Systematic comparison of Amber and Rosetta energy functions for protein structure evaluation.

Aliza B. Rubenstein[1,2], *Kristin Blacklock*[3,4], *Hai Nguyen*[3,4†], *David A. Case*[1,2,3,4]*, *Sagar D. Khare*[1,2,3,4]*

[1]Computational Biology & Molecular Biophysics Program

[2]Institute for Quantitative Biomedicine

[3]Department of Chemistry and Chemical Biology

[4]Center for Integrative Proteomics Research

Rutgers, The State University of New Jersey, Piscataway NJ 08854

KEYWORDS: Amber, ff14SBonlySC, Rosetta, talaris2014, REF2015, comparative analysis, structure prediction, benchmark, loop modeling

ABSTRACT An accurate energy function is an essential component of biomolecular structural modeling and design. The comparison of differently derived energy functions enables analysis of the strengths and weaknesses of each energy function, and provides independent benchmarks for evaluating improvements within a given energy function. We compared the molecular mechanics Amber empirical energy function to two versions of the Rosetta energy function (talaris2014 and REF2015) in decoy discrimination and loop modeling tests. Both Rosetta's talaris2014 and Amber's ff14SBonlySC energy functions performed well in scoring the native state as the lowest energy conformation in many cases. In 24/150 cases with Rosetta, and in 2/150 cases using Amber, a false minimum is found that is absent in the alternative landscape. In 21/150 cases, both energy function-generated landscapes featured false minima. The newest version of the Rosetta energy function, REF2015, which has more physically-derived terms than talaris2014, performs significantly better, highlighting the improvements made to the Rosetta scoring approach. To take advantage of the semi-orthogonal nature of these energy functions, we developed a technique that combines Amber and Rosetta conformation rankings to predict the most near-native model for a given protein. This algorithm improves upon predictions from either energy function in isolation, and should aid in model selection for structure evaluation and loop modeling tasks.

Introduction

Computational protein structure prediction is dependent on an accurate energy function. The native state of a protein is expected to be found uniquely at the minimum of the energy function[1]; therefore, the energy function must robustly discriminate between native and non-native conformations. A variety of energy functions to predict protein structure have been implemented over the past forty years[2-8]. These potentials largely fall into one of two categories: molecular

mechanics force fields that rely on the combination of various empirical potentials such as Lennard-Jones, torsional energies, Coulombic interactions, and desolvation penalties[3,4,7] and statistical or knowledge-based potentials that depend on characteristics of known protein structures[2,5,6]. While molecular mechanics force-fields are generally parameterized on small molecule properties[7,9–11], statistical potential parameter optimization is often guided by known biomolecular structures[12–14]. Each approach has its own drawback: since parameters in physically derived force-fields are fit based on small molecule properties, they may not be suited to macromolecules[15,16]: for example, force-fields will often display biases towards secondary structure propensities[15,17]. On the other hand, statistical potentials are trained on specific datasets of large biomolecules, and data sparseness may lead to overfitting[18].

The Rosetta macromolecular modeling program energy function combines elements of both categories; it contains physical force-field terms (Lennard-Jones interactions, electrostatic interactions, desolvation penalties, etc.) and statistical potentials (probability of amino acid identity given backbone angles, probability of backbone angles given amino acid identity, probability of backbone-dependent rotamer, etc.)[19]. The most recent Rosetta energy function (REF2015) is parameterized on both small molecule properties and large sets of biomolecular structures[18], although previous energy functions were generally parameterized on known biomolecular structures alone[13]. While efforts have been made to compare the performance of various empirical force-fields[17,20,21], little attention has been focused on the comparison between the Rosetta energy function and empirical force-fields.

The Amber ff14SBonlySC force field[10] uses a standard fixed-charge molecular mechanics potential, with torsion potentials based entirely on fits to quantum chemistry data. It is very like the more commonly-used ff14SB protein force field, but does not include the empirical

modifications to backbone torsion potentials that are present in ff14SB, and which provide an improved balance of secondary structure in explicit solvent simulations. Hence, ff14SBonlySC is more "physics-based" than is ff14SB, and it arguably better suited for the implicit solvent simulations used here, since the empirical backbone torsional potentials in ff14SB might be specific to its use of explicit solvent simulations. The ff14SBonlySC force field, in combination with a generalized Born implicit solvent model[22], has been shown to fold a variety of single-domain proteins using unrestrained molecular dynamics simulations[23].

Comparing the Amber force-field and Rosetta energy function performance at structure evaluation elucidates the strengths and areas of improvements for each energy function. As Rosetta energy functions have been developed based on improving performance for certain modeling datasets, testing their performance on the same macromolecular datasets may result in overfitting of the Rosetta energy function, while comparing their performance to that of a physics-based Amber energy function is a relatively unbiased comparison for evaluating performance improvements. Finally, selecting a correct near-native model for a given sequence is an elementary challenge; the combination of these two semi-orthogonal energy functions provides a method for model selection that is able to select more accurate models.

Methods

Benchmark Sets

To evaluate and compare the performance of Rosetta and Amber energy functions, we used two benchmark sets, a structure evaluation (decoy discrimination) set and a loop modeling set. The decoy discrimination benchmark set includes a total of 150 proteins, a combination of two independent decoy sets used in previous studies[18,24]. The proteins in the set are monomeric and have crystallographic native structures available in the RCSB PDB[25] with resolution < 2.0 Å. The

protein lengths range from 50 to 200 residues and have a diverse range of topologies. The decoy sets were originally generated using biased and unbiased ab-initio sampling runs[26] followed by parallel loophash sampling (PLS)[27]. This produced 40,000-200,000 decoys per protein, ~1000 representative low-energy structures of which were chosen for each protein to cover the range of possible C-α RMSD values.

The loop modeling benchmark set consisted of the 45-PDB dataset for 12-residue loops in the monomeric protein loops training set of the 2016 Collaborative Assessment and Development of Rosetta Energetics and Sampling (CADRES). This loop modeling benchmark set was obtained from Shane O'Connor and Tanja Kortemme (personal communication).

Structure Preparation

Rosetta

In the decoy discrimination benchmark, the native crystallographic structures for each protein set were downloaded from the RSCB PDB and residues were trimmed from the structure to match the sequence of the crystal with the decoy structure in the benchmark sets. Native structures were necessary to evaluate RMSD from native for decoy conformations. Native structures were then relaxed using FastRelax[26] with the talaris2014[13] scorefunction to relieve any clashes. One hundred relaxation trajectories were simulated to generate one hundred relaxed native-like decoys. These native-like decoys were used for false minima analysis. Then, these one hundred native-like decoys, along with the ~1000 pre-sampled decoys, were subjected to backbone and sidechain minimization using talaris2014 and the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) minimizer implementation with inexact line search conditions (lbfgs_armijo_nonmonotone) over a maximum of 2000 iterations for convergence. C-α atom RMSD was calculated for all decoys.

The REF2015 dataset was obtained from F. DiMaio and H. Park[18]. For this dataset, each decoy was relaxed with 3 cycles torsion-space minimization and 2 cycles Cartesian mode[24] using the REF2015 energy function[19]. Only 140 out of 150 protein systems were included in this set due to the lower quality of experimentally determined structures for 10 systems (H. Park and F. DiMaio, personal communication, July 5, 2017). Those 10 systems are ignored when comparing REF2015 to Amber.

In the loop modeling benchmark, the native crystal structures for each protein set were downloaded from the RCSB PDB and trimmed of excess residues that were not found in the decoy PDB structures. The backbone and sidechain geometries for residues in the loop region of each decoy structure were minimized in Rosetta using the talaris2014 scorefunction and the lbfgs_armijo_nonmonotone over a maximum of 2000 iterations for convergence. C-α RMSDs were calculated with respect to the crystal structure over loop residues only without fitting; since the protein scaffold was fixed during optimization, this statistic describes the extent of loop deviation. Loop residues are defined in supplementary file LoopDefs.xlsx.

Amber

Hydrogens were removed from the crystal structures and decoy PDBs, and initial structures were built using the tLEaP module of AmberTools[28] with the ff14SBonlySC[10] forcefield parameters. Minimizations were carried out for a maximum of 1000 steps under the LBFGS quasi-Newton algorithm[29] with a convergence criterion of 0.01 kcal/mol-A. In the loop modeling benchmark, positional restraints were added to all non-loop-residue atoms except for hydrogens with a force constant of 10.0 kcal mol-1 A-2. Solvent effects were treated with a generalized Born implicit solvent model (GB-Neck2[22]) implemented in the Amber16[28] package with mbondi3 radii and a cutoff value of 999A for nonbonded interactions. Total potential energies of

minimized structures and C-α RMSDs with respect to the crystal structure were obtained using the pytraj 2.0.0 interactive molecular dynamics simulation data analysis Python package[30], which is a Python interface for cpptraj in AmberTools16[28]. In the loop modeling benchmark, C-α RMSDs were calculated over the loop residues only. Six sets of decoy structures for the loop modeling benchmark were unable to be minimized in Amber due to missing residues, and those sets were not considered in subsequent analyses (1cb0, 1dts, 1m3s, 1ms9, 1t1d, and 2pia).

Energy Landscape Generation

Energy landscapes (RMSD vs. normalized energy scatterplots) were generated for all proteins for both Rosetta and Amber. The ideal shape of an energy landscape is that of a funnel (i.e. Figure 1A, turquoise plot) where the lowest-scoring decoy conformations are of near-native RMSD. We use the binned Boltzmann metric (see below) to evaluate the funnel shape of each energy landscape.

Energy Normalization

For each set of energies per scorefunction per protein, energies are normalized so that the gap between the 5th percentile of and the 95th percentile is equal to 1. This enables the comparison of energies between different structures and between different energy functions. This is accomplished via the following equation:

$$E_{i(norm)} = (E_i - E_{min})/(E_{95th} - E_{5th})$$

E_i refers to the raw energy of decoy i. E_{min} is the minimum energy value, E_{95th} is the 95th percentile energy, and E_{5th} is the 5th percentile energy

Funnel Evaluation Metric

We use the binned Boltzmann metric, *B*, for energy landscape evaluation, as described previously by Park et al.[18]. This metric finds the Boltzmann probability of selecting native-like

decoys over high-RMSD decoys based on their energy values. As in previous work[24], the metric is averaged over multiple thresholds for determining native-like status for each decoy.

$$B = \frac{\sum_j (\sum_i d_{ij} P_i / \sum_i P_i)}{N_j}$$

$$P_i = e^{-\beta E_{i(norm)}}$$

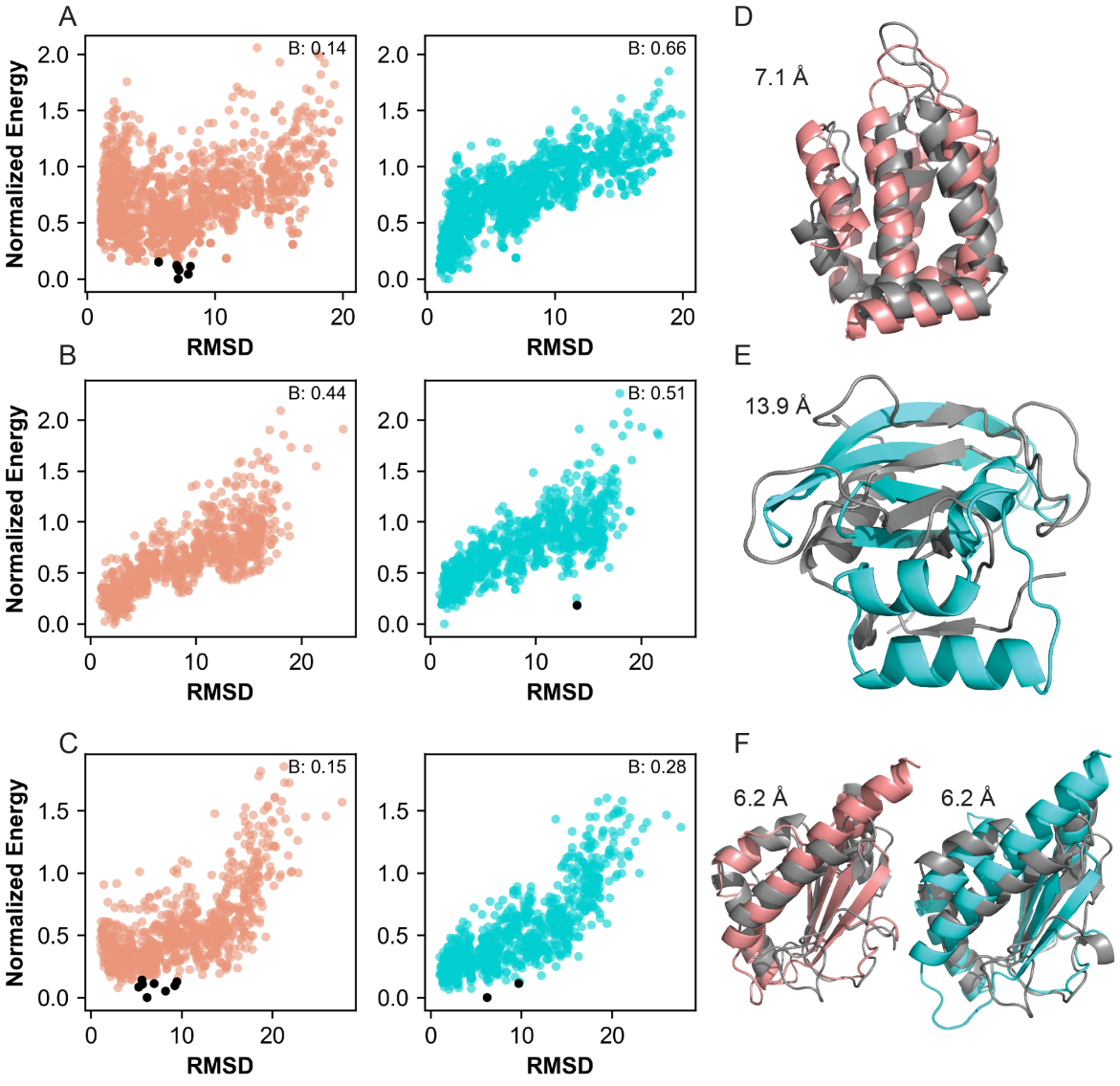


Figure 1. (A-C) Energy landscapes for 2QY7, 1T2I, and 1SEN respectively. Each dot on the plot represents one decoy conformation. The x-axis is RMSD from native and the y-axis is normalized energy. False minima (defined as decoys within top 10 energies but with RMSD > 5.0 Å) are depicted in black. The *B* metric, which represents the efficacy of the score-function at differentiating between native and non-native decoys, is shown at the top right corner of each plot. Rosetta plots are to the left, in salmon, and Amber plots are to the right, in turquoise. (D-F) Superimposed native (gray) and Rosetta lowest-ranking false minimum decoy (salmon) and/or superimposed native (gray) and Amber lowest-ranking false minimum decoy (turquoise) for 2QY7, 1T2I, and 1SEN respectively.

The conformation index is i and j is the native threshold definition index. Cutoffs are 0.5, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.5, 4.0, 5.0, 6.0 Å and N_j is thus 14. $E_{i(norm)}$ is the score of decoy i as determined in Rosetta or Amber and normalized as described above. The factor β refers to the Boltzmann factor and has a value of 0.1, as this was the value of β used by Park et al. previously[18]. d_{ij} determines whether decoy i is considered native at threshold j; it is set to 1 if it is native and 0 if it is not. As the sum of the probabilities of the non-native-like conformations approaches 0, the numerator ($\sum_i d_{ij} P_i$) approaches the value of the denominator ($\sum_i P_i$), so that the value of B approaches to 1. As mentioned in Park et al.[18], the B metric is better than the previously used S metric[24] due to a larger increase in the metric for a poor energy landscape vs. a good energy landscape than the increase from an already good energy landscape to a steeper energy landscape. Additionally, it is a smoother metric that is less affected by single-decoy outliers.

Model Selection

For each protein, a model was selected by finding the decoy that had the lowest sum of Amber and Rosetta ranks; this decoy also satisfies the criteria for Pareto-optimality. First, the Amber scores and Rosetta scores were converted into ranks so that the rank of decoy *a* was less than the rank of decoy *b* if the energy of decoy *a* was less than the energy of decoy *b*. Second, the Pareto-optimal solutions are found as follows. Decoy *a* is defined as dominating decoy *b* if both ranks (Rosetta and Amber) of decoy *a* are <= both ranks of decoy *b*. Pareto-optimal decoys are decoys that dominate at least one other decoy and are not dominated by any decoys. From among the set of Pareto-optimal decoys, the decoy that has the lowest sum of ranks is chosen as the solution. In the rare case that more than one decoy has a minimum sum of ranks, a decoy is arbitrarily chosen from the minimum-sum-ranks decoys. Technically, the minimum-sum decoy must be found within the Pareto set of solutions; however, the use of the Pareto-minimization allows for easier visualization and interpretation of the minimum-sum solution. Additionally, the Pareto-optimal set of decoys may be useful for selecting a set of top-scoring *n* (*n*>1) decoys when performing consensus scoring according to the two energy functions.

Results

Performance of Amber and Rosetta energy functions in discriminating between native and non-native structures

Protein free energy landscapes involve folding funnels[31–33] which enable the folding chain to efficiently find the native state[1,31], and their existence implies that the higher energy of non-native (decoy) structures compared to the native (e.g., crystallographically-determined) structure drives protein folding. Therefore, a common test used for evaluating[23,26] and improving[18] energy functions is the decoy discrimination test, in which the evaluated scores of decoy structures are compared to that of near-native structures. High-RMSD decoys which have comparable energies

to near-native structures are classified as “false minima”, and are indicative of inaccuracies in the energy function. The *B* metric[18], ranging from 0 to 1, quantifies the existence of false minima in a set of structures upon evaluation with a given energy function, with values close to 1 indicating a smooth folding funnel with no false minima. Conversely, a lower *B* value indicates that one or more false minima exist.

We compared the performance of the Amber energy function and Rosetta energy function at ranking native state structures lower than decoy conformations for a set of 150 proteins. Amber ff14SBonlySC generally performed better than Rosetta talaris2014, scoring significantly higher *B* metrics for many systems (Figure 2A). We also compared Amber to the newer default Rosetta energy function, REF2015[19], and found that while Amber did have a higher *B* metric for several systems, several other systems had a higher *B* metric when scored by REF2015, thus showing the improvement of REF2015 over talaris2014 when compared to Amber as an unbiased benchmark. Nonetheless, the comparative performance of the two energy functions (Amber and REF2015; Fig. 2B) shows that each has its strengths and limitations (Table 1). Our analysis was carried out with the talaris2014 energy function, and we refer to it as the Rosetta energy function in the remainder of this paper.

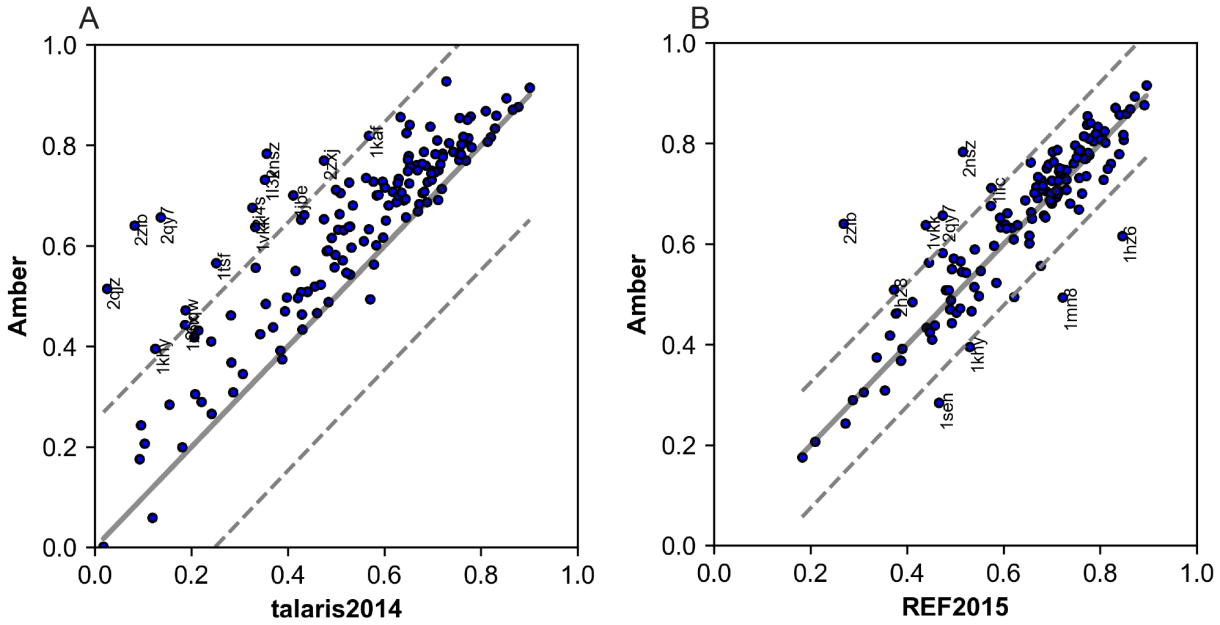


Figure 2. Scatterplots to depict general performance of Rosetta talaris2014 scoring function vs. Amber scoring function (A) and Rosetta REF2015 scoring function vs. Amber scoring function (B) over entire decoy discrimination set. Each dot represents the *B* metric for one system. The black line is x=y and the dashed line represents the 95% prediction interval. Any points that lie outside the 95% prediction interval are annotated with the PDB ID of that system.

We examined cases in which either Amber, Rosetta, or both were unable to correctly rank high-RMSD decoy conformations, scoring them as low-scoring instead of high-scoring. A false minimum is defined as a decoy within the top-10 ranked decoys that has a C-α RMSD from native of greater than 5 Å. Three of these cases are shown in Figure 1. 2QY7 (Figure 1A,D) has several false minima for Rosetta but none for Amber. Generally, Rosetta alone had at least one false minimum in 16% of structures. 1T2I (Figure 1B,E) has a false minimum for Amber but none for Rosetta; 1.32% of systems have at least one false minimum for Amber alone. 1SEN

(Figure 1C,F) has false minima for both Amber and Rosetta, as do 14% of overall structures (Table 1).

Table 1. *B* metric, false minima, and model selection summary comparisons for Amber ff14SBonlySC, Rosetta talaris2014, and Rosetta REF2015 energy functions.

	No. Cases/Total No. Proteins
Decoy Discrimination	
ff14SBonlySC B > talaris2014 B by 0.1	54/150
talaris2014 B > ff14SBonlySC B by 0.1	0/150
ff14SBonlySC B > REF2015 B by 0.1	6/140
REF2015 B > ff14SBonlySC B by 0.1	9/140
False minima in ff14SBonlySC only (not talaris2014)	2/150
False minima in talaris2014 only (not ff14SBonlySC)	24/150
False minima in ff14SBonlySC and talaris2014	21/150
False minima in REF2015 only (not ff14SBonlySC)	0/140
False minima in ff14SBonlySC and REF2015	10/140
Minimum-sum RMSD < ff14SBonlySC selected RMSD by 1 Å	10/150
Minimum-sum RMSD < talaris2014 selected RMSD by 1 Å	21/150
Minimum-sum RMSD < ff14SBonlySC selected and talaris2014 RMSD by 1 Å	1/150
Loop Modeling	
ff14SBonlySC B > talaris2014 B	15/39
talaris2014 B > ff14SBonlySC B	7/39

Superimpositions of false minima decoys with native decoys show their distinct non-native conformations involving both misprediction of secondary structure elements as well as their

incorrect relative placement in tertiary structures. In the case of 2QY7, a Rosetta false minimum, the four helical bundle found in the native structure is perturbed in the false minimum, as the order of the first two helices is reversed; thus, they do not contact the other two helices as tightly as that of the native structure (Supplementary Figure 1E, I-J). The difference between the native structure of 1T2I and its Amber false minimum is more subtle. While the contact maps for the native and false minimum conformations are similar, except for a small contact region in the native structure between residues 40 and 59 that does not appear in the false minimum (Supplementary Figure 1K-L), the false minimum is slightly more compact and has a more ordered secondary structure. Two beta sheet regions in the false minimum are beta strands/unordered in the native structure and two alpha helices in the false minimum are beta strands in the native structure (Supplementary Figure 1F-G).

The case of 1SEN, which has false minima for both Rosetta and Amber, is similar to 1T2I in that the false minima are more ordered than the native structure, although the native structure forms more contacts between remote regions than do the false minima (Supplementary Figure1A-D). Residues 85-96 form a tight beta hairpin in the false minima, whereas the native residues 85-96 has a longer loop between the beta strands, resulting in a shorter, less tight, beta hairpin. Additionally, residues 94-109 in the native are entirely disordered, while that of the false minimum begins as a beta strand and ends in an alpha helix (Supplementary Figure 1H). Decoys that are predicted as false minima often have the same overall structure and contact maps as native structures, yet secondary structure differences may result in large structural deviation. In some cases, false minima contain more ordered secondary structures yet fewer contacts than native conformations; the propensity away from disordered loops may result in lower energies for these false minima.

Per-residue Rosetta energy decomposition

To investigate whether certain residues, structural elements, or energy terms contribute more to false minima conformations, we analyzed the per-residue score decomposition for Rosetta scores for the three systems outlined above (2QY7, 1T2I, and 1SEN). We were unable to perform the same decomposition for Amber as the GB solvation term is not pairwise-decomposable[28]. We calculated the Z-scores for each residue over the lowest-scoring native and false minimum conformations. We identified residues as possibly implicated in false minima if the false minimum residue Z-score score was lower than the native residue Z-score by at least one (i.e. the distance between the two was greater than one standard deviation). We have highlighted these residues (Figure 3A-C). False minima contributing residues were distributed over the conformations and did not cluster to any particular region. Moreover, false minima contributing residues were found in various types of secondary structure: alpha helices, beta strands, and loops. It is therefore currently not possible to attribute Rosetta false minima to any single per-residue propensity, but as expected, several small errors in energy estimation may lead to the observed incorrect scoring.

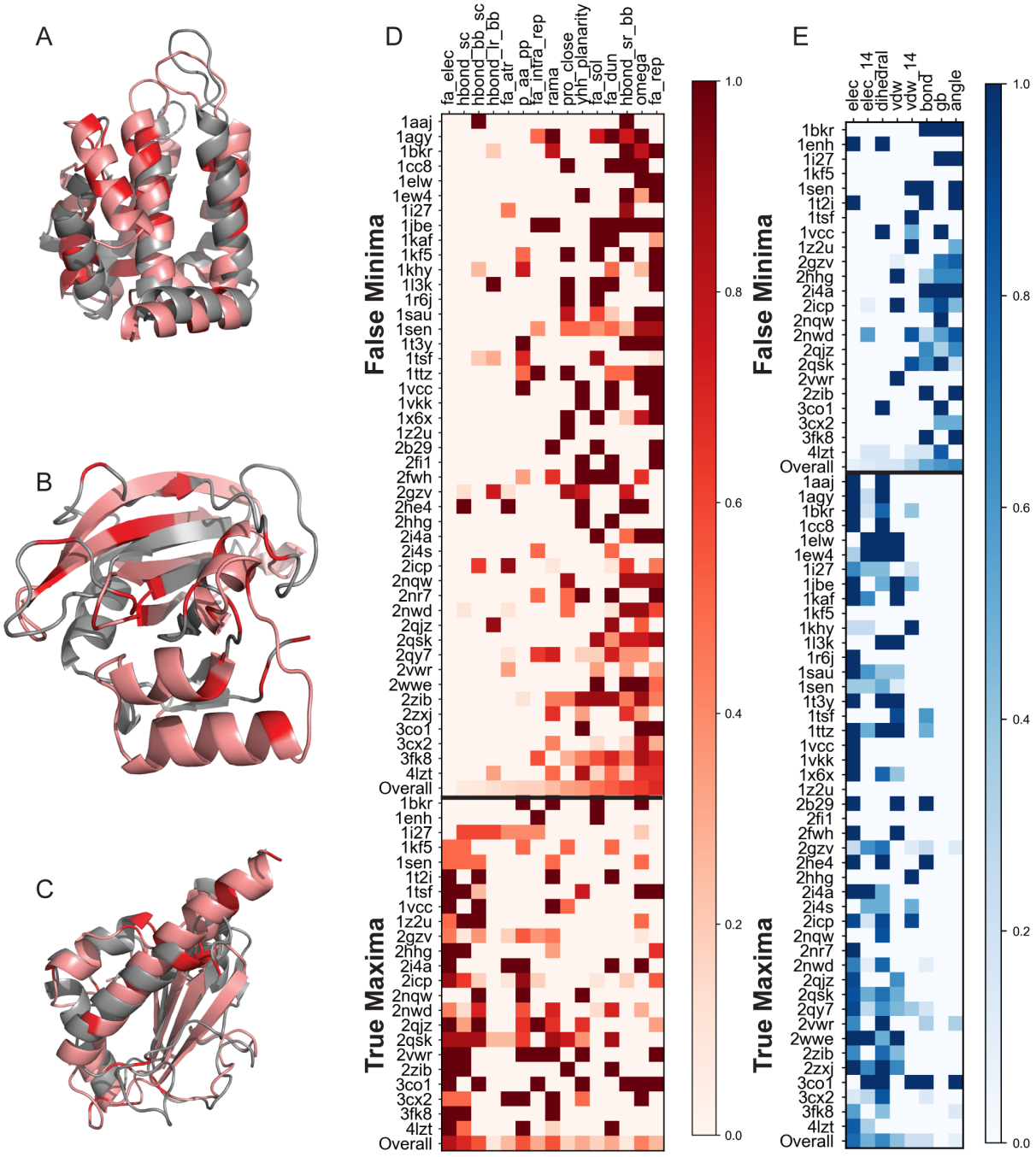


Figure 3. Per-residue and per-score-term propensity of score-functions toward false minima. (A-C) Native (gray) and Rosetta-minimized (salmon) structures of 2QY7, 1T2I, and 1SEN respectively. Rosetta-minimized residues that are scored by Rosetta as greater than 1 standard deviation away from the corresponding native residue are highlighted in red. Heatmaps of per-

structure, score-term contribution to Rosetta-determined (D) and Amber-determined (E) false minima and true maxima. The row marked Overall shows the percentage of structures that indicate some degree of implication for that score-term.

Per-scoreterm contributions of Amber and Rosetta

We reasoned that insight about the performance and pathologies of each energy function could be gained by identifying the energy terms that are responsible for correct and incorrect evaluations within the same energy function. For example, we asked which terms in the Amber energy function help it avoid mis-scoring a decoy (called Amber true maximum) that is identified as a false minimum in the Rosetta landscape (called Rosetta false minimum), and vice versa.

We identified terms that contribute to false minima and true maxima by calculating the Z-scores per decoy set and native set for each protein. If the lowest native score-term Z-score is greater than the false minimum score-term by at least one, that term is implicated in that false minimum. The reverse (i.e. true maximum score-term Z-score is greater than the lowest native score-term Z-score by at least one) is true for identifying true maximum contributing score-terms. The heatmap in Figure 3D depicts the fraction of Rosetta false minima decoys (top) and true maxima decoys (bottom) that show some degree of implication for each score-term. This is calculated both on a per-protein basis and over the entire false minima/true maxima sets. Several score-terms, including hbond_sr_bb, fa_dun, fa_rep and omega, are implicated in a majority of false minima in the Rosetta talaris2014 energy function. A set of other score-terms contribute to a majority of Rosetta true maxima (or Amber false minima), including rama, hbond_bb_sc, hbond_sc, p_aa_pp, and fa_elec. These are score-terms that are not usually implicated in Rosetta false minima, thus demonstrating that the score-terms that contribute to the two trends (towards

false minima and true maxima) are mutually exclusive. Except fa_elec, the other terms identified as helping "rescue" Amber false minima are all PDB-statistics derived, and it is not surprising that they are implicated in correcting the errors of the more physics-based Amber energy function.

We next performed a similar analysis on Amber score-terms for both Amber false minima and Amber true maxima (Figure 3E). We found that bond, angle, and gb are responsible for more than 50% of Amber false minima and that dihedral and elec are implicated in rescuing Rosetta false minima (Amber true maxima). We found that score-terms that are responsible for false minima are not implicated in true maxima and vice versa. Similar to the identification of statistically-derived terms in Rosetta as being responsible for correctly scoring Amber false minima, we find that physics-based terms, i.e., elec (which is counterbalanced by gb) and dihedral potentials, that are orthogonal to the talaris2014 Rosetta scorefunction, are implicated in the rescue of Rosetta false minima by Amber.

Combining rankings to select decoys improves decoy selection

Based on the results above indicating that the rescue of false minima in the landscape generated by one energy function can be effected by the use of the other energy function due to additional terms or different parameterization of terms, we sought to develop an approach to productively combine the two landscapes for model selection. In model selection (for example in protein structure prediction) the challenge is to select a near-native conformation from a set of decoy conformations based on one or more energy values or other features. Typically, an energy value obtained from a single energy function is used. In the current benchmark set, if model selection is performed by the Rosetta and Amber energy functions individually, the Rosetta

lowest-scored decoy has an RMSD of > 5.0 Å for thirteen out of 150 systems, while the lowest-scored Amber decoy has an RMSD of > 5.0 Å for seven systems (four of which overlap with the aforementioned Rosetta systems). We designed a minimum-sum based algorithm (see Methods) to select a decoy conformation based on both sets of ranks to improve the chances of selecting a near-native decoy.

We found that our minimum-sum algorithm improved model selection for both Rosetta and Amber rankings (Figure 4D-E), although it improved model selection for Rosetta to a greater extent. The minimum-sum selected decoy had a lower RMSD than the lowest-scoring Rosetta decoy by at least 1 Å for ten out of the thirteen cases mentioned above and a lower RMSD than the lowest-scoring Amber decoy by at least 1 Å for four out of the seven cases mentioned above. More generally, the minimum-sum selected decoy had a lower RMSD than the lowest-scoring Rosetta decoy for 22 out of 150 cases and a lower RMSD than the lowest-scoring Amber decoy for 11 out of 150 cases.

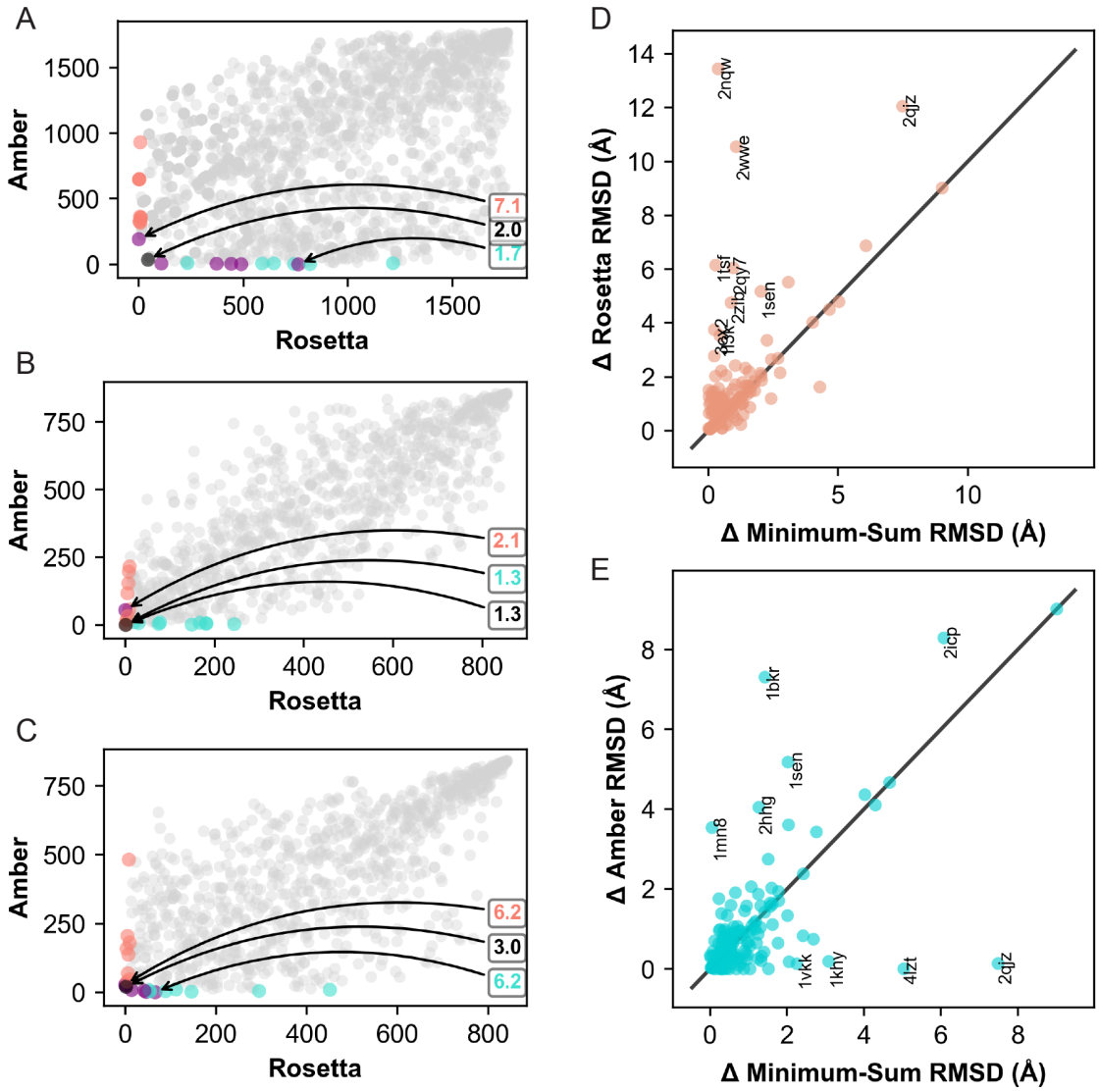


Figure 4. Minimum-sum model selection. (A-C) Scatterplots of Rosetta-rank vs. Amber-rank for all decoys of 2QY7, 1T2I, and 1SEN respectively. Each point represents one decoy conformation. The set of Pareto solutions is purple, the top-10 ranked Amber decoys are turquoise, the top-10 ranked Rosetta decoys are salmon, and the minimum-sum solution is black. Annotations represent the RMSD in Å from native for the top-ranked Amber decoy (turquoise), top-ranked Rosetta decoy (salmon), and minimum-sum solution (black). Scatterplots that show the efficacy of the minimum-sum solution at minimizing the distance from native relative to the

top-ranked Rosetta decoy (D) and top-ranked Amber decoy (E). Each point represents one system. The x-axis is the difference between the minimum-sum solution RMSD from native and the RMSD of the minimum-RMSD decoy conformation, while the y-axis is the difference between the Rosetta lowest-ranked conformation (D) or Amber lowest-ranked conformation (E) RMSD from native and the RMSD of the minimum-RMSD decoy conformation. Points that fall outside the 95% prediction interval are annotated.

We examined the false minima cases described above (2QY7, 1T2I, and 1SEN) and found that the minimum-sum decoy generally had a lower RMSD than that of Rosetta- or Amber-selected decoys (Figure 4A-C). However, for 2QY7, which contains a false minimum for Rosetta but not for Amber, the Amber-selected decoy had a slightly lower RMSD than that of the minimum-sum decoy (1.7 Å vs. 2.0 Å). Nevertheless, the minimum-sum selected decoy RMSD is significantly lower than that of the Rosetta-selected decoy (2.0 Å vs. 7.1 Å). Thus, a minimum-sum framework allows combining the two energy functions productively to select a near-native model.

Loop Modeling

The conformational variability of loops plays a multi-functional role in protein structure and function. They are implicated in stability and folding pathways[34], binding and active sites[35,36], and binding other proteins[37,38]. Efficient sampling algorithms have been developed[36–39], but loop structure prediction efforts can be limited by energy functions, as the energy gaps between loops are smaller and minima are narrow[40]. Therefore, we tested both Amber and Rosetta energy functions on a loop modeling benchmark obtained from T. Kortemme and S. O'Connor. In this benchmark, most of the structure remains the same over the set of decoys; the difference lies in a

small loop region, which can vary highly in RMSD. The energy gaps between structures are therefore smaller; thus, loop modeling provides a more stringent test to distinguish between energy functions.

We found that Amber ranked loops more accurately than did Rosetta (Figure 5C). Several systems had significantly higher *B* values with Amber than with Rosetta. Figure 5A depicts the energy landscapes for one of these structures (1TCA). The Amber funnel is steeper than that of Rosetta, which is reflected in its higher *B* (0.86 vs. 0.37). The lowest-energy and highest-energy loop conformations are shown for both Rosetta and Amber in Figure 5B. Both Rosetta and Amber rank the lowest-energy and highest-energy conformations correctly.

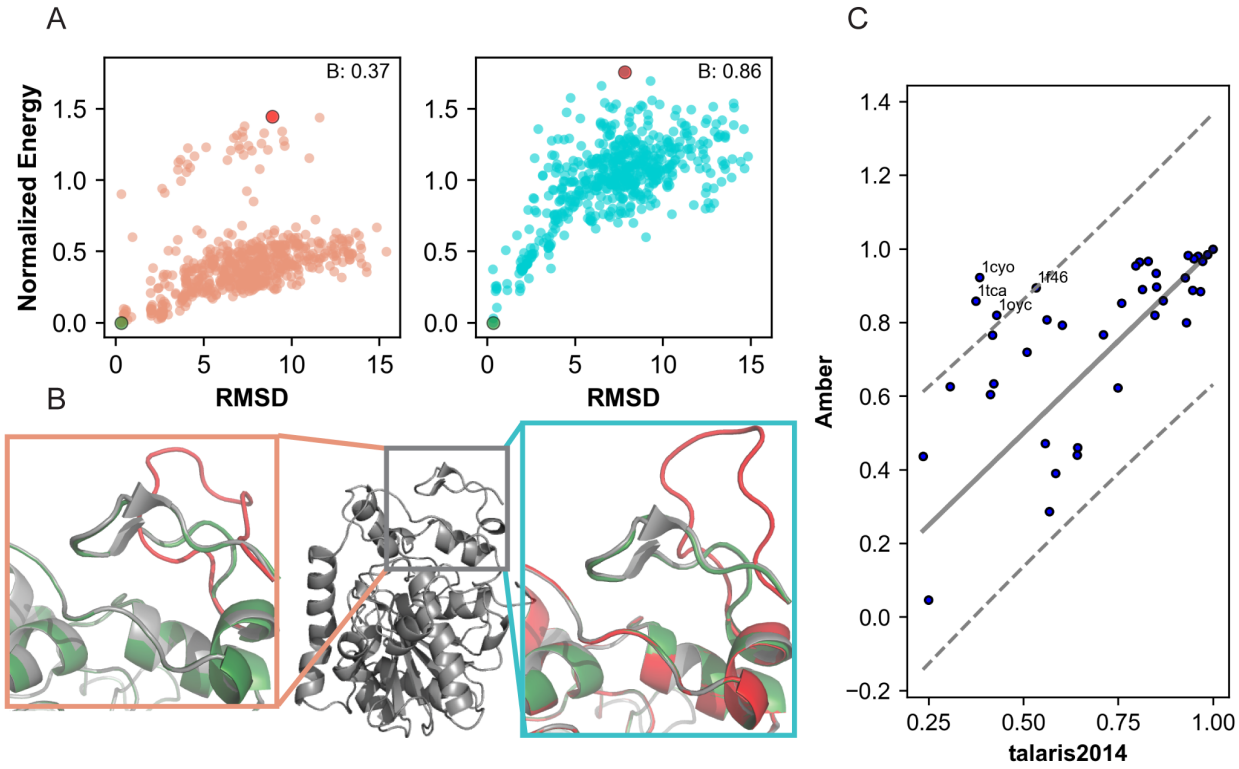


Figure 5. Loop modeling benchmark. (A) Energy landscape for 1TCA. Each dot on the plot represents one decoy conformation. The x-axis is RMSD from native and the y-axis is normalized energy. The *B* metric, which represents the efficacy of the score-function at

differentiating between native and non-native decoys, is shown at the top right corner of each plot. Rosetta plots are to the left, in salmon, and Amber plots are to the right, in turquoise. The lowest-energy decoy conformation in each plot is shown in green and the highest-energy decoy conformation is shown in red. (B) Native structure of 1TCA (gray) and close-ups of loop conformations for lowest-energy decoys (green) and highest-energy decoys (red) for Rosetta (salmon box) and Amber (turquoise box). (C) General performance of Rosetta talaris2014 scoring function vs. Amber scoring function over the entire loop modeling set. Each dot represents the *B* metric for one system. The black line is x=y and the dashed line represents the 95% prediction interval. Any points that lie outside the 95% prediction interval are annotated with the PDB ID of that system.

Discussion

Systematic comparison of Amber ff14SBonlySC (a physically-derived energy function) and Rosetta talaris2014 (both physical and statistical based) reveals the strengths and weaknesses of each energy function. Generally, Amber ff14SBonlySC performs better than Rosetta talaris2014 at both decoy discrimination and loop modeling. However, comparison of Amber ff14SBonlySC to Rosetta REF2015 (the newer, default Rosetta energy function) reveals that REF2015, which has more physically-derived terms than talaris2014, performs comparably well to Amber ff14SBonlySC. Examination of Rosetta talaris2014 score-terms that rescue Amber ff14SBonlySC false minima and Amber ff14SBonlySC score-terms that correct Rosetta talaris2014 false minima reveals two possible sources for the performance improvement of REF2015. While two of the Rosetta score-terms and two of the Amber score-terms that contribute to the correction of false minima are counterparts to each other (Amber dihedral and Rosetta rama, and Amber elec and Rosetta fa_elec), subtle nuances in their derivation and

parameterization appear to influence the propensity of each energy function toward false minima. Although rama and dihedral both score the propensity of the backbone dihedral angles, rama does so in a statistically-derived manner while dihedral is based on fits to quantum chemistry data. Both elec and fa_elec are derived from a Coulombic model, yet they are differently parameterized; the Amber elec is parameterized *via* small-molecule properties, whereas fa_elec is optimized on larger biomolecular structures. The improvement of Rosetta REF2015 over Rosetta talaris2014 may be caused by its greater inclusion of physical-derived terms (bond, angle, etc.) and/or its parameterization on both small-molecule properties and larger biomolecular structures.

Model selection, or the ability to select a near-native decoy from a set of decoy conformations is a general problem in protein structure prediction. If low-energy decoys exist in false minima in the energy landscape, it is difficult to identify conformations that are near-native. Since Amber and Rosetta provide different, semi-orthogonal information, a framework to combine the two rankings enable the identification of near-native decoys. The minimum-sum based algorithm that we have implemented improves model selection for 15% of structures over Rosetta model selection and 7.3% of structures over Amber model selection. The model selection algorithm is extensible to any two sets of energy functions or model ranks for one set of models and can thus be used to combine any two sources of information to produce meaningful improvements in near-native decoy selection.

The approach described here should enable comparative analysis and combination of future versions of both Amber and Rosetta scoring functions, and enable a variety of biomolecular modeling tasks.

ASSOCIATED CONTENT

Supporting Information.

The following files are available free of charge.

LoopDefs: definitions for the loops in the loop modeling benchmark (xlsx)

Supplementary_Software: descriptions of software and scripts used in Methods (docx)

Figures: S1, False minima contact maps and structures; S2, Energy landscapes for all decoy discrimination systems; S3, plot of minimum, minimum-sum, Rosetta, and Amber RMSDs; S4, Energy landscapes for all loop modeling systems. Tables: S1, B-metric values for Amber and Rosetta decoy discrimination systems; S2, values of RMSD for minimum-sum-selected, Rosetta-selected, and Amber-selected models as well as the actual lowest-RMSD value; S3, B-metric values for Amber and Rosetta loop modeling systems (PDF)

AUTHOR INFORMATION

Corresponding Author

*Emails: sagar.khare@rutgers.edu, david.case@rutgers.edu

Present Addresses

†Schrodinger, Inc. 120 West 45th Street, 17th Floor, Tower 45, New York, NY 10036

Author Contributions

ABR, KMB, SDK and DAC wrote the manuscript. All authors have given approval to the final version of the manuscript.

Funding Sources

This material is based upon work supported by RosettaCommons and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1433187 (ABR).

ACKNOWLEDGMENT

We thank F. DiMaio, H. Park for the REF2015 dataset and S. O'Connor and T. Kortemme for the loop modeling dataset.

ABBREVIATIONS

RMSD, root-mean-square-deviation; REF2015, Rosetta Energy Function 2015; LBFGS, Limited-memory Broyden-Fletcher-Goldfarb-Shanno; lbfgs_armijo_nonmonotone, LBFGS minimizer implementation with inexact line search conditions; GB-neck2, generalized Born implicit solvent model.

REFERENCES

(1) Anfinsen, C. B. *Science* **1973**, *181* (4096), 223–230.

(2) Lu, H.; Skolnick, J. *Proteins Struct. Funct. Genet.* **2001**, *44* (3), 223–232.

(3) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.

(4) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.;

Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.

(5) Jernigan, R. L.; Bahar, I. *Current Opinion in Structural Biology*. 1996, pp 195–209.

(6) Shen, M.-Y.; Sali, A. *Protein Sci.* **2006**, *15* (11), 2507–2524.

(7) Ponder, J. W.; Case, D. A. *Advances in Protein Chemistry*. 2003, pp 27–85.

(8) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268* (1), 209–225.

(9) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

(10) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.

(11) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(12) Xu, D.; Zhang, Y. *Proteins Struct. Funct. Bioinforma.* **2012**, *80* (7), 1715–1735.

(13) O'Meara, M. J.; Leaver-Fay, A.; Tyka, M. D.; Stein, A.; Houlihan, K.; Dimaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B. *J. Chem. Theory*

Comput. **2015**, *11* (2), 609–622.

(14) Shapovalov, M. V.; Dunbrack, R. L. *Structure* **2011**, *19* (6), 844–858.

(15) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *J. Chem. Theory Comput*. **2012**, *8* (9), 3257–3273.

(16) Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *Proteins Struct. Funct. Bioinforma*. **2012**, *80* (8), 2071–2079.

(17) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLoS One* **2012**, *7* (2).

(18) Park, H.; Bradley, P.; Greisen, P.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; Dimaio, F. *J. Chem. Theory Comput*. **2016**, *12* (12), 6201–6212.

(19) 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. *J. Chem. Theory Comput*. **2017**, *13* (6), 3031–3048.

(20) Beauchamp, K. A.; Lin, Y. S.; Das, R.; Pande, V. S. *J. Chem. Theory Comput*. **2012**, *8* (4), 1409–1414.

(21) Cino, E. A.; Choy, W. Y.; Karttunen, M. *J. Chem. Theory Comput*. **2012**, *8* (8), 2725–2740.

(22) Nguyen, H.; Roe, D. R.; Simmerling, C. *J. Chem. Theory Comput*. **2013**, *9* (4), 2020–2034.

(23) Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. *J. Am. Chem. Soc.* **2014**, *136* (40), 13959–13962.

(24) Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D. *Protein Sci.* **2014**, *23* (1), 47–55.

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

(26) Tyka, M. D.; Keedy, D. A.; Andre, I.; Dimaio, F.; Song, Y.; Richardson, D. C.; Richardson, J. S.; Baker, D. *J. Mol. Biol.* **2011**, *405* (2), 607–618.

(27) Tyka, M. D.; Jung, K.; Baker, D. *J. Comput. Chem.* **2012**, *33* (31), 2483–2491.

(28) Case, D. A.; Betz, R. M.; Cerutti, D. S.; Cheatham III, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; Kollman, P. A. *AMBER 2016*; University of California: San Francisco, 2016.

(29) Liu, D. C.; Nocedal, J. *Math. Program.* **1989**, *45* (1–3), 503–528.

(30) Nguyen, H.; Roe, D. R.; Swails, J. M.; Case, D. A. *Manuscr. Prep.* **2017**.

(31) Dill, K. a; MacCallum, J. L. *Science* **2012**, *338* (6110), 1042–1046.

(32) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267* (5204), 1619–1620.

(33) Shakhnovich, E. *Chemical Reviews*. 2006, pp 1559–1588.

(34) Xiang, Z.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 7432–7437.

(35) Fiser, A.; Kinh Gian Do, R.; Sali. *Protein Sci.* **2000**, *9*, 1753–1773.

(36) Murphy, P. M.; Bolduc, J. M.; Gallaher, J. L.; Stoddard, B. L.; Baker, D. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (23), 9215–9220.

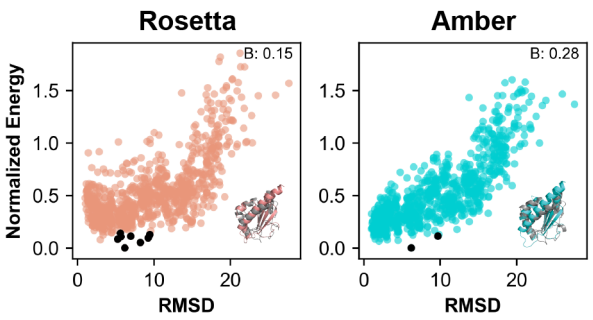
(37) Wang, C.; Bradley, P.; Baker, D. *J. Mol. Biol.* **2007**, *373* (2), 503–519.

(38) Mandell, D. J.; Coutsias, E. a; Kortemme, T. *Nat. Methods* **2009**, *6* (8), 551–552.

(39) Ollikainen, N.; Smith, C. A.; Fraser, J. S.; Kortemme, T. *Methods Enzymol.* **2013**, *18* (9), 1199–1216.

(40) Stein, A.; Kortemme, T. *PLoS One* **2013**, *8* (5).

For Table of Contents Only


Rosetta
Amber
B: 0.15
B: 0.28
Normalized Energy
RMSD
RMSD
0.0
0.5
1.0
1.5
0
10
20

1