This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

ACS OMEGA

Article

Cite This: ACS Omega 2018, 3, 3655−3664

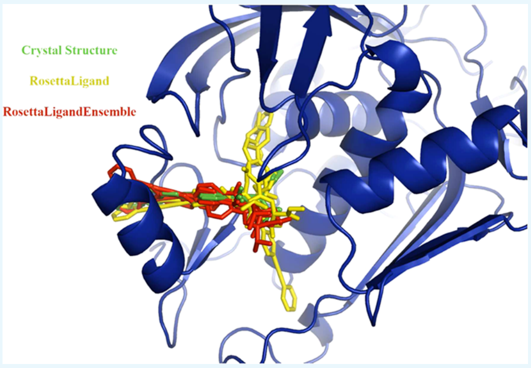
RosettaLigandEnsemble: A Small-Molecule Ensemble-Driven Docking Approach

Darwin Yu Fu and Jens Meiler*

Center for Structural Biology, Department of Chemistry, Vanderbilt University, 465 21st Avenue South, Nashville, Tennessee 37221, United States

S *Supporting Information*

ABSTRACT: RosettaLigand is a protein−small-molecule (ligand) docking software capable of predicting binding poses and is used for virtual screening of medium-sized ligand libraries. Structurally similar small molecules are generally found to bind in the same pose to one binding pocket, despite some prominent exceptions. To make use of this information, we have developed RosettaLigandEnsemble (RLE). RLE docks a superimposed ensemble of congeneric ligands simultaneously. The program determines a well-scoring overall pose for this superimposed ensemble before independently optimizing individual protein−small-molecule interfaces. In a cross-docking benchmark of 89 protein−small-molecule co-crystal structures across 20 biological systems, we found that RLE improved sampling efficiency in 62 cases, with an average change of 18%. In addition, RLE generated more consistent docking results within a congeneric series and was capable of rescuing the unsuccessful docking of individual ligands, identifying a nativelike top-scoring model in 10 additional cases. The improvement in RLE is driven by a balance between having a sizable common chemical scaffold and meaningful modifications to distal groups. The new ensemble docking algorithm will work well in conjunction with medicinal chemistry structure−activity relationship (SAR) studies to more accurately recapitulate protein−ligand interfaces. We also tested whether optimizing the rank correlation of RLE-binding scores to SAR data in the refinement step helps the high-resolution positioning of the ligand. However, no significant improvement was observed.



■ INTRODUCTION

Ligand Docking and Structure-Based Drug Discovery. Structure-based drug discovery and optimization is a critical technique at the intersection of pharmacology and structural biology. Structure-based computer-aided drug discovery (SB-CADD) is a powerful way to create hypotheses based on ligand-binding poses and specific predicted protein/ligand interactions that guide the design of improved small molecules.[1] These hypotheses can be tested by a variety of experimental approaches including fluorescence-binding studies, calorimetric measurements, NMR spectroscopic studies, or X-ray crystallography, often comparing multiple ligands and/or wild-type with mutant proteins.[2] For SB-CADD to maximize its impact on drug discovery, it is necessary for computational ligand docking methodologies to effectively identify correct protein−ligand-binding positions.

Structure−activity relationships (SARs) refer to differences in binding affinity or biological efficacy following chemical scaffold derivatizations. Medicinal chemistry makes use of such minor modifications to optimize lead compounds for desired affinity and other pharmacological properties. This creates a massive wealth of SAR data on related ligands for a single protein target. The PubChem database alone contains over 200 million measurements of biological activities on approximately 10 000 protein targets.[3] BindingDB specifically organizes a portion of its database into collections of congeneric ligands with at least one co-crystallized with the common protein target.[4] It is generally expected that highly similar ligands form similar interactions when binding to the same target.[5] We hypothesize that a docking algorithm that leverages this information can eliminate a portion of false-positive binding poses, i.e., poses that score well, but are incorrect.

Inconsistent Performance of Existing Protein−Ligand Docking Tools. RosettaLigand,[6,7] a small docking tool within the Rosetta structural biology modeling software suite,[8] is one of several algorithms developed for this purpose in the last few decades. AutoDock,[9] DOCK,[10] and Glide[11] are other popular methods, all of which differ in sampling and/or scoring techniques. The performance of these docking tools is not always consistent across systems. A 2013 docking study using the PDBBind data set evaluated scoring functions for decoy discrimination and scoring correlation. The success rate for identifying correct binding modes from decoys was significantly higher than that for identifying weak, middle, and strong

Received: December 24, 2017
Accepted: March 20, 2018
Published: April 2, 2018

binders within a related ligand series.[12] Similar results were obtained in the 2012 Community Structure Activity Resource (CSAR) evaluation, which found that even when docking software was able to recover correct binding poses for a given ligand, few could consistently rank order active ligands.[13] The recent D3R Grand Challenge reaffirmed these findings and noted that docking performance varied even within the same congeneric series. In addition, the overall success of a docking method was dependent on its preparatory workflow.[14] This performance gap between docking and ranking is likely due to the steep energy landscape observed near-native binding modes for high-affinity protein−ligand complexes. Small perturbations in these regions generally resulted in drastic scoring changes.[15]

Use of Structure Ensembles in Docking. Ensemble methods have traditionally been independently approached from the protein and ligand sides. Protein ensembles are a common way of capturing conformational diversity during rigid receptor docking simulations. This need for a structure ensemble can be due to the inherent flexibility of the protein (conformational selection) and/or an induced fit effect on ligand binding. Protein structural ensembles can be generated from experimental determination such as NMR or through computational methods such as molecular dynamics. One such preparation is the relaxed complex scheme that generates a set of receptor targets for docking.[16] To emulate induced fit with ligand binding, Glide docking can be used to convert all interface residues into alanine to allow for sampling the binding pocket without bias from initial side-chain orientations.[17] For scoring purposes, protein ensembles can be handled by an "average energy grid" that scores over the ensemble[18] or by using a selection method to identify a single template mid-simulation.[19] Feixas et al. and Sinko et al. further review the use of multiple receptor structures in drug discovery and design.[20,21]

Ligand structural ensembles are used to represent both ligand conformations and pharmacophore information from multiple ligands. Molecular mechanics or fragment-based sampling can be used to generate conformations before docking.[22] Hybrid methods incorporate information from multiple ligands to better position a given target. For example, HybridDock performs predocking alignment via pharmacophore matching with similar molecules.[23] However, these methods require related co-crystal structures to be readily applicable.

It has been observed that use of well-chosen structural ensembles is advantageous over docking with a single structure, particularly when ensemble proteins are co-crystallized with molecules of similar chemical structure.[24,25] In this manuscript, we developed a two-stage algorithm for ensemble docking of multiple related ligands into a single protein structure.

Incorporating Ligand Ensemble Docking into RosettaLigand. RosettaLigand models protein−ligand interactions with full ligand- and protein-binding pocket flexibility. This is achieved with pregenerated ligand conformations and protein side-chain rotamer libraries.[6,7] RosettaLigand is currently capable of docking multiple ligands simultaneously, but only in the sense that they bind the protein jointly (e.g., a small molecule together with a key bridging water molecule or a cofactor with metal ion bound).[26] Here, we have extended RosettaLigand to RosettaLigandEnsemble (RLE), an algorithm that can identify a binding mode favorable to a superimposed ensemble of congeneric ligands. This allows users to simultaneously dock a series of ligands in unison instead of individually as single ligands. We hypothesize that this will increase the efficiency and accuracy of sampling. We illustrate the hypothesized sampling advantage of RLE in Figure 1. Due to the presence of functional groups of varying sizes found within a SAR series, there may be binding modes available to certain molecules, but not others. RLE is capable of eliminating binding orientations not available to the ensemble as a whole. Furthermore, highly similar ligands are expected to bind in a similar fashion with common interactions to the chemical core.[5,27] The RLE scoring function emphasizes favorable positioning for the common scaffold, shown by the red outline. The greater the number of molecules that share a common substructure, the greater the scoring emphasis on that particular substructure. It is not anticipated that RLE will significantly improve docking for congeneric ligands that exhibit significantly different binding modes. Malhotra et al. reviewed receptor and ligand characteristics that tend to exhibit these alternate binding modes.[28]

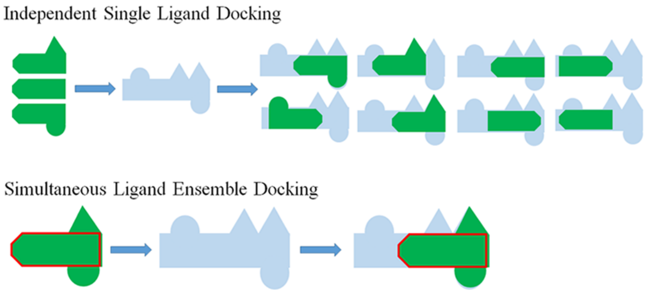

Independent Single Ligand Docking
Simultaneous Ligand Ensemble Docking

Figure 1. Hypothesized mechanism of the sampling advantage of RLE. (Top) Three small molecules (green) are independently docked by RosettaLigand into the protein-binding pocket (blue). Multiple docked orientations are possible for each small molecule. (Bottom) The same three molecules are first aligned using their common scaffold (red). Docking in concert using RLE then yields a single, unambiguous binding orientation.

EXPERIMENTAL METHOD

Benchmark Data set. A data set of 109 protein−ligand complexes across 20 systems (Supporting Information (SI) Table S1) is curated from the combination of the Community Structure−Activity Relationship (CSAR),[29] BindingDB Protein−Ligand Validation Sets,[4] PDBBind,[30] D3R docking resource, and individual crystallographic studies.[31−33] Each data set consisted of at least four chemically related ligands with experimental data and X-ray crystallography determined structures against a common protein target. A single receptor structure was selected from each data set as the primary docking target on the basis of crystallographic resolution, density in the ligand-binding pocket, and experimental affinity/activity. To test the potential of an ensemble docking approach, the data set favors cases wherein congeneric ligands bind in a similar fashion and an improvement using RLE docking is expected. Figure S3 shows the distribution of congeneric ligand root-mean-square deviations (RMSDs) and common scaffold sizes observed in the data set. The selected protein receptor structure is energy minimized using the Rosetta FastRelax protocol with a knowledge-based all-atom energy function.[34] The details of the Rosetta energy function have been covered extensively by Alford et al.[35] This minimization is performed in the apo state to remove the bias of side-chain positioning for the co-crystallized ligand. All other molecules in the series are

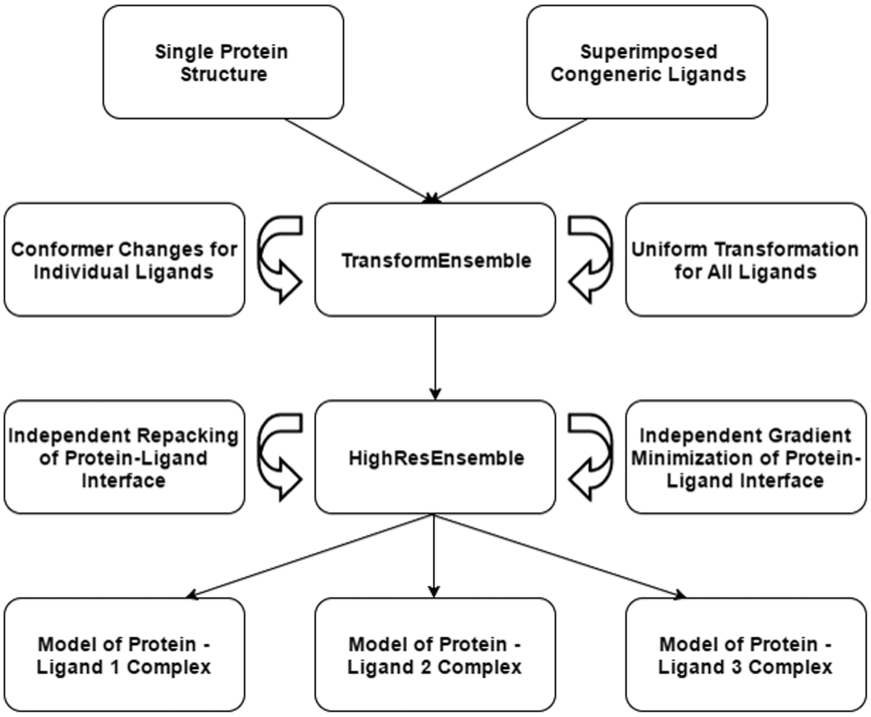

Single Protein Structure
Superimposed Congeneric Ligands
Conformer Changes for Individual Ligands
TransformEnsemble
Uniform Transformation for All Ligands
Independent Repacking of Protein-Ligand Interface
HighResEnsemble
Independent Gradient Minimization of Protein-Ligand Interface
Model of Protein - Ligand 1 Complex
Model of Protein - Ligand 2 Complex
Model of Protein - Ligand 3 Complex

Figure 2. Illustration of the RLE algorithm. The algorithm is divided into the low-resolution TransformEnsemble step and the high-resolution HighResEnsemble step. Curved arrows represent repeated moves accepted or rejected based on the Metropolis Monte Carlo criterion. Individual models of each protein−ligand pair are outputted from a single protein structure and superimposed congeneric ligands as input.

cross-docked to the energy minimized target using either traditional RosettaLigand docking or simultaneous RLE docking. Ligand conformations are generated using the in-house BioChemicalLibrary fragment-based conformer sampling methodology.[22] The co-crystallized ligand is excluded from docking with RosettaLigand or RLE to avoid self-docking bias, leaving a total of 89 test cases across all systems.

RLE Algorithm. Figure 2 illustrates the two-stage RLE algorithm. RLE takes as input a single protein structure and a congeneric series of molecules superimposed by a chemical scaffold. In the low-resolution TransformEnsemble phase, the same three-dimensional translations and rotations are applied to all molecules to maintain the superposition and find a common binding mode. Step sizes and direction for both translation and rotation are taken from a Gaussian distribution centered on a user-provided value. Scoring is done using a pregenerated shape complementarity energy grid and moves are accepted/rejected by a metropolis Monte Carlo criterion based on the sum of scores for all ligands in the ensemble. The protein structure remains static, but ligand conformers are changed by swapping out individual ligands with alternate conformations from pregenerated libraries. The benchmark used the fragment-based BCL::Conf small-molecule conformer generator.[22] During the high-resolution HighResEnsemble phase, only small perturbations to the ligand are applied, with the focus on optimizing the protein−ligand interface. As side-chain orientation differences are observed even for binding of related ligands, each protein−ligand interface is optimized independently. In a single simulation run, RLE generates x models, where x is the number of ligands in the ensemble. Over the course of n simulation runs, RLE generates $n \times x$ total models, the same quantity as x independent RosettaLigand runs of n trajectories each.

The bulk of the computation time in both RosettaLigand and RLE is due to protein side-chain rotamer sampling during the high-resolution docking phase. As RLE generates individual protein−ligand models for the high-resolution stage, the computation time is not significantly altered.

Experimental Model Generation. Initial parameters for RLE are derived from the latest features of the RosettaLigand algorithm[36,37] and optimized for sampling efficiency. Additional sampling cycles and a decreased rotational barrier were necessary to counteract the increased sampling space involved in finding an optimal position for all molecules simultaneously. The exact number of sampling steps was calculated on-the-fly based on the difference between the current step score and the maximum possible score, assuming that all atoms formed favorable interactions. Meanwhile, the repulsive score term was halved to allow the entire ensemble to rotate through clashes. Ligand atoms are forbidden from moving outside of the defined docking sphere as was the case in RosettaLigand.

Following optimization, docking was performed with both RosettaLigand and RLE, and evaluated for native ligand pose recovery. For each system, individual molecules were docked independently and as an ensemble into the same receptor structure. For each run, 2500 models were produced and the top 10% were selected based on ligand interface energy for subsequent analysis.

To make the docking simulation resemble actual use, a uniform volume random translation within a 5 Å sphere and a random full rotational orientation are performed before docking. A random conformer is selected from the ligand conformer library. This avoids biasing the starting position and orientation to that observed in the crystallographic complex. An example of how to generate models for one system is provided in the Supporting Information.

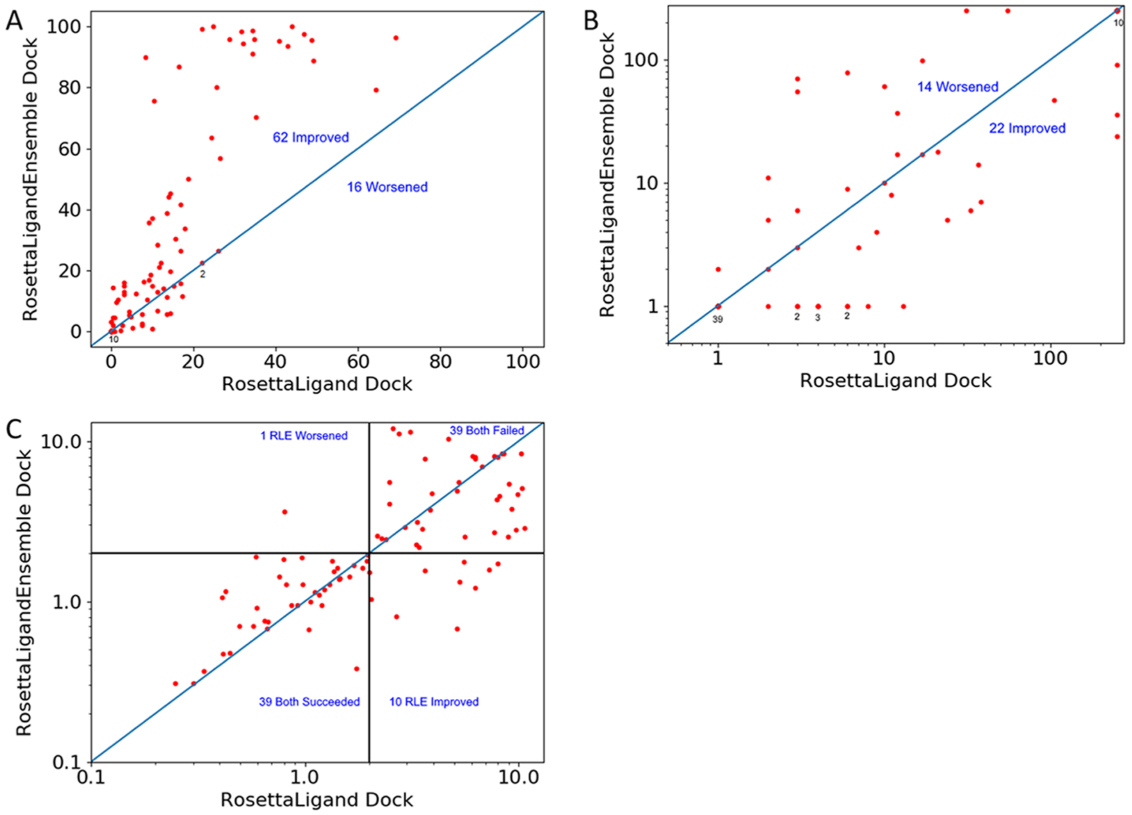


Figure 3. Comparison of sampling efficiency and scoring discrimination among the top 10% of models by score from individual RosettaLigand docking versus ensemble RLE docking. Overlapping dots are indicated by the number of overlapped points below it. The blue diagonal line shows when RosettaLigand and RLE performance are identical. (A) Percentage of nativelike models from single and ensemble docking. (B) Scoring rank of the best-scored nativelike model from single and ensemble docking. (C) Small-molecule RMSD of the top-ranked model from single and ensemble docking. The 2.0 Å success cutoff is marked out in black lines.

■ RESULTS AND DISCUSSION

We examine the top 10% of scoring models by the ligand interface score for each ligand cross-docking case. The top 250 models are analyzed for both sampling efficiency and scoring discrimination of nativelike models. Nativelike models are defined as having a ligand root-mean-square deviation (RMSD) of less than 2 Å compared with the co-crystal structure. Sampling efficiency is represented as the percentage of models that are nativelike, whereas scoring discrimination is represented as the scoring rank of the first nativelike model. A higher sampling percentage of nativelike models and a lower scoring rank for the best-scored nativelike model indicate an improvement.

RLE Improves Sampling and Scoring among the Top Models. Among the top 10% of models by score, RLE improved both the percentage of nativelike models and the scoring rank of the first nativelike model compared with RosettaLigand. Increased sampling efficiency was observed in 62 out of 89 cases, whereas an improved scoring rank was observed in 22 out of 89 cases, as shown in Figure 3. In three cases, RLE produced a nativelike model, whereas RosettaLigand did not. In 10 cases, neither RLE nor RosettaLigand could find a nativelike model in the top 10%.

Among the cases where RLE improved sampling efficiency, nearly half showed an improvement of at least 25%. In contrast, in no case did RLE decrease sampling efficiency by more than 9%. For scoring discrimination, RLE recovered a nativelike top-scoring model in 10 cases, where RosettaLigand failed to do so. This is important as RLE would have still produced an accurate model in an application scenario even for these cases. There is a single case where only RosettaLigand produced a nativelike top-scoring model. Here, RLE still produced a nativelike model in the top 10 scoring.

Although the increase in sampling efficiency was significant, there does not appear to be a direct translation between the number of nativelike models and the ability to discriminate them from non-native-like models. As both algorithms utilize the same knowledge-based scoring function during the high-resolution docking and the final ranking, it is expected that they may have a similar model discrimination power. This is illustrated in Figure 3c, where, in the large majority of cases, RLE and RosettaLigand either both succeeded or both failed at ranking a nativelike model as the best scoring. However, there are 10 cases where RLE was able to rescue the performance of RosettaLigand by producing a nativelike best-scoring model. Averaged across all 89 cases, the sampling efficiency improved by 18%, and, in 20 of these cases, both sampling and scoring metrics improved.

RLE Eliminates Alternate Binding Modes. The final binding location and orientation of the ligand is primarily determined by the low-resolution docking stage. Perturbations of the ligand in the high-resolution stage are minimal as the bulk of the computational time is spent toward conformational energy minimization of protein side chains. The RLE low-resolution phase moves all molecules in unison, maintaining superimposition, and therefore forces molecules to adopt a common binding mode. This coordinated movement is the process that eliminates binding volume available to some, but

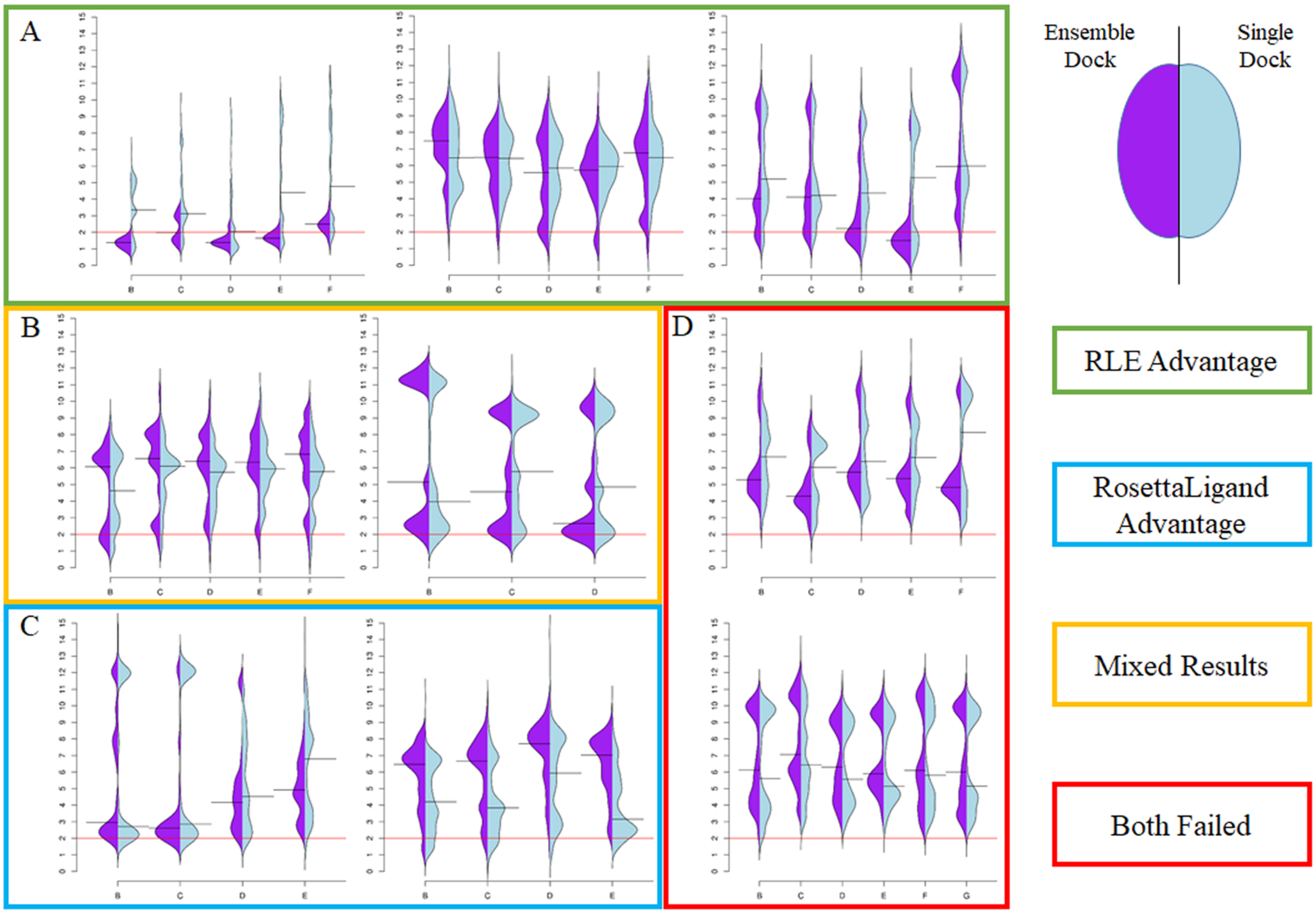


Figure 4. Ligand RMSD distribution observed among the top 10% of the models for RosettaLigand and RLE. Nine example systems have been divided into four qualitative categories of sampling and scoring change. For each system, data are split by individual protein–ligand pairs, with RLE docking on the left and RosettaLigand on the right. The black line shows the median and the red line shows the 2 Å cutoff. (A) RLE improved docking sampling and scoring for CTAP, HCV, and TPPHO (left, mid, right). (B) The improvement in RLE varied from ligand to ligand within the system for CDK2 and P38 (left, right). (C) RosettaLigand performed better for LPXC and THROM (left, right). (D) Both methods failed to perform well for CATB and THERM (top, bottom).

not all, members of the group. Figure 4 shows the ligand RMSD distributions observed among the top 10% of scoring models for both RLE and RosettaLigand docking. Each protein–ligand pair of the system is plotted separately so that effects across the system can be observed. Higher density at the low RMSD end of the distribution indicates success. The red line in each subplot shows the 2 Å ligand RMSD cutoff for nativelike binding modes. The systems have been sorted qualitatively into broad categories based on whether or not RLE generally improved both the sampling efficiency and the scoring discrimination. The RMSD distribution pattern for RLE is much more consistent within a system than the RosettaLigand distribution patterns for the same system. This is the aforementioned "forced common binding mode" effect. However, there remain individual protein–ligand pairs within a system where the distribution was not significantly improved.

In the systems where RLE drives both a sampling efficiency and a scoring discrimination improvement (green), RLE eliminated a significant number of high RMSD-binding modes observed in the RosettaLigand results. In the CTAP example, RLE ligand RMSDs are all within a similar range, whereas the outliers produced by RosettaLigand are eliminated. It remains possible for ensemble docking to be more successful for certain ligands within a group than others. Ligand C for CTAP has a smaller second peak that is not consistently eliminated by ensemble docking. One reason for this is that the high-resolution stage considers ligand conformers in addition to protein conformers. For larger, more flexible molecules, RMSD may be relatively high, even if the correct binding location and orientation is recovered. This is due to ligand conformational flexibility in the distal regions. Alternatively, in the HCV example, the majority of models from both RosettaLigand and RLE are not nativelike, but only RLE generates a batch of nativelike models. This is the aforementioned "rescue" scenario in which RLE is able to produce a correct model when RosettaLigand cannot.

A limitation to the RLE algorithm occurs when the alternate, high RMSD-binding mode is available to all molecules within a system, as seen in the P38 system with mixed results (orange). RLE does not provide a significant advantage in scoring discrimination when both methods have a similar sampling efficiency. The emphasis on the placement of the common scaffold means that an incorrectly identified common binding mode will result in poor performance across the system, as seen in THROM (blue). RosettaLigand was able to produce good results for two members of this system because its docking runs are independent. Whether or not the different binding modes will be correctly consolidated in an actual application depends

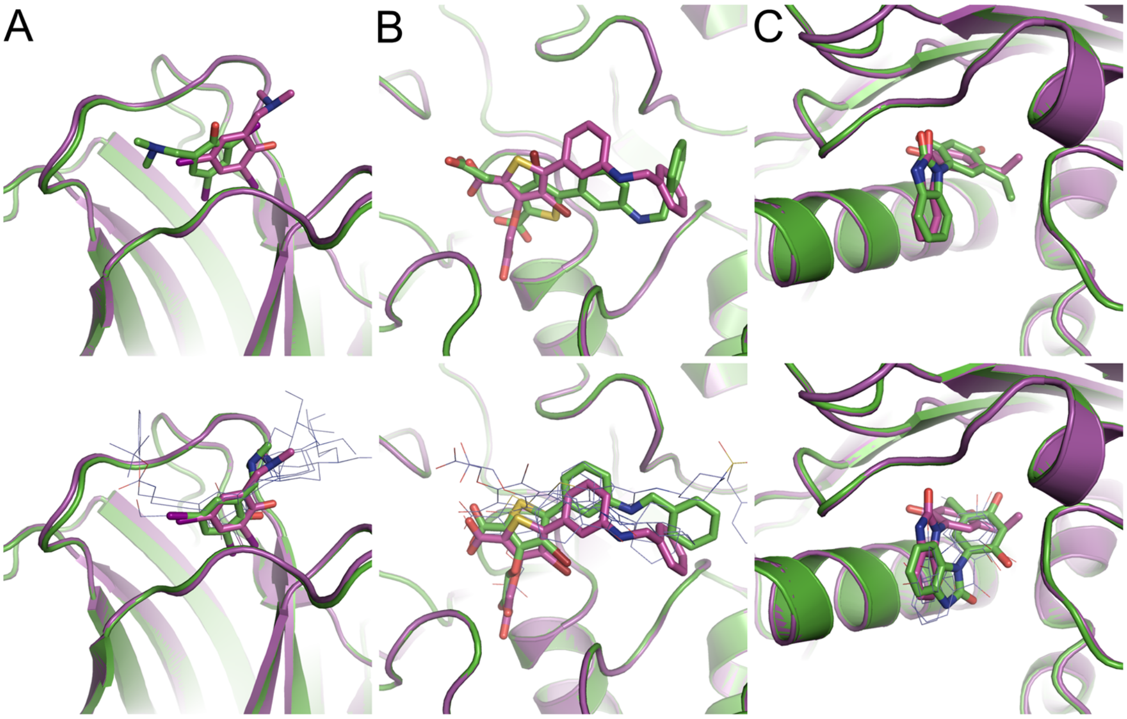


Figure 5. Illustrative examples of success and failure in recovering a nativelike best-scoring model. The top panels show the best-scoring model from RosettaLigand and the bottom panels show the best-scoring model from RLE. The co-crystal structure is shown (purple) aligned with the model (green). The remaining ligands of the RLE ensemble are shown as blue lines. (A) CTAP system, ligand ID B (PDB: 4AGL); (B) TPPHO system, ligand ID C (PDB: 2QBR); (C) HSP90 system, ligand ID B (4YKQ).

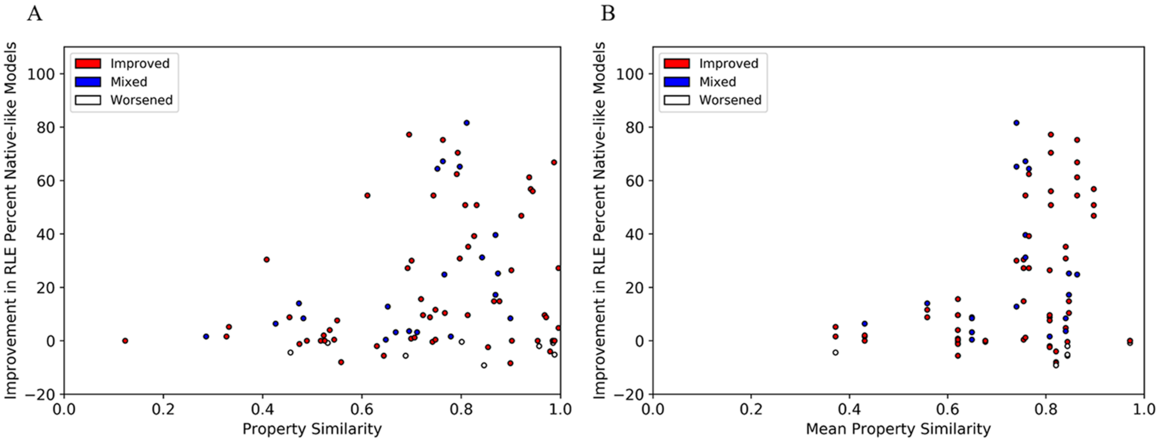


Figure 6. Sampling efficiency vs PropertySimilarity for the top 10% scoring models. Protein−Ligand pairs are divided into cases where RLE improved both sampling and scoring (red), worsened both sampling and scoring (white), or improved one, but not the other (blue). (A) PropertySimilarity of each test ligand to the ligand co-crystallized with a receptor structure vs the improvement in RLE docking, as calculated by RLE percent nativelike minus RosettaLigand percent nativelike. (B) The mean PropertySimilarity for each protein system vs the improvement in RLE docking. Each vertical set of dots represents a single protein−ligand system.

on the particular post-hoc analysis chosen. This incorrect placement of the common scaffold is repeated in the CATB and THERM systems, where the distribution peaks fall out of the nativelike RMSD range (red). One reason for this is the lack of chemical diversity in the functional group modifications within the group. These systems are difficult cases that neither algorithm can dock well.

Illustrative Examples of Success and Failure. The binding pockets for several illustrative examples are shown in Figure 5. Both CTAP ligand B and TPPHO ligand C show a significant improvement in sampling and scoring. The best-scoring RLE model is nativelike in both cases and sampling efficiency was 2.1× and 3.3× better for TPPHO and CTAP, respectively.

In the CTAP example, ligand B is a small ligand in a relatively open binding pocket. This made it difficult for RosettaLigand to determine the proper orientation, generating three equal possibilities, as shown in the RMSD distribution in Figure 4. However, the remaining ligands built off of the core scaffold have large chemical modifications off of two sites. The interactions formed by the distal groups with the bordering protein loops allow RLE to identify the proper orientation of the common core. Another example of this orientation flip is illustrated in the TPPHO example. Although the RMSD

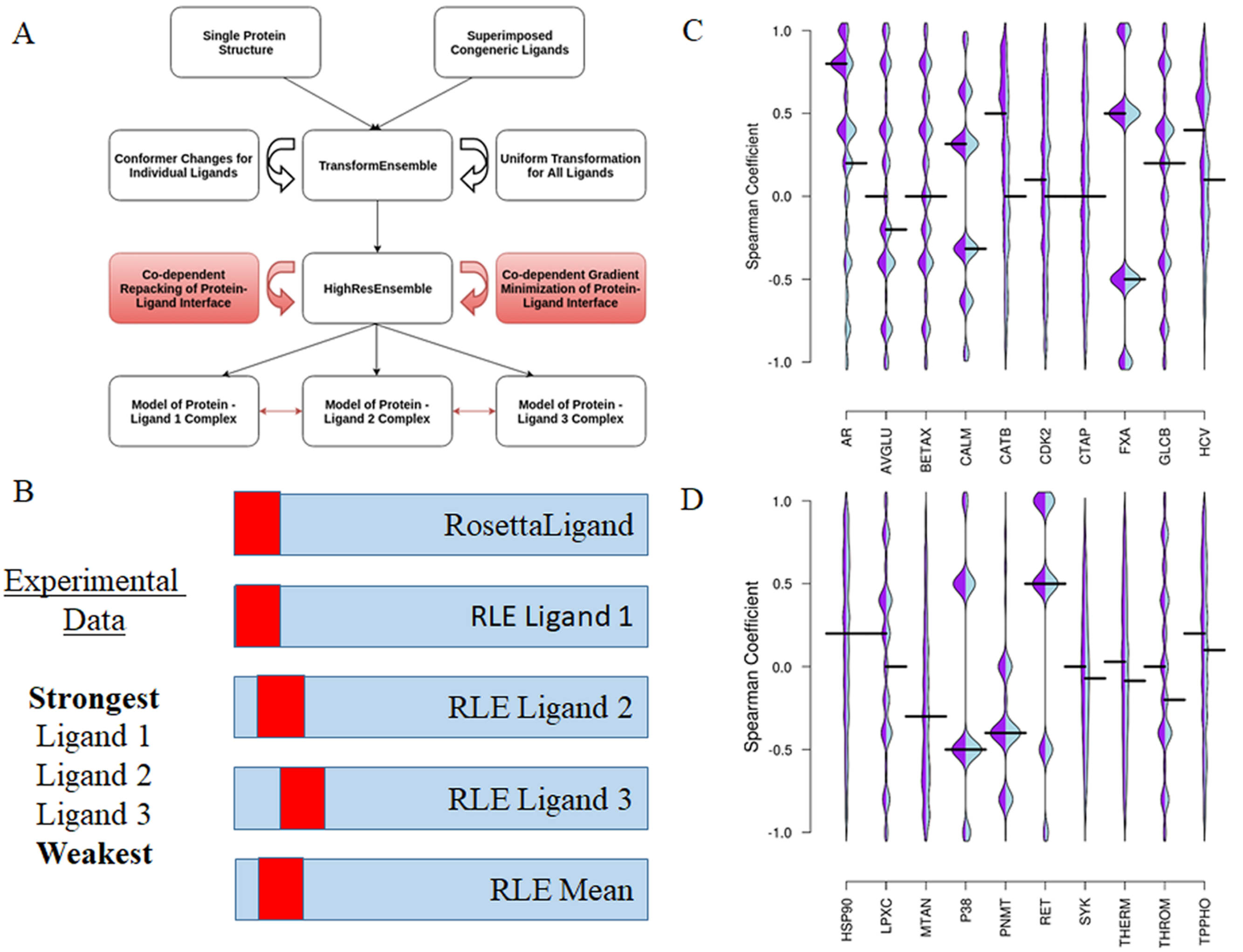


Figure 7. RLE Spearman correction during high-resolution docking to favor binding modes that correlate with experimental data. (A) Spearman-corrected RLE high-resolution docking steps are shown in red; (B) model selection dilemma resulting from inaccuracies in the docking scoring function; (C, D) distribution of Spearman correlation in generated ensembles for each system with a corrective factor (left, purple) and without a corrective factor (right, blue). The mean line is shown in black.

distribution for TPPHO ligand C is more distributed, a clear binding mode is available for ligands D and E in the TPPHO system. This confers RLE a modest increase in successfully orienting TPPHO ligand C, with the major deviation due to conformation rather than orientation.

Although RLE showed a slight sampling improvement for HSP90, the system proved to be a relatively challenging case for ensemble docking due to the small size of the ligand system. RosettaLigand produced a nativelike best-scoring pose, whereas RLE generated a flipped conformation. The RMSD distribution, however, favors RLE with more nativelike models. Across the system, there is a persistent alternate binding mode suggested by RLE (SI Figure S1) due to the fact that the binding pocket is much larger than the ligand. RLE is unable to rule out alternative binding modes of the common scaffold without distal groups that can eliminate conformational space.

Higher Chemical Similarity Promotes Higher Sampling Efficiency up to a Limit. To better understand indicators of successful and unsuccessful systems, we sought to characterize the similarity of the cross-dock small molecules compared with the co-crystallized molecule. The traditional Tanimoto similarity coefficient is not particularly robust for more complex substitutions as it focuses on atom identity within a common substructure. We compared molecules using the in-house BioChemicalLibrary to calculate a PropertySimilarity.[38] PropertySimilarity measures similarity based on atomic charges, van der Waals volume, bond types, and the presence of hydrogen bond donors/acceptors. For this data set, PropertySimilarity has a general positive correlation with Tanimoto similarity (SI Figure S2).

The relationship between sampling and scoring improvement to property similarity is shown in Figure 6. Ligands are classified based on whether both sampling and scoring improved (red), both worsened (blue), or a combination of the two (white). There is a general tendency for molecules docked with high sampling efficiency to have a high chemical similarity to the co-crystallized molecule. However, there are a number of highly related molecules in Figure 6 that are poorly sampled, suggesting that chemical similarity is a necessary, but not sufficient, condition for docking success. This is in agreement with previous studies that have shown that docking success increases with chemical similarity.[24,25]

To predict performance on a system level, we computed the mean PropertySimilarity for each system and plotted this value against improvement in sampling efficiency in each ligand when docked with RLE. This is shown in Figure 6, with each vertical line comprising of a congeneric set of protein−ligand pairs. The largest improvement occurs around a mean PropertySimilarity measure of 0.8, suggesting that there is a "sweet spot" for improvement. Systems that are too different (P38, mean = 0.37) or too similar (THERM, mean = 0.97) exhibit limited benefits from ensemble docking. In particular, the THERM system consists of a chemical scaffold to which the primary modification is the switching of various hydrocarbon groups. Furthermore, the molecule interacted with two separate hydrophobic ends and a network of water molecules in the binding account, which makes the determination of orientation difficult.[31]

Identifying Favorable Binding Poses Corresponding with SAR Data. The inaccuracy of ranking despite accurate docking remains problematic. One post-hoc solution is to select sets of binding modes that correlated with experimental data. We sought to address this deficiency during docking by adding a corrective factor to drive high-resolution docking toward binding modes. Following each cycle of optimization, we modified the scoring difference based on the Spearman correlation to experimental data as an adjustment before applying the Metropolis criterion. This correction is shown in eq 1. ρ_B and ρ_A refer to spearman rank correlation before and after the current step, respectively. The weight is a user-provided value and the adjusted score is used to evaluate whether the Monte Carlo step is accepted or rejected.

$$\text{adjusted score} = \text{score} - (\rho_A - \rho_B) \times \text{weight} \tag{1}$$

As smaller Rosetta scores are considered more favorable, eq 1 is written with the default assumption that smaller experimental values are preferred. The adjustment provides an additional bonus to perturbations that improved the score of stronger binding or more active ligands, and to perturbations that worsened the score of weaker binding or less active ligands. The low-resolution docking stage remains the same and does not account for experimental correlations. The adjusted co-dependent algorithm is shown in Figure 7A, with the score adjustment being applied in the highlighted step.

Although the corrective factor does improve correlation with experimental affinity, it does not improve the sampling efficiency or docking accuracy. This is in part due to the fact that the binding orientation is primarily determined in the low-resolution phase that does not account for correlation. The Spearman correlation coefficient is defined as the Pearson correlation based on only the ranks of the models. Therefore, the Spearman correlation has a discrete distribution, with limited values available for a small data set. This makes it difficult to significantly improve the correlation in many cases. The results are in agreement with previous results showing that improvements in the Pearson scoring correlation using machine learning-based scoring functions only translated to a moderate increase in accurate ranking.[39]

One additional hindrance to a more successful corrective method is the dilemma in selecting the models illustrated in Figure 7B. To maintain the experimental correlation, entire ensembles of ligand models must be selected. However, as the Monte Carlo sampling method is stochastic, it is unlikely that each ensemble will contain low-energy conformations of every protein−ligand interface. Selecting models by a mean metric across the entire ensemble may select the best-scoring models for one ligand, but not for others. Even with improvements in scoring functions, this selection dilemma may prevent RLE from simultaneously selecting the best models for each ligand.

Comparing RosettaLigandEnsemble with Other Protein−Ligand Docking Tools. We used AutoDock[9] with a Lamarckian Genetic Algorithm to test its performance in the 89 cross-docking systems in the benchmark set. Standard protocol settings and a docking volume comparable to RLE docking were used to generate the models. The AutoDock simulations were performed using a rigid receptor model.

Figure 8 shows the small-molecule RMSD of the top-scoring model from RLE and AutoDock docking. RLE recovered a nativelike small molecule pose in 23 cases, where AutoDock did not. By contrast, there were only 13 cases where AutoDock had a nativelike best-scoring model when RLE did not.

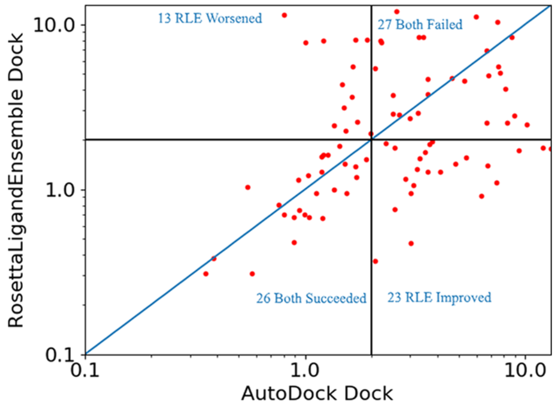


Figure 8. Small-molecule RMSD of the top-ranked model from RLE and AutoDock docking. The 2.0 Å success cutoff is marked out in black lines. Equal performance of the two software is indicated by the blue line.

Wang et al. evaluated 10 docking software across the PDBBind data set, including 18 cross-docking cases from the present benchmark, in which at least one tested method did not recover a nativelike top-scoring model.[40] RLE rescued the performance in 15 out of 18 of these cases (SI Table S2). Most notably, RLE was able to generate nativelike models across the calcium-dependent protein kinase CDPK1 (CALM) and *Helicobacter pylori* nucleosidase (MTAN) systems. However, the 2B07 test case from a series of protein tyrosine phosphatase inhibitors (TPPHO) remains challenging. This is likely related to the orientation flip discussed with respect to Figure 5. RLE was also able to recover nativelike top-scoring models for all five spleen tyrosine kinase compounds also tested in CSAR 2014, matching the performance of the best available docking tools.[41] However, RLE performed worse on an deacetylase (LPXC) test system that was part of CSAR 2012,[13] only recovering a near-native model in one out of four cases. The generated best-scoring models had an RMSD of 2.23 Å in the worst test case, suggesting only a minor decrease in performance in the LPXC system. It should be noted that these comparisons do not account for additional protein flexibilities accounted for by RLE, nor does it include the effects of differences in the starting ligand conformation. However, there does not appear to be a strong induced fit or conformational selection component in these structures.

■ CONCLUSIONS AND FUTURE DIRECTIONS

Needed Improvements in Decoy Discrimination. The improved sampling efficiency did not directly translate into improved scoring ranking partly due to the inaccuracies in discriminating between nativelike and non-native-like models.

Better decoy discrimination in conjunction with the more efficient sampling will allow for fewer models to be produced before converging on a nativelike binding mode. The reduced number of models will greatly reduce the time and computational resources necessary for docking. Furthermore, the SAR-correlated docking would benefit greatly from a more accurate scoring function capable of ranking ligands. RLE in combination with such a method would generate binding modes in accordance with SAR data without the need for post-hoc filtering.

Consideration of Alternate Binding Modes among Congeneric Ligands. RLE docking is generally designed for docking in cases where similar ligands exhibit a common binding mode. This is the case for the vast majority of known protein−ligand crystallographic complexes.[5,28] Presently, a priori assumptions are made for a given system, even if single-ligand docking is used, as initial placement is often based on previously seen binding modes. A future development of RLE docking would allow for minor shifts in the binding mode, while maintaining the general placement and orientation, a sort of soft ensemble docking. Furthermore, the use of a property-based alignment method such as PropertySimilarity will allow for common scaffolds based on chemical similarity as opposed to identity. Cases wherein similar ligands bind in completely different pockets or to different protein conformations will remain challenging for ensemble-based methods.

Ensemble Approaches from the Protein Structure-Based Direction. A similar approach can be used to drive ensemble docking improvements in the use of protein mutation data. Current approaches to protein ensembles generally focus on accounting for conformational diversity. Mutational data on proteins are used to identify potential protein−ligand interaction sites as a distance restraint to docking. An alternate ensemble approach would utilize SARs based on multiple protein mutants to determine how the ligands may bind to each mutant within the series. A further step would be in combining protein ensemble and ligand ensemble methods to improve docking accuracy by considering how ligand modifications fit into the different pockets of protein mutants. Multi-target virtual screening, in particular, with biologically relevant mutants, can be performed with such an algorithm. The Platinum database of small-molecule interactions with protein mutants[42] provides an excellent source of data for training an algorithm in this approach.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b02059.

> Protein−Ligand systems used to benchmark RosettaLigandEnsemble (Table S1); RosettaLigandEnsemble performance against previous benchmark (Table S2); HSP90 ligand RMSD distribution, Tanimoto similarity versus property similarity; RMSD of common scaffold in dataset (Figures S1−S3); a protocol capture for RosettaLigandEnsemble in Omega_Supplemental.pdf (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jens.meiler@vanderbilt.edu.

ORCID

Darwin Yu Fu: 0000-0003-1407-1689

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by a Predoctoral Fellowship from the PhRMA foundation. Work in the Meiler laboratory was supported through NIH (R01 GM099842, R01 GM073151) and NSF (CHE 1305874). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

■ REFERENCES

(1) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(2) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.

(3) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(4) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

(5) Fu, D. Y.; Meiler, J. Predictive Power of Different Types of Experimental Restraints in Small Molecule Docking: A Review. *J. Chem. Inf. Model.* **2018**, 225−233.

(6) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins* **2006**, *65*, 538−548.

(7) Davis, I. W.; Baker, D. RosettaLigand Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* **2009**, *385*, 381−392.

(8) Bender, B. J.; Cisneros, A.; Duran, A. M.; Finn, J. A.; Fu, D.; Lokits, A. D.; Mueller, B. K.; Sangha, A. K.; Sauer, M. F.; Sevy, A. M.; Sliwoski, G.; Sheehan, J. H.; DiMaio, F.; Meiler, J.; Moretti, R. Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry* **2016**, *55*, 4748−4763.

(9) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

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

(11) 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.

(12) Li, Y.; Han, L.; Liu, Z.; 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.

(13) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B.; 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.

(14) Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K. D3R Grand Challenge 2015: Evaluation of Protein-ligand Pose and Affinity Predictions. *J. Comput. Aided. Mol. Des.* **2016**, 651−668.

(15) Wang, J.; Verkhivker, G. M. Energy Landscape Theory, Funnels, Specificity, and Optimal Criterion of Biomolecular Binding. *Phys. Rev. Lett.* **2003**, *90*, No. 188101.

(16) Amaro, R. E.; Baron, R.; McCammon, J. A. An Improved Relaxed Complex Scheme for Receptor Flexibility in Computer-Aided Drug Design. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 693−705.

(17) 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.

(18) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated Docking to Multiple Target Structures: Incorporation of Protein Mobility and Structural Water Heterogeneity in AutoDock. *Proteins* **2002**, *46*, 34−40.

(19) Huang, S. Y.; Zou, X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins: Struct., Funct., Genet.* **2007**, *66*, 399−421.

(20) Feixas, F.; Lindert, S.; Sinko, W.; McCammon, J. A. Exploring the Role of Receptor Flexibility in Structure-Based Drug Discovery. *Biophys. Chem.* **2014**, *186*, 31−45.

(21) Sinko, W.; Lindert, S.; McCammon, J. A. Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem. Biol. Drug Des.* **2013**, *81*, 41−49.

(22) Kothiwale, S.; Mendenhall, J. L.; Meiler, J. BCL::Conf: Small Molecule Conformational Sampling Using a Knowledge Based Rotamer Library. *J. Cheminform.* **2015**, *7*, 47.

(23) Huang, S.-Y.; Li, M.; Wang, J.; Pan, Y. HybridDock: A Hybrid Protein-Ligand Docking Protocol Integrating Protein- and Ligand-Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1078−1087.

(24) 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.

(25) 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.

(26) Lemmon, G.; Meiler, J. Towards Ligand Docking Including Explicit Interface Water Molecules. *PLoS One* **2013**, *8*, No. e67536.

(27) Boström, J.; Hogner, A.; Schmitt, S. Do Structurally Similar Ligands Bind in a Similar Fashion? *J. Med. Chem.* **2006**, *49*, 6716−6725.

(28) Malhotra, S.; Karanicolas, J. When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? *J. Med. Chem.* **2017**, *60*, 128−145.

(29) Dunbar, J. B., Jr.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y.-N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A. CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *J. Chem. Inf. Model.* **2013**, *50*, 1842−1852.

(30) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(31) Krimmer, S. G.; Betz, M.; Heine, A.; Klebe, G. Methyl, Ethyl, Propyl, Butyl: Futile but Not for Water, as the Correlation of Structure and Thermodynamic Signature Shows in a Congeneric Series of Thermolysin Inhibitors. *ChemMedChem* **2014**, *9*, 833−846.

(32) Baum, B.; Muley, L.; Smolinski, M.; Heine, A.; Hangauer, D.; Klebe, G. Non-Additivity of Functional Group Contributions in Protein-Ligand Binding: A Comprehensive Study by Crystallography and Isothermal Titration Calorimetry. *J. Mol. Biol.* **2010**, *397*, 1042−1054.

(33) Krieger, I. V.; Freundlich, J. S.; Gawandi, V. B.; Roberts, J. P.; Gawandi, V. B.; Sun, Q.; Owen, J. L.; Fraile, M. T.; Huss, S. I.; Lavandera, J. L.; Ioerger, T. R.; Sacchettini, J. C. Structure-Guided Discovery of Phenyl-Diketo Acids as Potent Inhibitors of M. Tuberculosis Malate Synthase. *Chem. Biol.* **2012**, *19*, 1556−1567.

(34) Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D. Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling. *Protein Sci.* **2014**, *23*, 47−55.

(35) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031−3048.

(36) Lemmon, G.; Meiler, J. RosettaLigand Docking with Flexible XML Protocols. *Methods Mol. Biol.* **2012**, *819*, 143−155.

(37) DeLuca, S.; Khar, K.; Meiler, J. Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLoS One* **2015**, *10*, No. e0132508.

(38) Nguyen, E. D. Structural Studies of the Interaction between mGlu5 and Allosteric Modulators. Ph.D. Dissertation, Vanderbilt University, 2013.

(39) Ashtawy, H.; Mahapatra, N. Does Accurate Scoring of Ligands against Protein Targets Mean Accurate Ranking? *Bioinf. Res. Appl.* **2013**, 298−310.

(40) Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T. Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein-ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power. *Phys. Chem. Chem. Phys.* **2016**, *18*, 12964−12975.

(41) Carlson, H. A.; Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G.; Peishoff, C. E.; Lambert, M. H.; Dunbar, J. B. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* **2016**, *56*, 1063−1077.

(42) Pires, D. E. V.; Blundell, T. L.; Ascher, D. B. Platinum: A Database of Experimentally Measured Effects of Mutations on Structurally Defined Protein-Ligand Complexes. *Nucleic Acids Res.* **2015**, *43*, D387−D391.

DOI: 10.1021/acsomega.7b02059