem within the docking programs.

## METHODS

**High-Quality Data Set for Docking Power Assessment.** The ligands in Table 1 (a subset from Greenidge et al.[44] with high-quality ligand electron density) were used for self-docking and cross-docking studies. To minimize the difficulties associated with automatic assignment of both protein and ligand tautomeric and protonation states, we manually checked the data set. We selected this particular subset because of the availability of ligand density information. However, with respect to the assignment of protonation and tautomeric states it is a rather challenging selection. For example, the ligand from PDB ID 1fhd[49] was protonated on the imidazole NH to interact with the carboxylate of Glu127. In the same reference, for PDB ID 1fh9, the tautomeric state of the piperidine-2-one-oxime as shown in Figure 1 enables two hydrogen bond donor interactions to be made by the oxime moiety with Gln203 and Glu127. For the ligand from PDB ID 1n46[50] the ionized imide moiety of the 6-azauracil ring, as depicted in Figure 1, interacts with Arg320, which is supported by the fact that *N*-methylation of N-3 (elimination of the acidic proton) leads to a 50-fold reduction in binding affinity. The phenoxy group in PDB ID 1br6 is in close proximity to Arg180, suggesting a deprotonated acid that can make a salt bridge. The protonation and tautomeric states for the other ligands were relatively straightforward to assess and are shown in Figure 1.

**Larger Data Set for Scoring Power Assessment.** For comparing the score-in-place power of common scoring functions to the more computationally intensive MM-GBSA binding free energy prediction scheme, we used the same 855 membered data set as in our previous MM-GBSA analysis.[44] In


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

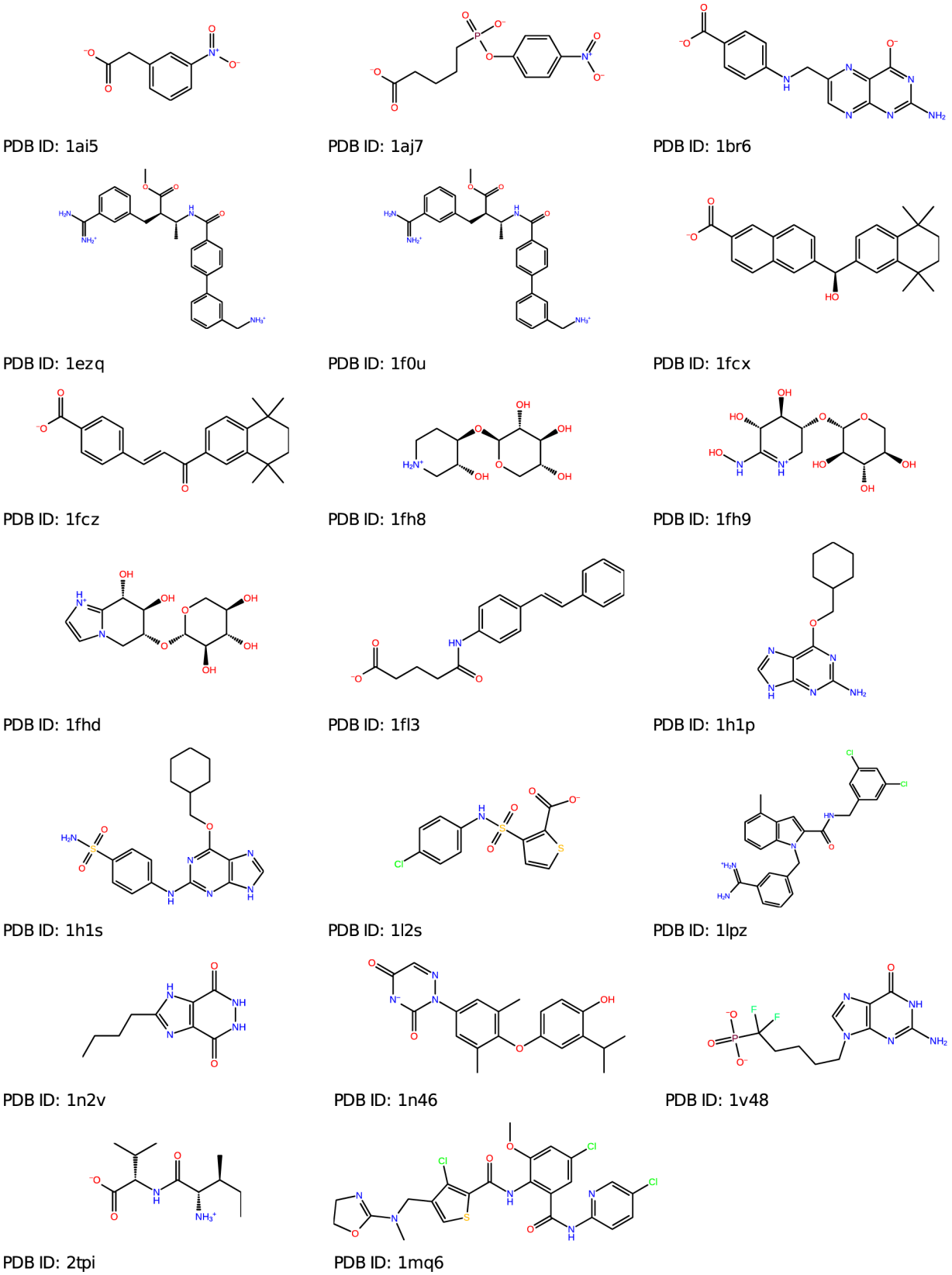



**Figure 1.** Structures of cognate ligands and corresponding PDB ID from Table 1. The ligand structures correspond to those in complex with their protein structures and are consistent with the primary citations of the complexes.

brief, this is a drug-like subset of the PDBbind 2009 refined set (resolution better than 2.5 Å, MW< 900 g/mol, < 20 donors/acceptors, maximum of 1 phosphorus atom). The structures have been prepared[51] using Schrödinger's Protein Preparation Wizard (Maestro, version 9; Schrödinger: New York, NY, 2010),[51] and only structures that passed this process without


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

errors have been kept. In short, appropriate bond orders were assigned to all residues and hydrogens were added using the default options in PrepWizard. Then the hydrogen bond network was optimized by rotating all groups unambiguously defined by electron density (Asn, Gln, and His), His tautomer/ionization states were varied, and hydroxyl/thiol hydrogens were optimized. Finally, an all-atom minimization was performed with a termination criterion of 0.3 Å on the non-hydrogen atoms. All waters and ions were removed before the aforementioned procedure. During this procedure any missing side chains were added. All complexes with ions in the binding site have been removed to exclude possible problems arising from incorrectly accounting for the covalent nature of some interactions with ions. Details about the data set preparation can be found in a previous publication.[44]

**GOLD Calculations.** *Rescoring/Score in Place.* The complexes according to the preparation described above were used for scoring in place the extracted ligands. Local minimization of the ligands was performed for rescoring using Gold v5.1. Cole et al.[25] recommend that comparisons based on rescoring require local optimization of the ligand with respect to the scoring function being assessed. The scoring functions used in this work were ChemPLP,[18] GoldScore,[16,36,37] and ChemScore.[52]

*Docking.* Default options were used, as described below, unless otherwise noted.

(i) In the first round of docking up to 25 genetic algorithm (GA) runs (docking attempts) were allowed with the possibility of early termination. All atoms within 6 Å of any atom of the cognate ligand were used to define the binding site. The slow option was used for docking. All results were written out and clustered, with the distance between the centroids of the clusters set to 0.75 Å. The scoring function used throughout was ChemPLP.[18]

(ii) In a second round of docking up to 50 GA runs were allowed without the possibility of early termination.

*Ensemble Docking.* Ensemble docking (50 GA runs) was used for the three factor Xa (fXa) complexes (PDB ID 1mq6, 1lpz, and 1eqz). The inhibitor from 1lpz was used to define the binding site of the aligned proteins. The CDK2 inhibitor from PDB ID 1h1s (NU6102) was used to define the ensemble binding site including PDB ID 1h1p. The hRAR ligand from PDB ID 1fcx was used to define the ensemble binding site including PDB ID 1fcz. M77 was the common reference ligand used to define the binding site in all five proteins of Case Study 2 (PDB IDs 3DND, 3AGM, 1Q8W, 1CDK, and 1CMK), and GoldScore was the scoring function used.

**Glide Calculations.** All Glide calculations were performed with version 5.8 using the Standard Precision (SP) mode available within Maestro v9.3.515 with default options. PDB structures were prepared with the Protein Preparation Wizard (PrepWizard) in Maestro prior to docking.[51] In short, appropriate bond orders were assigned to all residues and hydrogens were added using the default options in PrepWizard. Then the hydrogen bond network was optimized by rotating all groups unambiguously defined by electron density (Asn, Gln, and His), His tautomer/ionization states were varied, and hydroxyl/thiol hydrogens were optimized. Finally, an all-atom minimization was performed with a termination criterion of 0.3 Å on the non-hydrogen atoms. All waters and ions were removed before the aforementioned procedure. Docking grids were generated using default settings, and the cognate ligand was used to define the center of the grid box. No constraints were used, and only the best pose per ligand based on Emodel was saved. For score in place, *Ligand Sampling* (refine only) was used with and without a correction for ligand strain.

**Input Ligand Conformations.** Two sets of ligand conformations were used as input for docking the 20 ligands in Table 1. Set 1 includes the conformation of the ligand as extracted from the refined complexes in Greenidge et al.[44] The second set was obtained from the lowest energy conformer resulting from the Bioactive Search option within the Conformational Search engine (ConfGen) in Maestro.[30] The *Fast(CF)*[53] option was used, and the input and output structures were minimized. Case Study 2 ligands, LL1 (PDB ID 3DNE), LL2 (PDB ID 3DND), M77 (PDB ID 1Q8W), PTV (PDB ID 4IJ9), and PZX (PDB ID 4IE9), were built in Maestro and prepared using LigPrep with default settings. The lowest energy conformers, generated as above (Bioactive Search), were used for docking.

**MM-GBSA.** MM-GBSA binding affinities for the full 855 complex data set have previously been computed as described in Greenidge et al.[44] using a KNIME[54] workflow and the VSGB2.0 solvent model[48] with the OPLS2005 force field.[55−57]We originally created the workflow to handle a large number of complexes and because the GUI option at that time forced minimization of protein residues within a fixed radius of a reference ligand, whereas we wanted to minimize the ligand but keep the protein fixed. Additional computations performed here also used the VSGB 2.0 energy model[48] using the Maestro GUI "Binding Energy Estimation" panel in Prime with only the ligand minimized. MM-GBSA energies are computed with and without the inclusion of ligand strain. The ligand strain energy is the difference between two energies: the energy of the ligand as it is in the complex and the energy of the extracted ligand, minimized, starting from the geometry in the refined complex. Calculations to assess the strain were performed in implicit solvent.

**Decoys.** We checked the DUD[6,7] database and found appropriate decoys for the two CDK2 inhibitors (PDB ID 1h1p and 1h1s); 35 decoys for each of the inhibitors were selected. LigPrep with default options was used to generate structures for docking from the decoys in SMILES format. This led to a total of 62 and 58 decoys for the PDB ID 1h1p and 1h1s decoy sets, respectively. LigPrep is the recommended method for preparing ligands for use with Glide. The inhibitor from PDB ID 1h1s was used to define the binding site for GOLD docking using ChemPLP as the scoring function (cf. GOLD ensemble docking above). All members of the output clusters (cf. above) were scored in place with both ChemPLP[18] and MM-GBSA VSGB 2.0[48] in both PDB ID 1h1p (Chain C) and PDB ID 1h1s.

## RESULTS AND DISCUSSION

**Scoring Power.** Based on the results of their study, Warren et al.[15] concluded that docking scoring functions performed poorly. This was because even when the RMSD of docked poses relative to cognate poses decreased, the correlation of the empirical score with experimental binding affinity did not improve. In light of the above, here we first establish if a correlation does in fact exist between the docking score of the crystallographic pose with the experimental binding affinity. However, it should be noted that GOLD[16,36,37] and Glide[22,38] scoring functions were not parametrized for predicting binding energies — they were developed for pose prediction and/or the separation between active and inactive compounds in virtual


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

screening. Thus, there seems to be some confusion in the literature between scoring power and docking power; the performance of one does not necessarily predict the performance of the other.

We "score-in-place" the data set of 855 curated complexes from PDBbind[44] using two popular docking programs (GOLD[16,36,37] and Glide[22,38]) and the scoring functions contained therein. The score-in-place results give a limit for the best obtainable correlations between experimental data and scores for well-docked solutions, assuming the crystallographic pose is correct. Previously, a coefficient of determination ($R^2$) of 0.63 was obtained between experimental and calculated binding affinities using MM-GBSA with the VSGB 2.0 energy model[48] as the scoring function for the data set of 855 complexes.[44] For the present study, GlideScore,[38] ChemPLP,[18] ChemScore,[52]and GoldScore[16,36,37] were used to score in place this data set. $R^2$ values of 0.36 to 0.48 were obtained for the different scoring functions, indicating that the scoring functions in the docking programs used here are not as good as the MM-GBSA model used for rescoring. See the Supporting Information for more details.

Thus, the examined docking scoring functions did not produce correlations with experimental binding activity as high as the MM-GBSA method used in our previous work.[44] There are many reasons for the lack of correlation, such as not accounting for protein strain energy, inadequate treatment of solvation effects, no protein sampling, and neglect of entropy. These results are not unexpected, given the objective of docking programs is not to rank-order compounds binding to different targets. We next examine how these programs fared with respect to pose prediction, which is a primary objective of most docking programs.

**Docking Power.** One motivation for this study was the belief that if the correct pose in the most appropriate protein conformation can be identified, then it will be more productive to explore deficiencies in scoring functions and how to improve them. For example, with an accurate pose it is possible to apply a more computer intensive scoring function for pose ranking, such as MM-GBSA, which has been shown by some to be a superior scoring function as compared to docking scoring functions.[9,12,13] In fact, many authors suggest that MM-GBSA should routinely be used to rescore docking results. However, as previously discussed, any such protocol will reflect more upon the quality of the poses than upon the new scoring function if accurate protein−ligand complex structures cannot routinely be obtained.

The 20 compounds used for assessing docking power represent the intersection of the Iridium data set[45] complexes (mainly high trust category) present in our larger data set of PDBbind compounds used for investigating scoring power (see Table 1). Ideally, poses that deviate from the experimental pose should be penalized with worse scores. While the results derived from a data set of this size can be limited in statistical significance, the results can provide valuable insights and case studies. Furthermore, the value of data set quality can be at least as important as quantity, as demonstrated in studies by Plewczynski et al.[35] (1300) and Li et al.[39] (195), which use subsets of PDBbind[58] for assessing the docking power of multiple scoring function but arrive at different conclusions. Plewczynski et al.[35] report that when the ligand conformational space is searched more thoroughly, that this is not reflected in the improvement in docking power of the ligand. This contrasts with the more recent findings of Li et al.[39] who concluded that generally docking scoring functions performed well in docking power for redocking the native pose. The success rates were greater than 70% for GoldScore, ChemPLP, and GlideScore. They note that the docking power has improved from the previous study[34] as more diverse ligand poses are used. They evaluate docking power by scoring in place previously generated ligand poses (ensembles). Up to 100 conformations were allowed to be associated with each ligand.[39] The different conclusions about the docking power of the scoring functions reached by these studies can arise from a number of factors. Among them is the method of generation of the ligand conformations for docking and the crystallographic quality of the protein-complexes used for docking. Li et al.[39,59] perform a visual quality check of the ligand electron density and any nearby residues — they make the point that data set quality should not be sacrificed just to obtain a larger test set. In a separate study, Hawkins et al.[60] have found that there is a less clear difference in performance between docking scoring functions when using Iridium-HT versus Iridium-MT protein−ligand complexes. Also Hawkins et al.[60] have argued that statistical differences may be of more importance to software developers than to software users, and they caution against using poorly solved ligand structures for pose prediction, since the selection of a metric to classify how well a crystallographic binding mode is reproduced by a docking program is outweighed by how to best select a suitable data set. Nominal resolution alone of a crystal structure complex is not a sufficient selection criterion, and it is generally acknowledged that the use of RMSD as the measure of the reproduction of the correct pose of ligands may not always be the most appropriate metric.[45,61,62] In X-ray structures of complexes, ligand coordinates are often less precise than those for the protein. Even in high-resolution structures it is not uncommon to find ligand geometries that are of poor quality (unnatural distortions, wrong bond length and angles, and high-energy conformations that cannot be attributed to induced fit) when comparing them to high-quality X-ray structures of the isolated ligands or of closely related structures.[63] Many X-ray structures have alternate binding modes for the ligand, i.e., the pose is not unique, even in the X-ray structure. In addition, a decent RMSD ($\leq$2 Å) can be obtained even when some key interactions are not captured.

In our study, we score in place the X-ray pose and compare the docking scoring function value with that of the top ranked docked pose. Also by comparing the RMSD between the docked and X-ray pose we determine if large deviations from the crystallographic pose are penalized by the docking scoring function. We do not perform a black box assessment; instead we also discuss the issues/give input upon the parameters, which affect the ligand conformational sampling of the software. To fairly compare the self- and cross- docking performance of the docking functions water molecules have been removed, although future work should consider the inclusion of explicit water molecules. The majority of the ligands of the data set do have at least one water mediated interaction with the protein. However, the number of direct protein−ligand interactions dominate the bridged hydrogen bond interactions in most cases. This information is summarized in the Supporting Information. There are four exceptions where this is not the case - the complexes with PDB IDs 1fl3, 1lpz, 1mq6, and 1n2v, respectively. This might indicate that in the absence of explicit solvent, docking software might struggle to replicate the crystallographic pose of the


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

ligand. However, the binding of the 1mq6 inhibitor is driven by shape constraints (see discussion in the Supporting Information), and the 1fl3 ligand is divided into hydrophobic and solvent exposed moieties by an amide group, hence helping to orient the binding mode. Based on RMSD values, neither of these molecules proved to be problematic to redock (Tables 2 and 3a). 1mq6 is however noteworthy for being discussed as requiring increased sampling by GOLD in order to reproduce the crystallographic pose. The 1n2v ligand posed some difficulty for GOLD (see later discussion), and the 1lpz ligand presented some difficulty for Glide (see later discussion). The 1h1p inhibitor represents an ambiguous case. Three direct but no bridged hydrogen bonds are made in chain A. In chain C, in addition to the direct hydrogen bonds, there are also two water-mediated hydrogen bonds. The correct docking of the 1h1p inhibitor in chain A proved to be difficult for Glide and in chain C problematic for GOLD.

**Table 2. Glide Docking Results Using Restraint-Optimized Crystallographic (Set 1) and Randomized Ligand Conformations (Set 2) with the SP Scoring Function[b]**

| | | Set 1 | | Set 2 | |
|---|---|---|---|---|---|
| PDB ID | score in place | GlideScore | RMSD | GlideScore | RMSD |
| 1ai5 | −7.14 | −8.44 | 0.86 | −8.59 | 0.63 |
| 1aj7 | −7.86 | −8.64 | 0.52 | −9.00 | 0.54 |
| 1br6 | −7.22 | −6.84 | 0.10 | −6.84 | 0.11 |
| 1ezq | −12.7 | −12.70 | 0.16 | −12.69 | 0.17 |
| 1f0u | −9.38 | −9.91 | 0.31 | −9.39 | 0.52 |
| 1fcx | −15.93 | −15.97 | 0.08 | −15.95 | 0.09 |
| 1fcz | −15.27 | −15.45 | 0.20 | −15.40 | 0.19 |
| 1fh8 | −7.94 | −8.11 | 0.43 | −8.12 | 0.43 |
| 1fh9 | −7.67 | −7.88 | 0.38 | −7.88 | 0.38 |
| 1fhd | −7.56 | −7.38 | 0.42 | −7.39 | 0.42 |
| 1fl3 | −9.48 | −9.88 | 0.55 | −9.61 | 0.80 |
| 1h1p | −9.23 | **−7.30** | **4.94** | **−7.39** | **4.82** |
| 1hls | −10.76 | −10.85 | 0.24 | **−7.74** | **4.84** |
| 1l2s | −6.74 | −6.24 | 0.34 | −6.77 | 0.49 |
| 1lpz | −10.26 | **−4.62** | **7.78** | −10.86 | 0.55 |
| 1mq6 | −10.14 | −10.24 | 0.67 | −10.48 | 1.47 |
| 1n2v[a] | −7.57 | −7.23 | 2.00 | −7.22 | 2.00 |
| 1n46 | −12.98 | −13.44 | 0.19 | −13.45 | 0.14 |
| 1v48 | −10.14 | −10.03 | 0.21 | −10.03 | 0.22 |
| 2tpi | −6.91 | −7.32 | 1.13 | −7.04 | 0.49 |

[a] 1n2v RMSD is misleading due to the nonspecific interactions of the *n*-butyl moiety, which is responsible for the increased RMSD value. [b] GlideScore and RMSD values (Å) between the heavy atoms of the docked and restraint-optimized crystallographic ligand pose are reported. The docked compounds that have large heavy atom deviations from the crystallographic pose are highlighted in bold. In all cases, poor RMSD values are associated with poor scores relative to the score-in-place values, indicating sampling and not scoring problems.

*Ligand Conformations.* We commence by examining the influence of initial ligand conformations and sampling level on pose prediction. It has been noted in previous studies that using experimentally observed ligand geometries as starting structures increases success rates[25] for so-called self- or cognate-docking, since bonds and angles are fixed at ideal values for the correct pose. As such, success rates are intrinsically linked to the program used to generate the initial structure. By means of an extensive self-docking study (798 complexes), it has already been shown that for Glide, there is a marked relationship between how well a pose is reproduced and the manner in which the initial conformation is generated.[38] For the 20 complexes in Table 1, the ligands were scored in place for the purpose of comparison with self-docking or cross-docking scores. Tables 2 and 3a show the comparison between the score of the top scoring pose from fully flexible ligand docking using either the crystal structure ligand input conformation or that from ConfGen; the results confirm the sensitivity of docking results to initial ligand conformations.

*Docking Power in Self-Docking. Glide.* The two complexes PDB ID 1ai5 and 1aj7 produce an anomaly; their docking scores are notably better than the score-in-place results. Of these complexes, 1aj7 has poor electron density for the ligand[45] (nontrustworthy category of ligand electron density, Table 1), and the docked complex makes one more hydrogen bond with the protein than the ligand used for scoring in place (Tyr33), suggesting that there might be a better pose than the experimentally determined crystal structure. The carbonyl oxygen of the ligand in the complex PDB 1ai5 used for scoring in place makes a bidentate hydrogen bond with the side chains of Ser1 and Asn241. The docked ligands similarly make a bidentate hydrogen bond but only with Ser1 (side chain and backbone). Starting from the restraint-optimized crystal structure conformations of the ligands, Glide SP fails to accurately reproduce the crystallographic poses of the CDK2 inhibitor from 1h1p and the fXa inhibitor from 1lpz (Table 2 bold). Feher and Williams[64] have also found that starting from the X-ray ligand conformation does not always guarantee a pose with a low RMSD to the cognate pose. Because of the docking protocol (approximated exhaustive conformational search), the initial conformation of the ligand is never explicitly docked.[38] Since the conformational sampling is finite, different docking solutions can be obtained from even quite similar starting geometries. This reflects that the potential energy landscape has multiple minima — one might expect many hundreds or thousands of local minima for a given protein−ligand complex. Docking with randomized conformations of the ligands, the crystallographic poses of both CDK2 inhibitors (1h1p and 1h1s) cannot be reproduced (Table 2 bold). The docked compounds that have large heavy atom deviations from the crystallographic pose are penalized, as reflected by their lower scores (Table 2 bold).

We investigated various other settings available in Glide docking to see if it was possible to improve the RMSD values for the two CDK2 ligands (randomized conformations). Expanding the sampling and using a correction for strain, the top ranked Emodel pose of the inhibitor from 1h1p (NU2058) gave an improved RMSD of 2.68 Å. The purine base is flipped by 180° along the long axis with respect to the crystallographic pose but also makes three hydrogen bonds to the hinge (Figure 2). The energy of this flipped pose is calculated by MM-GBSA to be 1 kcal/mol better than the correct pose, indicating a near degeneracy in the scoring function for these two possible hinge interaction motifs. This flipping phenomenon was also experienced by Thomas et al.[4] when docking a series of purines also to CDK2 proteins. Using the same options for self-docking the inhibitor from 1h1s (NU6102) gave an excellent RMSD of 0.29 Å. These cases indicate that enhanced sampling and more accurate accounting of ligand strain can improve docking accuracy, although a study of more cases would be needed to draw statistically significant conclusions.


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 3a. GOLD Docking Results Using Restraint-Optimized Crystallographic (Set 1) and Randomized (Set 2) Initial Ligand Conformations and Using up to 25 and 50 Docking Attempts[a]**

| PDB ID | Score | Set1 (25) | RMSD | Set2 (25) | RMSD | Set2 (50) | RMSD |
|---|---|---|---|---|---|---|---|
| 1ai5 | 45.08 | 48.00 | 0.17 | 47.95 | 0.63 | 48.02 | 0.64 |
| 1aj7 | 75.71 | 82.69 | 0.58 | 80.41 | 0.67 | 83.12 | 0.66 |
| 1br6 | 73.49 | 74.77 | 0.38 | 71.69 | 1.62 | 74.97 | 0.49 |
| 1ezq | 123.68 | 118.7 | 2.60 | 100.14 | 1.10 | 96.98 | 1.86 |
| 1f0u | 100.39 | 85.93 | 1.58 | 82.95 | 1.37 | 83.95 | 1.91 |
| 1fcx | 119.83 | 123.36 | 0.32 | 118.90 | 0.44 | 119.18 | 0.44 |
| 1fcz | 114.88 | 117.94 | 0.27 | 110.00 | 0.48 | 111.18 | 0.49 |
| 1fh8 | 73.82 | 78.42 | 0.36 | **62.31** | **2.21** | **62.40** | **2.61** |
| 1fh9 | 74.26 | 75.37 | 0.48 | 54.28 | 1.54 | 54.70 | 1.55 |
| 1fhd | 80.80 | 87.92 | 0.32 | **58.14** | **4.79** | **58.94** | **5.75** |
| 1fl3 | 90.22 | 94.95 | 1.10 | 95.58 | 0.47 | 95.55 | 0.51 |
| 1h1p | 53.89 | 59.04 | 0.49 | 58.30 | 0.56 | 58.33 | 0.55 |
| 1hls | 74.51 | 76.72 | 0.87 | 75.54 | 0.90 | 76.92 | 0.87 |
| 1l2s | 55.25 | 54.57 | 1.00 | **54.00** | **2.71** | 55.90 | 1.41 |
| 1lpz | 108.20 | 106.83 | 0.57 | 103.13 | 0.83 | 109.81 | 0.58 |
| 1mq6 | 109.78 | 102.01 | 1.00 | 106.86 | 1.60 | 103.16 | 0.71 |
| 1n2v | 49.58 | 55.61 | 3.04 | 54.68 | 2.65 | 55.44 | 3.02 |
| 1n46 | 108.02 | 112.89 | 0.19 | 109.22 | 0.24 | 109.48 | 0.22 |
| 1v48 | 95.78 | 93.58 | 1.00 | 95.07 | 0.31 | 96.10 | 0.90 |
| 2tpi | 64.59 | 70.65 | 0.56 | 68.21 | 0.58 | 69.43 | 0.54 |

[a]ChemPLP scores and RMSD values (Å) between the heavy atoms of the docked and restraint-optimized crystallographic ligand pose are reported. The larger the score, the better the predicted protein-ligand complementarity. The docked compounds that have large heavy atom deviations from the crystallographic pose are highlighted in bold.

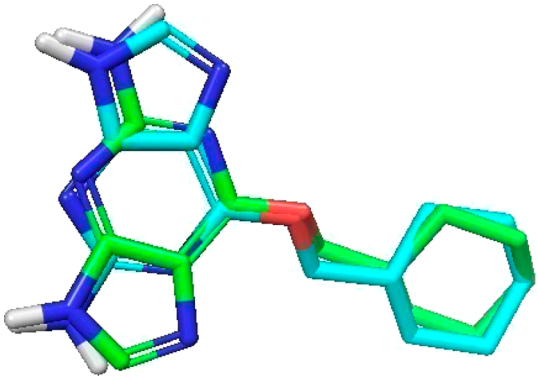


**Figure 2.** Overlay between the crystallographic pose of NU2058 (cyan, 1h1p) and the docked pose of the inhibitor (green).

*GOLD ChemPLP. Top Ranked Poses.* It has been stated that GOLD docking results should be relatively insensitive to the initial ligand conformation as long as it is a minimized structure.[24] However, in the work here it is seen that using randomized ligand conformations as input to GOLD generally produce poorer docking results than starting from optimized crystal structure conformations, likely because GOLD will not alter bond lengths or angles or rotate rigid bonds such as amide linkages, double bonds, and certain bonds to trigonal nitrogens. The degradation in results is reflected in their lower (i.e., poorer) scores (Table 3a), which is consistent with the findings of Corbeil and Moitessier,[32] who observed a drop in docking accuracy of up to 20% for noncrystal ligand structures as compared to crystal ligand structures. Initially 25 GA docking attempts were allowed (the default option is 10); the ligand from 1mq6 achieved its best score and lowest RMSD value (1.0 Å) on the 25th docking attempt — with the default of 10 docking attempts the RMSD for the lowest energy structure would have been 7.6 Å, demonstrating the value of additional sampling.

Redocking the 20 ligands with randomized geometries and increasing the allowed docking attempts to 50 (Table 3a) led to further improved RMSD values. In particular, the RMSD for the beta-lactamase ligand in 1l2s improved from 2.71 to 1.40 Å (best result from the 28th docking attempts), and the score for this pose was improved (Table 3a). Thus, the failure to reproduce native poses consistently points to inadequate sampling rather than scoring issues. The two CDK2 kinase inhibitors NU2058 and NU6102 (from 1h1p and 1hls, respectively), which were poorly docked by Glide (without expanded sampling), were well docked by GOLD (Table 3a). The three members of the xylanase data set, which contain sugar-like rings, are not well docked by GOLD and score poorly, possibly indicating inadequate sampling of the saturated rings. These molecules are better handled by Glide, which has more extensive ring sampling as a default option (Tables 2 and 3a). Previous studies have specifically chosen to omit such structures in docking data sets given the challenges associated with ring sampling,[17] although this does not represent a realistic test given the prevalence of saturated rings in pharmaceutically relevant screening sets.

*Pose Ranking.* GOLD uses hierarchical cluster analysis (complete linkage algorithm) to cluster the docking poses output from multiple docking attempts. A top-ranked pose (i.e., the cluster member with the best fitness score) is associated with each distinct cluster. This facilitates the identification of distinct binding modes of the ligand to the protein. The danger is that by only considering the individual top *n* ranked poses, all may belong to the same cluster and thus may represent very similar binding modes.[65] It occurred to us that lower RMSD values might exist for the poorly docked xylanase inhibitors. Therefore, we calculated RMSD values relative to the restraint-optimized crystallographic ligands, for all cluster rank


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

representatives (Set 2 (50)) of all complexes (Supporting Information). This was indeed the case. The third ranked pose of the 1fhd inhibitor has an RMSD of 1.9 Å versus 5.8 Å for the top ranked pose. The difference in ChemPLP score is only 0.5 (Table 3b). Similarly, the third ranked pose of the 1fh8 inhibitor has an RMSD value of 1.8 Å versus 2.6 Å for the top ranked one (Table 3c). The difference in ChemPLP scores is 0.6. However, both of these poses fail to make a hydrogen bond with His80 as in the X-ray, but the third ranked pose replicated the hydrogen bond between the ether oxygen atom and Lys47.

**Table 3b. Relationship between Ranking, ChemPLP, and MM-GBSA Binding Affinity (Including Ligand Strain) and RMSD to the Crystallographic Pose for the Ten Distinct Output Binding Modes for Docking of the Ligand from 1fhd Using GOLD[a]**

| cluster rank | ChemPLP | RMSD (Å) | MM-GBSA (kcal/mol) | MM-GBSA rank |
|---|---|---|---|---|
| X-ray | N/A | 0.00 | −84.5 | N/A |
| 1 | 58.9 | 5.75 | −46.9 | 7 |
| 2 | 58.7 | 4.86 | −38.0 | 9 |
| 3 | 58.3 | 1.88 | −60.7 | 1 |
| 4 | 55.5 | 6.59 | −54.4 | 4 |
| 5 | 55.3 | 5.56 | −56.4 | 2 |
| 6 | 54.7 | 6.40 | −53.9 | 5 |
| 7 | 54.5 | 4.53 | −49.2 | 6 |
| 8 | 53.9 | 3.88 | −37.5 | 10 |
| 9 | 52.9 | 5.41 | −54.6 | 3 |
| 10 | 51.4 | 4.07 | −41.6 | 8 |

[a]N/A not applicable.

**Table 3c. Relationship between Ranking, ChemPLP, and MM-GBSA Binding Affinity (Including Ligand Strain) and RMSD to the Crystallographic Pose for the Four Distinct Output Binding Modes for Docking of the Ligand from 1fh8 Using GOLD[a]**

| cluster rank | ChemPLP | RMSD (Å) | MM-GBSA (kcal/mol) | MM-GBSA rank |
|---|---|---|---|---|
| X-ray | N/A | 0.00 | −69.0 | N/A |
| 1 | 62.4 | 2.61 | −56.9 | 1 |
| 2 | 61.84 | 2.59 | −56.1 | 2 |
| 3 | 61.80 | 1.82 | −53.9 | 4 |
| 4 | 61.44 | 2.78 | −55.9 | 3 |

[a]N/A not applicable.

The fifth ranked pose of the 1n2v inhibitor has an RMSD of 0.8 Å. The top ranked pose (Table 3d) "slides" to make direct as opposed to water-mediated hydrogen bonds with Asp103 and Ser102. For docking, all water molecules have been removed. The third ranked pose of the 1ezq inhibitor has an RMSD value of 1.4 Å versus 1.9 Å for the top ranked one, but this lower RMSD is obtained by the loss of direct hydrogen bonds of the guanadinium group with Asp189 (Supporting Information). Thus, the superior ranking of the pose that deviates more from the X-ray one makes sense in this case. Generally, the poses that deviate the most from the crystallographic ones are penalized by being given lower scores than poses that show lower deviations. There are however some apparent anomalies. Closer inspection shows that some poses with low RMSD do in fact clash badly with the receptor. For example, two poses with RMSD values of 0.9 in 1v48 are given very different scores, 96 and 78, respectively; the latter makes unfavorable interactions with the receptor, underlining the poor descriptive content of RMSD values. Similarly, two poses of the 1aj7 ligand have equal RMSD values of 0.66 but are ranked first and fifth, respectively, with scores of 83 and 76, respectively (Supporting Information). The above suggests that the top three distinct poses should be considered as equally valid poses, and up to five poses should be retained if there are known to be key water-mediated interactions. Thus, while the scoring functions perform poorly with respect to correlation with experimental data (shown in the previous section), they do well with respect to pose prediction in self-docking studies with sufficient ligand sampling and retention of sufficient poses.

**Table 3d. Relationship between Ranking, ChemPLP, and MM-GBSA Binding Affinity (Including Ligand Strain) and RMSD to the Crystallographic Pose for the Six Distinct Output Binding Modes for Docking of the Ligand from 1n2v Using GOLD[a]**

| cluster rank | ChemPLP | RMSD (Å) | MM-GBSA (kcal/mol) | MM-GBSA rank |
|---|---|---|---|---|
| X-ray | N/A | 0.00 | −55.95 | N/A |
| 1 | 58.9 | 3.02 | −53.24 | 3 |
| 2 | 58.7 | 2.65 | −54.51 | 2 |
| 3 | 58.3 | 2.79 | −53.17 | 4 |
| 4 | 55.5 | 1.29 | −52.45 | 5 |
| 5 | 55.3 | 0.80 | −56.34 | 1 |
| 6 | 54.7 | 1.69 | −51.29 | 6 |

[a]N/A not applicable.

*Pose Ranking MM-GBSA.* Given the narrowness of the ChemPLP score window separating some of the poses, we went on to calculate MM-GBSA protein−ligand interaction energies (rigid receptor, flexible ligand) as described in methods. The X-ray pose of the ligand from 1fhd was clearly calculated to have the best binding energy, followed by the pose with the lowest RMSD to the crystallographic one (Table 3b). Even though the RMSD is within the generally accepted value of 2 Å, the calculated energy difference to the X-ray pose is >20 kcal/mol. There was not a strong correspondence between the ranking produced by the ChemPLP and MM-GBSA scoring schemes (Table 3b). The piperidine moiety of the ligand from 1fh8 is docked similarly but incorrectly for all four cluster representatives, and the main variation in docking arises from the xylopyranose moiety. Once again, MM-GBSA identifies the crystallographic pose as the most energetically favored one, and the MM-GBSA and ChemPLP rank order is pretty consistent (Table 3c). The docking pose with the lowest RMSD to the crystallographic one is however ranked as the least favorable pose (Table 3c). Presumably, the combination of incorrect docking of the piperidine part of the 1fh8 ligand is not energetically compatible with correct docking of the xylopyranose part (Figure 3). This example once again highlights the complexities of docking interactions that cannot be summarized by RMSD values alone. Table 3d summarized the MM-GBSA scoring results for the docking poses in 1n2v. There is some subtle shifting with respect to the ChemPLP ranking. Most notably, the pose ranked fifth by ChemPLP is energetically equivalent to the crystallographic pose according to MM-GBSA scoring following minimization of the ligand.

Overall there is a strong correspondence between the pose ranked top by both ChemPLP (Set 2, 50 docking attempts) and MM-GBSA. When this is not the case, then the top ranked


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

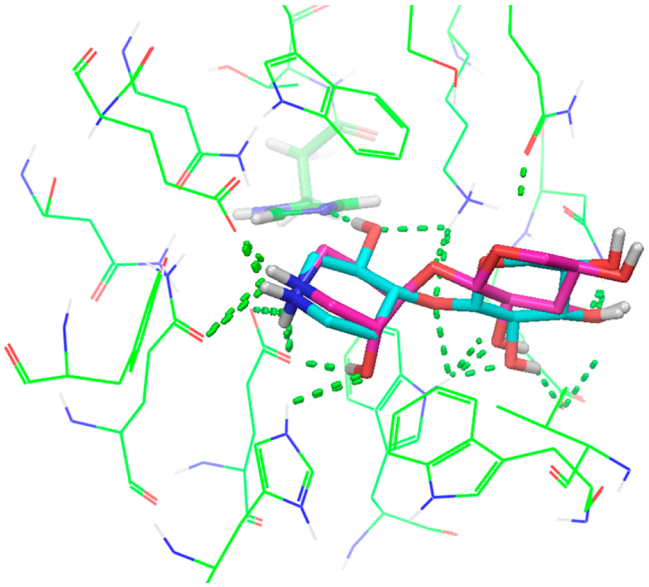


**Figure 3.** Comparison between the crystallographic pose of the 1fh8 ligand (cyan) and the docked pose with overall lowest RMSD (magenta). His80 is shown as sticks.

MM-GBSA pose is contained within the top 5 ChemPLP poses (Supporting Information). Glide has postdocking minimization as a default setting unlike GOLD. Hence, GOLD may benefit more from the combination with MM-GBSA scoring. For instance, the top two ranked docking poses of the 1ezq fXa inhibitor have an RMSD of 1.86 and 1.46 Å, respectively, relative to the restrained optimized crystallographic pose and energy differences of +24 kcal/mol and +12 kcal/mol, respectively (Supporting Information), following minimization. The pose that has the best calculated protein−ligand interaction energy is in fact ranked fourth according to its ChemPLP score with an RMSD of 2.3 Å. After minimization, the energy window to the restrained optimized crystallographic pose is only +3 kcal/mol. The benefits of a postminimization step after GOLD docking are further illustrated by the 1lpz fXa inhibitor. The top ranked docked pose has an RMSD of 0.58 Å; however, after minimization, MM-GBSA ranks this fifth as this pose is associated with high ligand strain energy (+14 kcal/mol). The third, fourth, and fifth ranked docked poses become the top ranked MM-GBSA poses with protein−ligand interaction energies equivalent to the X-ray pose.

*Docking Power in Cross Docking and Ensemble Docking.* The realistic scenario for the application of docking in the pharmaceutical industry involves docking of ligands to receptor structures with which they have not been cocrystallized (cross-docking).[19,24,28,66] This means that the protein structure has not adapted its conformation to match the ligands being docked.[67,68] Glide allows for soft-docking by scaling of the vdW radii of nonpolar atoms of both proteins and ligands. In addition, both Glide and GOLD allow for ensemble docking,[66,69,70] which simulates protein flexibility by combining the results from docking to individual rigid receptors.[71] Here, we include only the proteins from the high-quality subset for use in ensemble docking (see Methods). It is possible to explicitly sample the protein to account for induced-fit effects upon ligand binding,[72−75] and it has been found that using multiple structures, whether from experimental or other sources, tends to be beneficial for pose prediction and virtual screening studies.[4,67,76,77] According to a recent publication by Broccatelli and Brown,[47] the explanation may be that the use of multiple protein conformations indirectly compensates for poor ligand conformational sampling. However, Korb et al.[76] have concluded that an optimal protocol to select ensemble members does not yet exist. A generic ensemble selection procedure is complicated by a scoring function dependence.[76] Existing methods for choosing the protein structure to use for docking take into consideration ligand similarity,[28,47] binding cavity volume,[4,69,70] and the orientation of the residues forming the binding site. Currently, when relevant crystal structures are available, docking into such an ensemble with diversely shaped cavities yields better enrichment and diversity of active ligands than docking into structures derived from simulations.[70]

Here, we examine the performance of the GOLD scoring function ChemPLP[18] when used for cross-docking. There are three factor Xa (fXa), two CDK2 kinase, and two human retinoic acid receptor (hRAR) complexes that can be used for this purpose. We excluded the three xylanase complexes from this part of the study due to problems experienced by GOLD in reproducing the experimental pose of these compounds even in self-docking tests. We chose GOLD for the cross-docking study primarily because of our experience with this software with respect to ensemble-docking studies. The results for fXa and hRAR are given as Supporting Information.

**Case Study 1**: CDK2

It was clear from the outset that the two CDK2 structures were not conducive for cross-docking because of the flexibility of the glycine rich loop region (the position of Ile10 in particular) and Lys89 (Figure 4).[4] This was confirmed by the

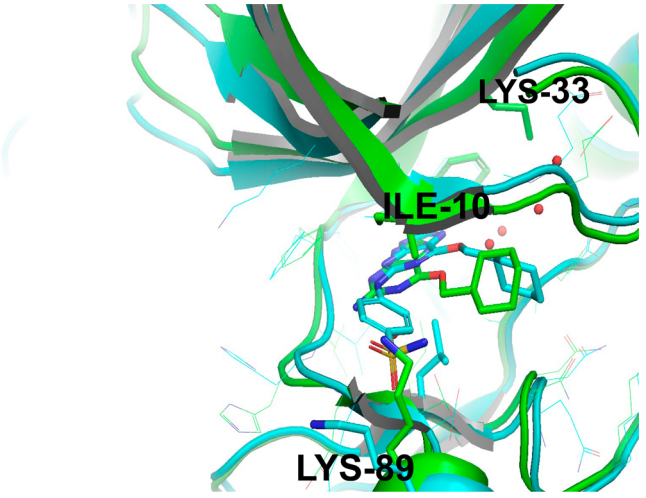



**Figure 4.** Overlay between the crystal structures 1h1s (cyan) and 1h1p (green) and their respective inhibitors.

poor docking poses for the inhibitors NU2058 (inhibitor from 1h1p) and NU6102 (inhibitor from 1h1s) when docking into the noncognate protein structure (Table 4a). The Lys89 side chain in the protein structure corresponding to 1h1p occupies the same spatial position as the sulfonamide moiety of the inhibitor in 1h1s. Thus, in cross-docking the inhibitor is unable to adopt the cognate pose.

The side chain of Lys89 in 1h1p occupies the same spatial position as the sulfonamide moiety of the inhibitor NU6102. Thus, in cross-docking the inhibitor from 1h1s is unable to adopt the cognate pose. However, the computed ChemPLP


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 4a. Docking of NU6102 (the Inhibitor from 1h1s), in Protein Structures from Chain C of 1h1p and 1h1s Using GOLD[a]**

| protein PDB ID | ChemPLP | RMSD | MM-GBSA |
|---|---|---|---|
| 1h1p (cross-dock) | 77.45 | 5.19 | −61.6 |
| 1h1s (native) | 76.16 | 0.52 | −76.8 |

[a]ChemPLP score and MM-GBSA binding affinity (kcal/mol, including ligand strain) and RMSD (Å) to the crystallographic pose for top ranked docking mode.

scores are within 1 unit of each other for the very different poses adopted in the proteins from 1h1s and 1h1p (as gauged by RMSD values, Table 4a), indicating that for this system the ChemPLP scoring function may not be able to differentiate the correct from incorrect pose. When MM-GBSA is applied to score the poses from GOLD, the correct pose with the lowest RMSD to the cognate inhibitor is clearly recognized (Table 4a). Ensemble docking of the 1h1s inhibitor resulted in 17 clusters; 7 had a preference for the cognate protein (Table 4b). The poses in 1h1p (cross-docking) were associated with much poorer MM-GBSA binding energies and higher strain energy than those in the cognate protein (1h1s). The top scoring poses of NU6102 in 1h1p tend to orient the sulfonamide in the vicinity of Lys88, which the ChemPLP score cannot distinguish from the correct pose, whereas MM-GBSA penalizes this partially solvent exposed polar interaction. Better treatment of solvation and electrostatic screening are clear advantages of an MM-GBSA approach over standard empirical scoring functions in docking programs. In a similar ensemble docking study, but using DOCK4.0 and ITScore, Huang and Zhou[78] found that NU6102 and NU2058 achieved significantly better scores and lower RMSD values in the cognate proteins. The deleterious effect of the prearrangement of ligand binding sites on cross-docking results has recently been discussed.[79]

**Pose Ranking and Ligand Strain.** The true binding mode of NU2058 is ranked fifth best in chain C of the protein based on ChemPLP score. However, when the poses are rescored using MM-GBSA, the pose closest to the crystallographic one (cluster rank 5) is unambiguously recognized to have the best score (greater than 4 kcal/mol better than the next best pose; see Table 4c). A possible contributor to the better ranking is the strain energy, which is more accurately accounted for in MM-GBSA as compared with ChemPLP score — the correct pose has the least strain energy (2.1 kcal/mol) as assessed by MM-GBSA, significantly less than the top score based on ChemPLP (strain energy of 5.4 kcal/mol). That such high strain in the ligands is permitted with ChemPLP score may be a reflection of the fast but simplistic ligand clash potential used by ChemPLP.[18] Chain C was used instead of chain A in self-docking calculations so that a common binding pocket could be defined for both inhibitors in ensemble docking. Of the four chains A to D, only chains A and C contain electron density for an inhibitor. Superimposition of chains A and C does not highlight any major differences between the binding sites, other than more water molecules in the binding site of chain C. The fact that the true binding mode is ranked fifth could reflect that this pocket, unlike the other one, was not subject to the bias that comes from minimizing with a cocrystallized ligand.[67,68] This would be consistent with the findings by Spitzer and Jain[79] and Corbeil and Moitessier,[32] which showed that performing self-docking on a protein that has been preminimized with the cognate ligand increases pose prediction success.

*Binding Site Definition.* Successful pose prediction, whether in self- or cross-docking studies, is crucially dependent on data preparation. Thus, perhaps a significant difference that arises between the self- and cross-docking protocol is the definition of the size of binding site. For example, a larger inhibitor (NU6102) is used to define the binding site in the cross-docking study. As such, we reran the cross-docking experiment with the smaller NU2058 ligand to define the ensemble binding site. Once again, the docked pose having the lowest RMSD (1.3 Å) to the cognate ligand was ranked fifth. The top ranked pose was similar to that found by Glide in chain A and described above −180° flip of the purine base along the long axis and three hydrogen bonds to the hinge (Figure 2). The cyclohexyl ring then occupies the space filled by three water molecules in the crystal structure but allowing it to have vdW interactions with Leu134, the gatekeeper residue Phe80, and the carbon side chain of the catalytic lysine (Lys33). Three out of five of the top ranked poses place the cyclohexyl ring in this region. In this particular example, failure to correctly reproduce the cognate pose may be related to inadequate treatment of the waters in the binding site, which were all removed prior to docking. In addition, the cyclohexyl ring is placed in close contact with residues at the start of the flexible glycine-rich loop (Ile10 to Glu12) when binding in the cognate pose (Figure 5), which is not sampled during the rigid docking calculations.

*Alternative Scoring Functions.* Docking success can be scoring function dependent due to the better representation of certain protein characteristics.[32] To test if the fifth place ranking comes from inadequate handling of ligand strain energy and/or unfavorable protein−ligand clashes in the cognate pose, we reran the cross-docking experiment with the slower GOLD scoring function GoldScore, rather than ChemPLP used earlier. It employs a 4−8 potential that allows for close protein−ligand contacts and has force field terms to account for ligand strain.[38] The second and third ranked poses accurately reproduced the cognate ligand structure (Table 4d). The top ranked pose again adopted a pose that fills the vacuum left after the removal of all water molecules (though different to previous poses shown in Figure 6). As shown before, the key energetic component of the MM-GBSA energy that distinguishes the correct pose is the ligand strain (Table 4d, Figure 4). The GoldScore function generated fewer poses with high ligand strain than the faster ChemPLP function,[18] and this enabled the cognate pose to be ranked better.

**Table 4b. ChemPLP Scores and MM-GBSA Binding Affinities for the 17 Clusters of NU6102 (the Inhibitor from 1h1s) in Chain C of Protein 1h1p and in Protein 1h1s**

| | ChemPLP | MM-GBSA (kcal/mol) | | |
|---|---|---|---|---|
| protein PDB ID | average | average | ligand strain (kcal/mol) | no. of clusters |
| 1h1p (cross-dock) | 71.69 ± 4.20 | −55.8 ± 6.76 | 13.34 ± 3.09 | 10 |
| 1h1s (native) | 70.75 ± 4.03 | −69.36 ± 6.49 | 4.64 ± 3.98 | 7 |


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 4c. Relationship between Ranking, ChemPLP, and MM-GBSA Binding Affinity (Including Ligand Strain) and RMSD to the Crystallographic Pose for the Six Output Binding Modes for Docking of NU2058 (the Inhibitor from 1h1p) in Chain C of 1h1p Using GOLD[d]**

| cluster rank | ChemPLP | RMSD (Å) | MM-GBSA (NS) (kcal/mol)[a] | MM-GBSA (kcal/mol) | MM-GBSA rank (NS)[a] | MM-GBSA rank |
|---|---|---|---|---|---|---|
| 1 | 56.40 | 5.00 | −55.34 | −49.97 | 6 | 6 |
| 2 | 55.27 | 4.38 | −67.88 | −54.62 | 1 | 3 |
| 3 | 55.12 | 4.81 | −62.54 | −53.06 | 3 | 5 |
| 4[b] | 53.84 | 2.10 | −58.03 | −53.71 | 5 | 4 |
| 5 | 53.60 | 1.27 | *−63.64* | *−61.50* | 2 | 1 |
| 6[c] | 53.01 | 3.01 | −60.71 | −57.15 | 4 | 2 |

[a]NS − no ligand strain. [b]Single hydrogen bond to the hinge, "partially correct pose" - the position of the cyclohexyl moiety is well produced, but the NH atom of the purine ring (N9) acts as the hinge donor to the carbonyl of Leu83 instead of the amino group as in the crystallographic pose. This results in the purine ring of the docked pose being somewhat displaced relative to the crystallographic one. [c]Three hydrogen bonds are made to the hinge, but the purine base is flipped by 180° relative to the crystallographic one. [d]The pose with the lowest heavy atom deviation from the crystallographic pose is highlighted in italics.

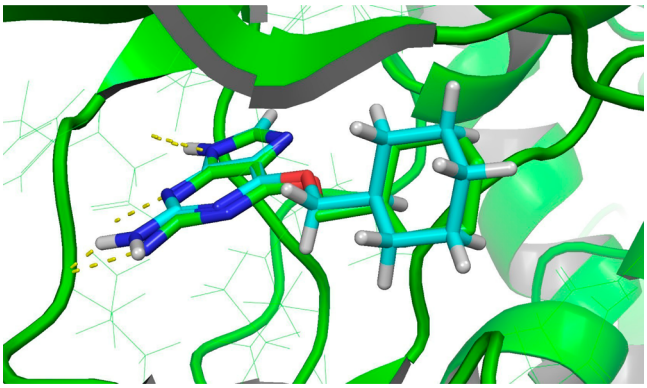


**Figure 5.** 1h1p inhibitor in Chain C (green). Crystal structure ligand is colored green and the pose closest to the cognate pose is in cyan (GOLD ChemPLP ensemble docking).

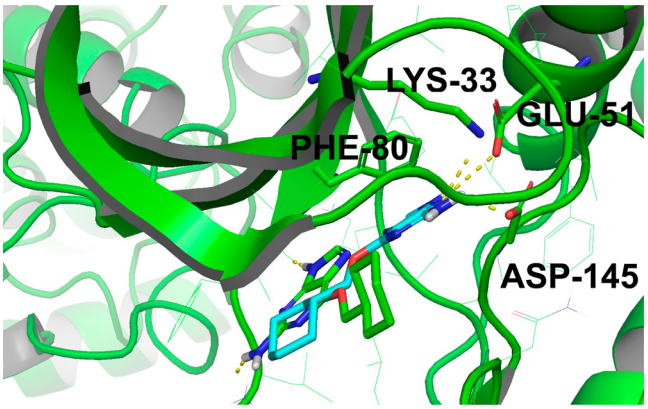



**Figure 6.** Crystal structure of the 1h1p inhibitor in Chain C (green) and top ranked GOLD pose (cyan). Instead of the purine base of the docked pose making hinge-binding interactions, it forms hydrogen bonds with the catalytic lysine (Lys 33), glutamate from helix-C (Glu51), and DFG-motif aspartate (Asp145). It also engages in $\pi$-stacking with the phenylalanine gatekeeper residue (Phe80) in this binding mode.

In addition, we used Glide SP to score-in-place the six clusters (Table 4e) output from the GOLD docking (ChemPLP) of the inhibitor from 1h1p in chain C of the same protein. Local minimization of the ligand was allowed, and calculations were performed with and without the inclusion of ligand strain (Table 4e). The inhibitor with the lowest RMSD to the cognate pose is only ranked first by Emodel when ligand strain is included (inclusion of ligand strain is a nondefault option in Glide). The pose ranks output by Glide mirror those from MM-GBSA scoring very closely (with and without a correction for ligand strain). However, the Emodel energy window between poses is much smaller than that by MM-GBSA scoring, which could be problematic given the concerns about the impact of numerical sensitivities upon docking recently highlighted by Feher and Williams.[80] The lack of a significant score separation between good and bad poses has also been found for other scoring functions.[61]

Properly accounting for ligand strain may be generally important for correct ranking in all scoring algorithms not just the four (MM-GBSA, ChemPLP, GoldScore, and GlideScore/Emodel) that we have looked at here for this particular example, assuming an accurate protein structure is used. However, if a protein structure that cannot accommodate the correct ligand pose is treated rigidly, then obtaining an accurate ligand pose will often require a degree of ligand strain to be accommodated. As recently pointed out by several authors,[24,79] accuracy in cognate pose prediction is often used as a test for comparing docking algorithms, but cross-docking studies would be a more realistic test. Ligand strain is likely to be less in a cognate protein as compared to the noncognate one,[67] so continually performing evaluations based on self-docking

**Table 4d. Relationship between Ranking, GoldScore, and RMSD to the Crystallographic Pose for the Three Output Binding Modes for Docking of NU2058 (the Inhibitor from 1h1p) in Chain C of 1h1p Using GOLD[c]**

| cluster rank | GoldScore | RMSD (Å) | MM-GBSA (NS) (kcal/mol)[a] | MM-GBSA (kcal/mol) | MM-GBSA rank (NS)[a] | MM-GBSA rank |
|---|---|---|---|---|---|---|
| 1 | 54.8 | 5.9 | −63.5 | −56.2 | 1[b] | 3 |
| 2 | 52.9 | 1.3 | −63.7 | −61.2 | 1[b] | 1 |
| 3 | 50.5 | 1.1 | −60.1 | −58.00 | 3 | 2 |

[a]NS − no ligand strain. [b]Equal rank. [c]The reported MM-GBSA binding energies account for ligand strain. Cluster rank 3 is a subset of cluster rank 2. Rotation of the cyclohexyl ring leads to it being defined as a separate cluster.


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 4e. Rescoring with Glide SP (Emodel; Ligand Sampling with Refine Only) of the Six Output Binding Modes from the Docking of NU2058 (the Inhibitor from 1h1p) in Chain C of 1h1p Using GOLD (Table 4c)[c]**

| Emodel rank (NS)[b] | Emodel (NS)[b] | Emodel rank | Emodel[a] | MM-GBSA rank (NS)[b] | MM-GBSA rank | RMSD (Å) |
|---|---|---|---|---|---|---|
| 1 | −82.8 | 3 | −78.5 | 1 | 3 | 4.38 |
| 2 | *−80.0* | 1 | *−80.0* | 2 | 1 | 1.27 |
| 3 | −78.9 | 2 | −78.9 | 4 | 2 | 3.01 |
| 4 | −66.8 | 4 | −66.7 | 5 | 4 | 2.10 |
| **5** | −65.0 | 5 | −64.9 | 6 | 6 | 5.00 |
| 6 | −60.0 | 6 | −59.7 | 3 | 5 | 4.81 |

[a]Apply strain correction terms with a threshold of 4 kcal/mol for strain correction. The excess strain energy is then scaled by 0.25. [b]NS − no ligand strain. [c]RMSD is based on comparison of the non-hydrogen atoms to the crystallographic pose. The pose rankings with and without corrections for ligand strain are reported. The pose with the lowest heavy atom deviation from the crystallographic pose is highlighted in italics.

studies may fail to highlight ligand strain as an important issue for the correct ranking of poses. Tirado-Rives and Jorgensen[81] have pointed out that correct handling of "conformer focusing" (a collective term for ligand strain and ligand conformational entropy) is essential for rank ordering success. This requires force fields capable of computing torsional and intramolecular energetics accurately.

The examples presented here regarding the benefits of including a correction for strain in the scoring function in order to improve pose ranking occur mainly with respect to the CDK2 example. However, the results appear to be consistent with respect to four scoring functions, ChemPLP, GoldScore, Glide SP, and MM-GBSA. We see this effect most clearly in CDK2 with 1h1p because we are considering chains A and C of the same protein with no obvious difference between protein conformations. We observed a difference in behavior between the chains and then explored the reasons including binding site definition and scoring function selection. The top ranked pose of the docking function (when ligand strain is not incorporated) being associated with relatively high strain was a consistent observation (Tables 4c−4e). The pose with the lowest RMSD to the cognate had a strain energy of ∼2 kcal/mol as opposed to the top ranked one from the docking scoring functions which varied between 5, 7, and 13 kcal/mol for ChemPLP, Goldscore, and Glide SP, respectively.

*Screening Power.* "Screening power" is the discrimination of true binders from random molecules.[39] The primary objective of virtual screening is to score large numbers of database compounds (proprietary and/or external vendor) and distinguish between actives and inactives (decoys). The scoring functions (GlideScore[22,38] and ChemPLP[18]) are generally able to recognize the cognate pose as top ranked in self-docking studies in the work presented here. However, ChemPLP shows poor differentiation between the alternative docking poses of CDK2 inhibitors in cross-docking studies. MM-GBSA scoring attributes this to poor handling of ligand strain energy (Table 4c). We decided to further investigate the sensitivity of ChemPLP by using it to dock known inactives with similar properties (decoys). Cutoff scores for activity based on the less active CDK2 inhibitor from 1h1p were assigned (Table 4c, averaged fitness, MM-GBSA, and ligand strain energies) for the 1h1p decoys.

Docking of 1h1p decoy inhibitors with GOLD resulted in 627 clusters; of these 459 had a preference for the protein of the cognate inhibitor (Figure 7). Jain[31] has postulated that true ligands of a binding site should have more good-scoring poses than decoys. It was found that considering the mean instead of the maximum score of a molecule could improve screening enrichments. Based on average scores, MM-GBSA would be expected to distinguish between actives and inactives, but ChemPLP would not. Docking of 1h1s decoy inhibitors with

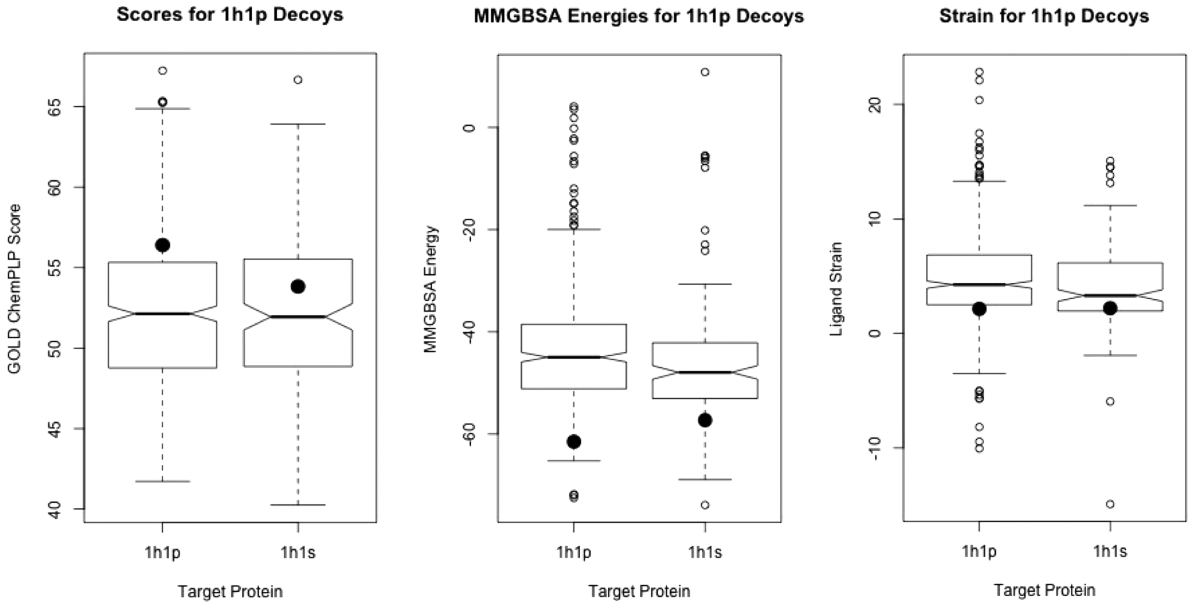



**Figure 7.** Boxplots of ChemPLP scores, MM-GBSA binding affinities (kcal/mol), and ligand strain (kcal/mol) for the 627 clusters from 1h1p decoys in chain C of 1h1p and in 1h1s. The notches represent the uncertainty of the estimation of the median, the boxes indicate the 25 to 75 quartile range, and the whiskers indicate maximum and minimum or 1.5 times the range of the next central quartile, depending on which one yields shorter whiskers. Empty circles represent outliers, and the filled black dot indicates the value for the docked active ligand.


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

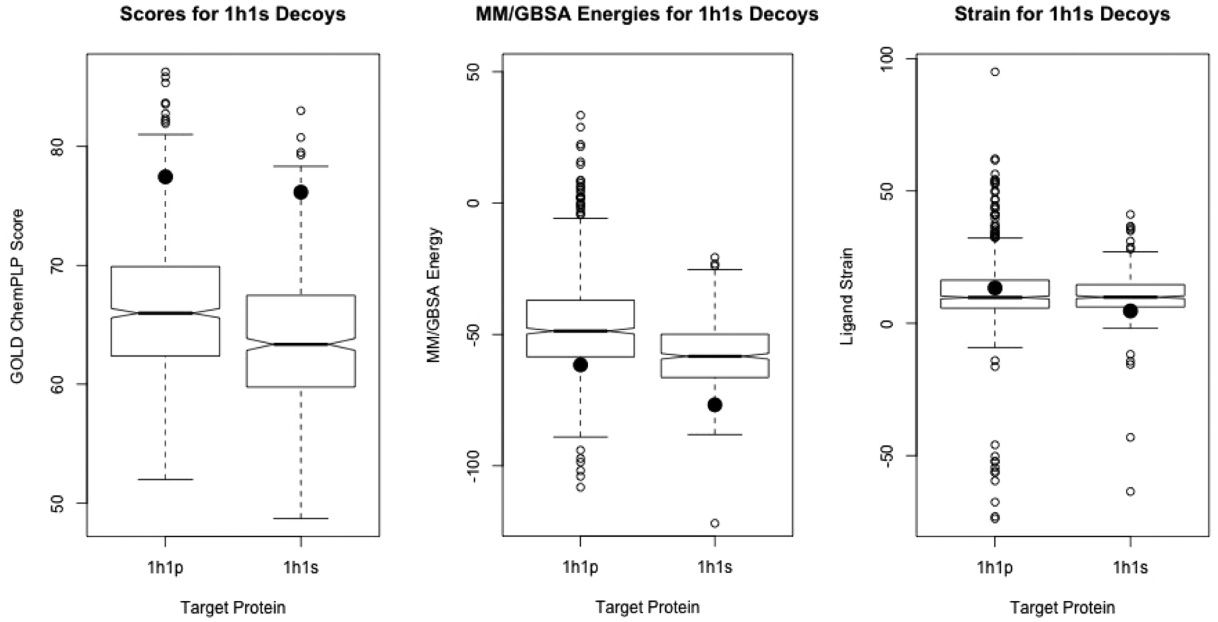



**Figure 8.** Boxplots of ChemPLP scores, MM-GBSA binding affinities (kcal/mol), and ligand strain (kcal/mol) for the 1511 clusters from 1h1s decoys in chain C of 1h1p and in 1h1s. The notches represent the uncertainty of the estimation of the median, the boxes indicate the 25 to 75 quartile range, and the whiskers indicate maximum and minimum or 1.5 times the range of the next central quartile, depending on which one yields shorter whiskers. Empty circles represent outliers, and the filled black dot indicates the value for the docked active ligand.

GOLD resulted in 1511 clusters; of these 592 had a preference for the protein of the cognate inhibitor (Figure 8). The box plots in Figure 7 indicate that ligand strain alone cannot be used as a metric to discriminate actives from inactives; however, ligand strain can be used as a filter to eliminate ligands that truly do not fit the binding site. The box plots in Figure 7 also confirm that ligand scores (docking and MM-GBSA) are generally better with the cognate protein than an alternate protein. With respect to 1h1p, MM-GBSA scoring works better than ChemPLP in ranking the true ligand higher than the decoys. For 1h1s, ChemPLP performs better in this respect. Interestingly, Li et al.[39] identified a strong connection between the docking and screening power of scoring functions. They found this to be a logical relationship, as a scoring function that can discriminate between alternate poses of the same molecule (docking power) should also have good ability to differentiate between the low energy pose of a true binder and the decoy poses of nonbinders (screening power). Overall they found the docking/screening power of scoring functions to be promising.

**Case Study 2**: Using multiple X-ray conformations to simulate protein flexibility

Here we examine the chemotype preferences of ligands for a particular protein conformation using multiple Protein Kinase A (PKA) crystal structure conformations as an example. In a recent docking study, Fischer et al.[82] exploited crystallographic refinement methods that enable the modeling of higher-energy protein conformational states using direct experimental observations. These conformations can be weighted using experimentally derived conformations (occupancies) as a guide, which ensures that no particular conformation of the protein dominates the others. By selecting molecules that score highly and to different protein-receptor conformations, Fischer et al.[82] were able to identify diverse chemotypes for cytochrome *c* peroxidase. They make the observation that using a single experimental structure ensures an accessible state; however, selecting a protein conformation to perform rigid docking based on enrichment biases the retrieval to known chemotypes. Finally, they note that there are over 800 unique proteins in the PDB, each with the requisite density maps to which this method could be applied today.[82]

The aim of this second and final case study is to demonstrate the general applicability and robustness of our MM-GBSA pose reranking approach in noncognate conformations of a protein. For this reason we searched the literature for a sufficiently challenging test set and chose the PKA data set used by Skjærven et al.[83] PKA is a suitable model system to further explore the method, because of the highly flexible nature of the protein upon ligand binding and the large number of crystal structures. With respect to ligands, the data set consists of five chemically diverse, low affinity kinase ligands ($K_D$ 6 to 30 $\mu$M; Figure 9). Skjærven et al.[83] compiled their protein selection by extracting all PKA structures with a sequence identity higher than 98% (88 structures) to the human sequence from the RCSB protein databank.[84] From these 88 structures, 5 representative protein conformation for cross-docking were chosen (PDB IDs 3DND, 3AGM, 1Q8W, 1CDK, and 1CMK). These protein structures encompass fully open apo (1CDK), intermediate (PDB IDs 3DND, 3AGM, and 1Q8W), and fully closed (PDB ID 1CMK) conformational states of the protein. By focusing on a limited number of data points, it allows us to make more detailed structural observations and to interpret docking behavior under varied conditions. The diversity of these structures resulted in poor docking accuracy for both Surflex and Glide in cross-docking studies with rigid receptor structures.[83] Here, we use GOLD as an alternative docking method for this case study.

The results of docking the five ligands to five different X-ray conformations of the same protein (PKA) are tabulated in Table 5. If these are compared with the equivalent Surflex and Glide results,[83] the consensus pose approach as advocated by Houston and Walkinshaw[85] is not beneficial for pose prediction, at least not with respect to significantly different


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

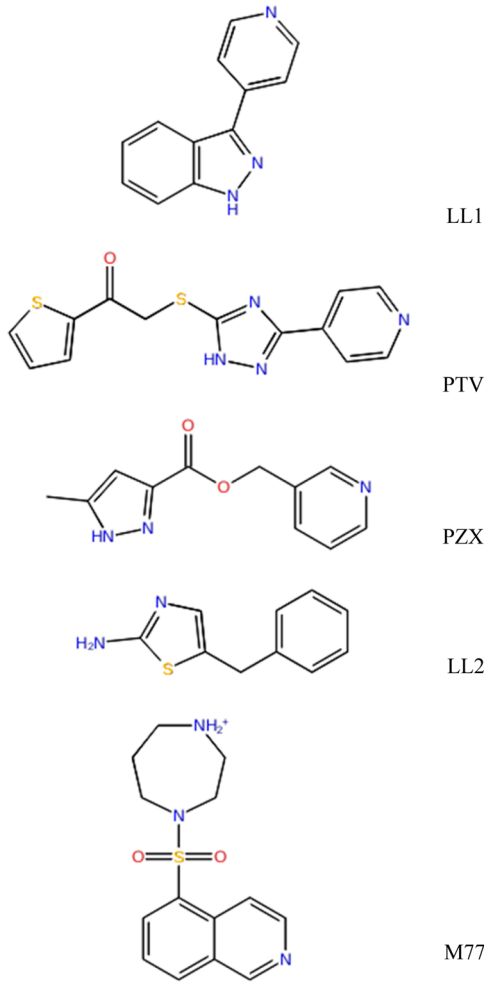



**Figure 9.** Structures of the PKA ligands used for docking.

crystal structure conformations. In addition, there is little consistency with the top ranked poses produced by the different software packages. The ensemble docking results are consistent with the finding of Voigt et al.[86] that the GOLD scoring function is able to distinguish between correct and incorrect poses of a ligand binding to a given protein conformation, but the score is not comparable between different crystal structures of the same target. Different overall top ranked poses are found when a ligand is docked to a single protein at a time, as compared to multiple conformations of the same protein (ensemble) at the same time. This is because some protein conformations generate an intrinsically higher (better) docking score than others and hence can disproportionately populate the final poses. However, the offset in scores does not necessarily reflect a more favorable pose if the relative energetics associated with the different protein conformations is not accounted for.

*Diverse Pose Option.* We observed that 1CDK chain B was better able to replicate cognate poses than chain A (see Score 2 in Table 5). Score 2 reports the percentage of poses within an RMSD to the crystal structure of 2 Å or better. Based purely on visual inspection of the superimposed protein structures, it was not clear to us why this was the case. However, we also noticed that docking to 1CDK chain B produced more distinct docking poses (larger number of clusters) than 1CDK chain A. We hypothesize that this might be the reason for the different docking behavior between chains A and B of 1CDK. Hence, we reran the docking calculations using the diverse option in GOLD, which forces docking poses that are distinct from each other to be sampled. With this option, more cross-docked poses with a RMSD <2 Å to the restraint minimized crystallographic one are produced for all proteins (Table 5B, Score 2). Since we used default settings with the diverse option, 49 clusters were produced for each protein−ligand docking run. All protein structures produced one pose with a RMSD <2 Å to the restraint minimized crystallographic one (Table 5B, Score 2)

**Table 5. Docking Summary**[a]

| | | | | (A) | | | |
|---|---|---|---|---|---|---|---|
| | LL1 | LL2 | M77 | PTV | PZX | Score 1 | Score 2 |
| 3DND | 4.8/- | 5.4/2 | 4.2/2 | 1.0/1 | 0.3/1 | 40% | 80% |
| 3AGM | 5.1/3 | 2.7/3 | 1.8/1 | 1.0/1 | 0.6/1 | 60% | 100% |
| 1Q8W | 5.4/- | 5.2/- | 1.1/1 | 6.6/3 | 1.0/1 | 40% | 60% |
| 1CDK chain A | 5.1/- | 5.1/3 | 3.0/- | 3.6/- | 4.7/- | 0% | 20% |
| 1CDK chain B | 4.9/5 | 8.9/3 | 3.2/- | 2.5/6 | 3.4/- | 0% | 60% |
| 1CMK | 5.5/- | 3.7/- | 3.1/2 | 7.7/- | 1.0/1 | 20% | 40% |
| ensemble | 5.0/- | 5.1/4 | 2.9/2 | 3.6/15 | 4.7/5 | 0% | 80% |
| | | | | (B) | | | |
| | LL1 | LL2 | M77 | PTV | PZX | Score 1 | Score 2 |
| 3DND | 4.8/22 | 5.9/4 | 4.1/2 | 6.1/4 | 0.4/1 | 20% | 100% |
| 3AGM | 5.5/4 | 2.6/6 | 1.8/1 | 2.9/2 | 0.6/1 | 40% | 100% |
| 1Q8W | 5.4/3 | 3.3/4 | 1.1/1 | 6.2/2 | 0.9/1 | 40% | 100% |
| 1CDK chain A | 5.5/4 | 5.1/5 | 3.9/2 | 3.4/31 | 4.6/19 | 0% | 100% |
| 1CDK chain B | 5.0/4 | 8.8/4 | 1.1/1 | 2.5/2 | 2.2/5 | 20% | 100% |
| 1CMK | 3.5/18 | 3.7/- | 3.1/4 | 7.5/- | 8.2/2 | 0% | 60% |

[a]In each cell within the table, the RMSD (Å) of the top ranked pose to the restraint minimized X-ray pose is given first, followed by the ranking of the first docking pose with an RMSD < 2 Å. Score 1 and Score 2 are defined as the percentage of ligands in which the correct binding mode is ranked as number 1 and have an RMSD < 2 Å to the restraint minimized X-ray pose, respectively. (A) Cross-docking of five ligands (columns) using GOLD (clustering of 50 docking attempts) to a set of five representative PKA structures (rows). (B) Cross-docking (50 diverse poses) of five ligands (columns) using GOLD to a set of five representative PKA structures (rows). Ensemble refers to ensemble docking results (simultaneous docking to all 6 protein structures at once).


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

**Table 6. Docking Summary**[a]

LL1 Is the Ligand in Complex with 3DNE: Reference Value 3DNE X-ray Pose −66.5 kcal/mol Window of −6 kcal/mol

| LL1 | cluster size | MM-GBSA binding energy (kcal/mol) X-ray pose | window to (kcal/mol) non-X-ray pose | RMSD of distinct pose (Å) | lowest RMSD (Å) |
|---|---|---|---|---|---|
| 3DND | 2 | −50 | <1 | 5.0 | 2.6 |
| 3AGM | 4 | −49 | −3 | 2.4 | 0.7 |
| 1Q8W | 1 | −42 | −3 | 5.3 | 5.3 |
| 1CDK chain A | 4 | −54 | −3 | 5.5 | 5.5 |
| 1CDK chain B | 7 | −53 | −4 | 2.4 | 1.0 |
| 1CMK | 3 | −39 | +8 | 3.5 | 3.5 |

PZX Is the Ligand in Complex with 4IE9: Reference Value 4IE9 X-ray Pose −63.8 kcal/mol Window of −13 kcal/mol

| PZX | cluster size | MM-GBSA binding energy (kcal/mol) X-ray pose | window to (kcal/mol) non-X-ray pose | RMSD of distinct pose (Å) | lowest RMSD (Å) |
|---|---|---|---|---|---|
| 3DND | 7 | −58 | −15 | 3.7 | 0.3 |
| 3AGM | 9 | −46 | +1 | 2.7/3.5 | 1.0 |
| 1Q8W | 7 | −42 | <1 | 3.1 | 0.9 |
| 1CDK chain A | 4 | −49 | +9 | 5.2 | 5.2 |
| 1CDK chain B | 14 | −48 | +7 | 3.0 | 1.1 |
| 1CMK | 6 | −52 | −6 | 2.0/6.0 | 0.7 |

M77 Is the Ligand in Complex with 1Q8W: Reference Value 1Q8W X-ray Pose −73.5 kcal/mol Window of −17 kcal/mol

| M77 | cluster size | MM-GBSA binding energy (kcal/mol) X-ray pose | window to (kcal/mol) non-X-ray pose | RMSD of distinct pose (Å) | lowest RMSD (Å) |
|---|---|---|---|---|---|
| 3DND | 8 | −75 | −17 | 3.1 | 0.6 |
| 3AGM | 16 | −71 | −15 | 2.6 | 0.8 |
| 1CDK chain A | 8 | −74 | −20 | 8 | 3.2 |
| 1CDK chain B | 9 | −74 | −21 | 3.9 | 3.9 |
| 1CMK | 4 | −54 | −4 | 3.8 | 2.0 |

LL2 Is the Ligand in Complex with 3DND: Reference Value LL2 X-ray Pose −52 kcal/mol window of −6 kcal/mol

| LL2 | cluster size | MM-GBSA binding energy (kcal/mol) X-ray pose | window to (kcal/mol) non-X-ray pose | RMSD of distinct pose (Å) | lowest RMSD (Å) |
|---|---|---|---|---|---|
| 1Q8W | 4 | −43 | +4 | 5.3 | 1.6 |
| 3AGM | 9 | −48 | +2 | 5.4 | 1.5 |
| 1CDK chain A | 12 | −53 | +4 | 5.3 | 0.8 |
| 1CDK chain B | 13 | −53 | +3 | 5.4 | 1.2 |
| 1CMK | 12 | −45 | +6 | 3.7 | 3.1 |

PTV Is the Ligand in Complex with 4IJ9: The Reference Value for the PTV X-ray Pose Is −91 kcal/mol with an Energy Window of −26 kcal/mol

| PTV | cluster size | MM-GBSA binding energy (kcal/mol) X-ray pose | window to (kcal/mol) non-X-ray pose | RMSD of distinct pose (Å) | lowest RMSD (Å) |
|---|---|---|---|---|---|
| 1Q8W | 18 | −62 | +3 | 1.6 | 0.8 |
| 3AGM | 11 | −65 | −9 | 2.3 | 1.6 |
| 1CDK chain A | 14 | −71 | −6 | 3.2 | 2.1 |
| 1CDK chain B | 22 | −71 | −8 | 3.1 | 1.7 |
| 3DND | 17 | −66 | +3 | 1.6 | 0.6 |
| 1CMK | 24 | −46 | +19 | 6.7 | 3.6 |

[a]MM-GBSA scoring of cross-docking results (GOLD) of five ligands to a set of five representative PKA structures (rows). The reference energy of the crystallographic pose in the cognate protein is given and the energy difference (window) to the next distinct pose identified by RMSD value. The alternate poses for comparison to the cognate pose are the ensemble docking results for that ligand.

for all ligands with the exception of 1CMK (fully closed conformation). In contrast, without this enforced diversity docking option, only 3AGM (intermediate conformation) produced a score of 100% for Score 2 (Table 5A). Given the stochastic nature of GOLD docking, the results for the top ranked poses are very consistent, with and without the diverse docking option (Table 5 A and B).

*Pose Ranking MM-GBSA*. Graves et al.[10] have concluded that the MM-GBSA scoring function should be considered as complementary but not superior to current rigid docking/scoring protocols. They opine that rescoring the results of rigid docking via a MM-GBSA scoring protocol in which the complex is allowed to be flexible not only can rescue false negatives but also can introduce false positives. The rescued molecules tend to be large and poorly accommodated by the rigid receptor but have a geometry that resembles the crystallographic pose. The false positives arise from permissive binding sites and inaccurately accounting for protein strain. Mobley and Dill[87] have described strain as an "invisible" energy cost and point out that ligand-induced changes in protein conformation are not uncommon. Indeed, in a study of 206 binding sites Boström et al.[88] showed that 83% of cases displayed significant changes in the binding sites between pair members (each pair was made up of the same protein, but similar although different ligand), the most frequent differences being water architecture and side-chain conformations. To counter this, an offset penalty to account for the internal energy differences between distinct protein conformations is used.[89]

In Table 6 we establish that the cognate pose can be distinguished from decoy poses (ensemble docking results) by the MM-GBSA scoring function. The MM-GBSA binding energy of the cognate pose is derived from the alignment of all the protein structures, including those not used for docking but that contain the cognate ligand. The ligand with these starting coordinates is then minimized in the respective proteins. Thus, in this way a benchmark for the energy of the cognate pose can be obtained, even when the docking protocol does not replicate the cognate pose. Poses with a low RMSD to the cognate pose, and that are top-ranked by the docking scoring function, have a propensity toward a large MM-GBSA energy difference with any alternate noncognate pose. Protein conformations that do not give a high rank to the cognate pose (Tables 5 and 6) tend to have a smaller MM-GBSA energy window. This is an especially striking effect for pose ranking with the diverse option (Table 5B). We use Table 5B as the comparator since poses with a low RMSD to the cognate pose are sampled. Thus,


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

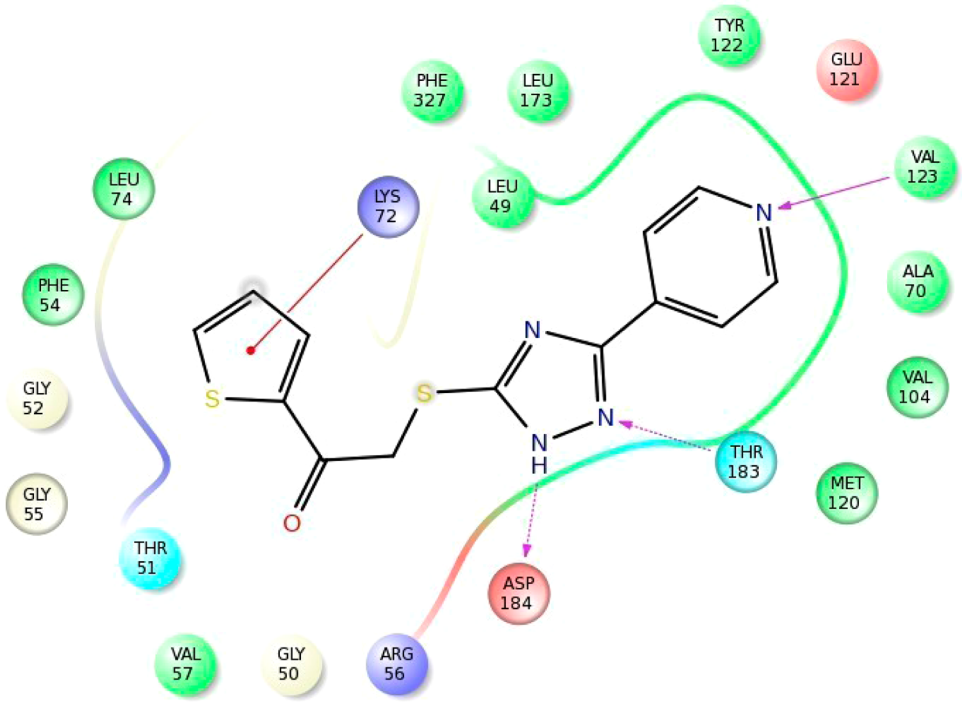



**Figure 10.** Interactions of PTV in PDB structure 1CMK.

the scoring function has assessed them and ranked them against alternate poses. In Table 5A, the cognate pose may not have been sampled. In Table 6, the ensemble docking results from Table 5 (nondiverse docking option) are scored using MM-GBSA.

The cognate pose of ligand LL1 is ranked 22 and 18 out of 49 poses in PDB IDs 3DND and 1CMK, respectively (Table 5B). In 1CMK, the cognate pose of LL1 is the worse pose by 8 kcal/mol according to MM-GBSA scoring (Table 6). In 3DND, the ligand LL1 is within an energy window of 1 kcal/mol of an alternate pose. In contrast, the other protein structures favor the cognate pose over alternate poses with a larger energy window (Table 6), consequently giving the cognate pose a better rank (Table 5B). The ligand PZX is ranked 19 and 5 in 1CDK, chains A and B, respectively (Table 5B), and the cognate pose is less favorable than an alternate pose by at least 7 kcal/mol (Table 6). The cognate pose is ranked either first or second by the other proteins with respect to GoldScore and the MM-GBSA scoring scheme. The cognate pose of the ligand M77 displays a large energy window to alternate poses and is highly ranked in all protein structures (Tables 5 and 6). In contrast, the ligand LL2 is ranked 4 or worse in all proteins, and the X-ray pose is disfavored by between 2 and 6 kcal/mol in these proteins. That the ligand PTV is not top ranked in 1Q8W, 3DND, and 1CMK is consistent with the MM-GBSA calculations. The MM-GBSA result for 1CDK chain A less clearly explains the low pose ranking. However, for this protein structures only one hydrogen bond is made between the cognate pose of the ligand with the rigid protein (pyridine nitrogen atom and NH of Val123) as opposed to three in the cognate protein (Figure 10). Thus, there are "many" similar poses that make this hydrogen bond interaction but show variations in the rest of the ligand structure.

Based on the above analysis, it could be argued that the docking scoring functions are performing correctly because poses with a low RMSD to the cognate pose only get a good MM-GBSA score in the context of the cognate receptor structure — accurate ligand poses with low RMSD get a poor score in the noncognate receptor structure, suggesting that analysis of the ligand pose alone is not sufficient to determine if a pose is good enough to yield a good score. The critical, and perhaps obvious, finding here is that the entire protein−ligand complex must be correct to produce a good score. In some cases, the pose of a ligand in a noncognate protein conformation differs from the correct pose, which may in fact be the best possible pose in the context of the noncognate receptor structure, again emphasizing the importance of protein sampling. An accurate assessment of ligand strain may be a deciding factor in determining the most energetically favorable pose; therefore, such a pose should not be regarded as a pose reproduction failure.[90]

Whereas in a self-docking study the criteria for defining success are clear, in a cross-docking study such criteria are harder to define when the receptor is kept rigid. Broadly speaking, it can be said that for the purposes of enrichment the ranking is more important, whereas in design accurate correct pose reproduction is more important. However, in practice both are important in most scenarios. There is a growing consensus that since docking scoring functions struggle to consistently rank poses correctly, poses should be rescored with a higher level scoring function such as MM-GBSA or QM/MM.[90−92] Irrespective of the level of the scoring function, in order to function optimally, the correct pose is required, so multiple poses need to be retained as well as including protein ensembles in the docking.[92,93] The publication of Zhang et al.[94] demonstrates that with high performance computing rescoring multiple poses using MM-GBSA is computationally feasible.

## ■ CONCLUSION

The work presented here suggests that sampling of the system as a whole (ligand, protein, and waters) is necessary to achieve optimal docking results and that current docking scoring


dx.doi.org/10.1021/ci5003735 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

functions may be sufficiently accurate given that an adequate structure for the complex can be found. A significant amount of time and effort is spent postprocessing hit lists generated by various means;[95,96] perhaps if the significant increases in computational capacity (both the speed of processors and the availability of more computers) are devoted to improved ligand sampling, it will be possible to considerably reduce or eliminate post processing steps. The expectation that docking scoring functions should give a good correlation with experimental binding data may be unrealistic.[7,15,76] They should simply prioritize the poses that complement the protein binding site the best. As found by Li et al.,[39] we also find the docking and screening power of the docking functions studied here to be far superior to the scoring power. That is reassuring, given that the intended applications of docking programs are primarily pose prediction and virtual screening.

It has been said that progress in docking has reached a plateau,[97] which may be true if we do not more accurately account for receptor flexibility. The ensemble docking results show that without a "correct" protein conformation, ligands cannot be docked well. Thus, it seems essential to better represent protein flexibility[88,98] in the quest for better docked ligands. This point cannot be overemphasized, since no amount of scoring function optimization and fitting can account for an incorrect protein conformation.[71] We showed that MM-GBSA can be used as an independent scoring function to assess the energetically preferred pose as generated with multiple scoring functions, and in multiple protein conformations. Thus, the role of MM-GBSA may be to distinguish between true and decoy poses of a ligand in addition to the rescoring of data sets[14,99] The CDK2 example showed that it is important to incorporate *ligand strain* into the scoring function.

Machine learning and consensus methods have demonstrated improved pose ranking in self-docking studies. Their utility for use in cross-docking has been less rigorously evaluated however. Hence, scoring an ensemble of ligand poses with e.g. MM-GBSA in multiple protein conformations may be the preferred tried and tested method. Our studies suggest that as few as five distinct ligand poses may be sufficient when using optimal internal software settings in an appropriate receptor conformation.[100] This is significantly less than for externally generated ligand ensembles and may be more practical for screening large data sets.[77] Ready experimental access to multiple protein conformations from crystallographic data may increase the value and significance of our investigation.[82]

## ■ ASSOCIATED CONTENT

### Supporting Information

Supporting Information for (i) Two Cross-docking case Studies (fXa and hRAR), (ii) Compilation of water-mediated interactions of Table 1 complexes, (iii) Summary of MM-GBSA protein-ligand interaction energies for GOLD (CHEMPLP) dockings of Table 1 ligands. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: Paulette.Greenidge@novartis.com.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors would like to thank Dr. Tjelvar Olsson from CCDC (GOLD) for providing the scripts to enable us to perform scoring in place calculations. We also thank Romain Wolf for helpful initial discussions about the manuscript and the reviewers for their critical feedback.

## ■ REFERENCES

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

(2) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.

(3) Wallach I. *Improving Posing and Ranking of Molecular Docking.* PhD Thesis.http://hdl.handle.net/1807/34955 (accessed Sept 14, 2014).

(4) Thomas, M. P.; McInnes, C.; Fischer, P. M. Protein Structures in Virtual Screening: A Case Study with CDK2. *J. Med. Chem.* **2006**, *49*, 92−104.

(5) Korb, O.; Ten Brink, T.; Victor Paul Raj, F. R.; Keil, M.; Exner, T. E. Are predefined decoy sets of ligand poses able to quantify scoring function accuracy? *J. Comput.-Aided Mol. Des.* **2012**, *26*, 185−197.

(6) Mysinger, M. M.; Shoichet, B. K. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561−1573.

(7) Huang, N.; Schoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(8) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J. Chem. Inf. Model.* **2008**, *48*, 1411−1422.

(9) Rastelli, G.; Degliesposti, G.; Del Rio, A.; Sgobba, M. Binding Estimation after Refinement, a New Automated Procedure for the Refinement and Rescoring of Docked Ligands in Virtual Screening. *Chem. Biol. Drug Des.* **2009**, *73*, 283−286.

(10) Graves, A. P.; Shivakumar, D. M.; Boyce, S. E.; Jacobson, M. P.; Case, D. A.; Shoichet, B. K. Rescoring Docking Hit Lists for Model Cavity Sites: Predictions and Experimental Testing. *J. Mol. Biol.* **2008**, *377*, 914−934.

(11) Guimarães, C. R. W.; Cardozo, M. MM-GB/SA Rescoring of Docking Poses in Structure-based Lead Optimization. *J. Chem. Inf. Model.* **2008**, *48*, 958−970.

(12) Lyne, P. D.; Lamb, M. L.; Saeh, J. C. Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring. *J. Med. Chem.* **2006**, *49*, 4805−4808.

(13) Thompson, D. C.; Humblet, C.; Joseph-McCarthy, D. Investigation of MM-PBSA Rescoring of Docking Poses. *J. Chem. Inf. Model.* **2008**, *48*, 1081−1091.

(14) Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P. Physics-Based Scoring of Protein-Ligand Complexes: Enrichment of Known Inhibitors in Large-Scale Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 243−253.

(15) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

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

(17) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225−242.

(18) Korb, O.; Stuetzle, T.; Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.

(19) Tuccinardi, T.; Botta, M.; Giordano, A.; Martinelli, A. Protein Kinases: Docking and Homology Modeling Reliability. *J. Chem. Inf. Model.* **2010**, *50*, 1432−1444.
(20) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47−57.
(21) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.
(22) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(23) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599−1608.
(24) Liebeschuetz, J. W.; Cole, J. C.; Korb, O. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 737−748.
(25) Cole, J. C.; Murray, C. W.; Nissink, J.W. M.; Taylor, R. D.; Taylor, R. Comparing Protein-Ligand Docking Programs Is Difficult. *Proteins* **2005**, *60*, 325−332.
(26) Sousa, F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 15−26.
(27) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore$^{CSD}$-Knowledge-Based Scoring Function derived from Small Molecule Crystal data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **2005**, *48*, 6296−6303.
(28) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in Molecular Recognition. 2. Assessing and Improving Cross-Docking Accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293−2302.
(29) Lupyan, D.; Abramov, Y. A.; Sherman, W. Close intramolecular sulfur−oxygen contacts: modified force field parameters for improved conformation generation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 1195−1205.
(30) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534−546.
(31) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modelling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281−306.
(32) Corbeil, C. R.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model.* **2009**, *49*, 997−1009.
(33) Hsieh, J. H.; Yin, S.; Wang, X. S.; Liu, S.; Dokholyan, N. V.; Tropsha, A. Cheminformatics meets molecular mechanics: a combined application of knowledge-based pose scoring and physical force field-based hit scoring functions improves the accuracy of structure-based virtual screening. *J. Chem. Inf. Model.* **2012**, *52*, 16−28.
(34) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.
(35) Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2010**, *32*, 742−755.
(36) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(37) Jones, G.; Willet, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(38) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(39) Li, Y.; Han, L.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717−1736.
(40) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(41) Sastry, G. M.; Inakollu, V. S. S.; Sherman, W. Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking. *J. Chem. Inf. Model.* **2013**, *53*, 1531−1542.
(42) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.
(43) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.
(44) Greenidge, P. A.; Kramer, C.; Mozziconacci, J.-C.; Wolf, R. M. MM/GBSA Binding Energy Prediction on the PDBbind Data Set: Successes, Failures, and Directions for Further Improvement. *J. Chem. Inf. Model.* **2013**, *53*, 201−209.
(45) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.
(46) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B., Jr.; Stuckey, J. A.; Carlson, H. A. CSAR Benchmark Exercise 2011−2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. Model.* **2013**, *53*, 1853−1870.
(47) Broccatelli, F.; Brown, N. Best of Both Worlds: On the Complementarity of Ligand-Based and Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2014**, *54*, 1634−1641.
(48) Li, J.; Abel, R.; Zhu, K.; Cao, Y.; Zhao, S.; Friesner, R. A. The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins* **2011**, *79*, 2794−2812.
(49) Notenboom, V.; Williams, S. J.; Hoos, R.; Withers, S. G.; Rose, D. R. Detailed Structural Analysis of Glycosidase/Inhibitor Interactions: Complexes of Cex from *Cellulomonasfimi* with Xylobiose-Derived Aza-Sugars. *Biochemistry* **2000**, *39*, 11553−11563.
(50) Dow, R. L.; Schneider, S. R.; Paight, E. S.; Hank, R. F.; Chiang, P.; Cornelius, P.; Lee, E.; Newsome, W. P.; Swick, A. G.; Spitzer, J.; Hargrove, D. M.; Patterson, T. A.; Pandit, J.; Chrunyk, B. A.; LeMotte, P. K.; Danley, D. E.; Rosner, M. H.; Ammirati, M. J.; Simons, S. P.; Schulte, G. K.; Tate, B. F.; DaSilva-Jardine, P. Discovery of a novel series of 6-azauracil-based thyroid hormone receptor ligands: potent, TR$\beta$ subtype-selective thyromimetics. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 379−382.
(51) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221−234.
(52) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions. 1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(53) Chen, I.-J.; Foloppe, N. Drug-like Bioactive Structures and Conformational Coverage with the LigPrep/Confgen Suite: Comparison to Programs MOE and Catalyst. *J. Chem. Inf. Model.* **2010**, *50*, 822−839.
(54) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin, Heidelberg, 2008; pp 319− 326.
(55) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins,

energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(56) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(57) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509−1519.
(58) Wang, R.; Xueliang, F.; Yipin, Lu.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(59) Li, Y.; Liu, Z.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.
(60) Hawkins, P. C. D.; Kelley, B. P.; Warren, G. L. The Application of Statistical Methods to Cognate Docking: A Path Forward ? *J. Chem. Inf. Model.* **2014**, *54*, 1339−1355.
(61) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 775−786.
(62) Baber, J. C.; Thompson, D. C.; Cross, J. B.; Humblet, C. GARD: a Generally Applicable Replacement for RMSD. *J. Chem. Inf. Model.* **2009**, *49*, 1889−1900.
(63) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179−190.
(64) Feher, M.; Williams, C. I. Effect of input differences on the results of docking calculations. *J. Chem. Inf. Model.* **2009**, *49*, 1704−1714.
(65) GOLD User Guide. www.ccdc.cam.ac.uk/Lists/DocumentationList/gold.pdf (accessed Sept 14, 2014).
(66) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214−2225.
(67) Saranya, N.; Jeyakanthan, J.; Selvaraj, S. Impact of protein binding cavity volume (PCV) and ligand volume (LV) in rigid and flexible docking of protein-ligand complexes. Bioorg. *Med. Chem. Lett.* **2012**, *22*, 7593−7597.
(68) Jain, A. N. Effects of Protein Conformation in Docking: Improved Pose Prediction through Protein Pocket Adaption. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 355−374.
(69) Osguthorpe, D. J.; Sherman, W.; Hagler, A. T. Generation of receptor structural ensembles for virtual screening using binding site shape analysis and clustering. *Chem. Biol. Drug Des.* **2012**, *80*, 182−193.
(70) Osguthorpe, D. J.; Sherman, W.; Hagler, A. T. Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. *J. Phys. Chem. B* **2012**, *116*, 6952−6959.
(71) Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.
(72) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534−553.
(73) Totrov, M.; Abagyan, R. Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. *Proteins: Struct., Funct., Bioinf. (Suppl. 1)* **1997**, *129*, 215−220.
(74) Nabuurs, S. B.; Wagener, M.; De Vlieg, J. A Flexible Approach to Induced Fit Docking. A Flexible Approach to Induced Fit Docking. *J. Med. Chem.* **2007**, *50*, 6507−6518.
(75) Moitessier, N.; Therrien, E.; Hanessian, S. A Method for Induced-Fit Docking, Scoring, and Ranking of Flexible Ligands. Application to Peptidic and Pseudopeptidic $\beta$-secretase (BACE 1) Inhibitors. *J. Med. Chem.* **2006**, *49*, 5885−5894.
(76) Korb, O.; Olsson, T. S. G.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and Limitations of Ensemble Docking. *J. Chem. Inf. Model.* **2012**, *52*, 1262−1274.
(77) Kotasthane, A.; Mulakala, C.; Viswanadhan, V. N. Applying conformational selection theory to improve crossdocking efficiency in 3-phosphoinositide dependent protein kinase-1. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 436−451.
(78) Huang, S. Y.; Zou, X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 399−421.
(79) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking Benchmark and Real-World Application. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 687−699.
(80) Feher, M.; Williams, C. I. Numerical Errors and Chaotic Behavior in Docking Simulations. *J. Chem. Inf. Model.* **2012**, *52*, 724−738.
(81) Tirado-Rives, J.; Jorgensen, W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **2006**, *49*, 5880−5884.
(82) Fischer, M.; Coleman, R. G.; Fraser, J. S.; Shoichet, B. K. Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. *Nature* **2014**, *6*, 575−583.
(83) Skjærven, L.; Codutti, L.; Angelini, A.; Grimaldi, M.; Latek, D.; Monecke, P.; Dreyer, M. K.; Carlomagno, T. Accounting for Conformational Variability in Protein-Ligand Docking with NMR-Guided Rescoring. *J. Am. Chem. Soc.* **2013**, *135*, 5819−5827.
(84) Berman, H.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.
(85) Houston, D. R.; Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.* **2013**, *53*, 384−390.
(86) Voigt, J. H.; Elkin, C.; Madison, V. S.; Duca, J. S. Cross-Docking of Inhibitors into CDK2 Structures. 2. *J. Chem. Inf. Model.* **2008**, *48*, 669−678.
(87) Mobley, D. L.; Dill, K. A. Binding of Small Molecule Ligands to Proteins: "What You See" Is Not Always "What You get". *Structure* **2009**, *17*, 489−498.
(88) Boström, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* **2006**, *49*, 6716−6725.
(89) Kufareva, I.; Abagyan, R. Type-II Kinase Inhibitors for Docking, Screening, and Profiling Using Modified Structures of Active Kinase States. *J. Med. Chem.* **2008**, *51*, 7921−7932.
(90) Gleeson, M. P.; Gleeson, D. QM/MM As a Tool in Fragment Based Drug Discovery. A Cross-Docking, Rescoring Study of Kinase Inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 1437−1448.
(91) Sippl, W.; Wichapong, K.; Rohe, A.; Platzer, C.; Slynko, I.; Erdmann, F.; Schmidt, M. Application of Docking and QM/MM-GBSA Rescoring to Screen for Novel Myt1 Kinase Inhibitors. *J. Chem. Inf. Model.* **2014**, *54*, 881−893.
(92) Slynko, I.; Scharfe, M.; Rumpf, T.; Eib, J.; Metzger, E.; Schüle, R.; Jung, M.; Sippl, W. Virtual Screening of PRK1 Inhibitors: Ensemble Docking, Rescoring Using Binding Free Energy Calculation and QSAR Model Development. *J. Chem. Inf. Model.* **2014**, *54*, 138−150.
(93) Duca, J. S.; Elkin, C.; Madison, V. S.; Voigt, J. H. Cross-Docking of Inhibitors into CDK2 Structures. 1. *J. Chem. Inf. Model.* **2008**, *48*, 659−668.
(94) Zhang, X.; Wong, S. E.; Lightstone, F. C. Towards Fully Automated High Performance Computing Drug Discovery: A Massively Parallel Virtual Screening Pipeline for Docking and Molecular Mechanics/Generalized Born Surface Area Rescoring to Improve Enrichment. *J. Chem. Inf. Model.* **2014**, *54*, 324−337.
(95) Greenidge, P. A.; Carlsson, B.; GöranBladh, L.-G.; Gillner, M. Pharmacophores Incorporating Numerous Excluded Volumes Defined by X-ray Crystallographic Structure in Three-Dimensional Database Searching: Application to the Thyroid Hormone Receptor. *J. Med. Chem.* **1998**, *41*, 2503−2512.

(96) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235−249.
(97) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.
(98) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *J. Mol. Biol.* **2004**, *337*, 1161−1182.
(99) Zou, X.; Sun, Y.; Kuntz, I. D. Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized Born Model. *J. Am. Chem. Soc.* **1999**, *121*, 8033−8043.
(100) Lindström, A.; Edvinsson, L.; Johansson, A.; Andersson, C. D.; Andersson, I. E.; Raubacher, F.; Linusson, A. Postprocessing of Docked Protein-Ligand Complexes Using Implicit Solvation Models. *J. Chem. Inf. Model.* **2011**, *51*, 267−282.