SCIENTIFIC REPORTS
natureresearch

OPEN

Neighborhood Preference of Amino Acids in Protein Structures and its Applications in Protein Structure Assessment

Siyuan Liu[1,2], Xilun Xiang[1,2], Xiang Gao[1,2] & Haiguang Liu[1,3*]

Amino acids form protein 3D structures in unique manners such that the folded structure is stable and functional under physiological conditions. Non-specific and non-covalent interactions between amino acids exhibit neighborhood preferences. Based on structural information from the protein data bank, a statistical energy function was derived to quantify amino acid neighborhood preferences. The neighborhood of one amino acid is defined by its contacting residues, and the energy function is determined by the neighboring residue types and relative positions. The neighborhood preference of amino acids was exploited to facilitate structural quality assessment, which was implemented in the neighborhood preference program NEPRE. The source codes are available via https://github.com/LiuLab-CSRC/NePre.

Despite the advances in protein structure determination methods, the discovery rate of new proteins greatly exceeds the rate of experimental structure determination. New proteins can be discovered by high-throughput genome sequencing using sophisticated genome analysis tools[1–3]. In contrast, the protein structure determination requires complicated procedures to obtain high-quality protein samples that produce sufficiently good experimental signals. For example, the target protein must have a reasonably high expression rate to obtain enough sample, after which the protein is purified, followed by the optimization of crystallization cocktail recipes to yield high-quality crystals for X-ray crystallography[4,5]. Alternatively, the molecules must be labeled using isotopes for specific atoms for nuclear magnetic resonance[6,7]. Recent breakthroughs in cryogenic electron microscopy methods have indicated that structure determination can be achieved without tedious crystallization or isotopic labelling[8]. However, the technology is not yet highly automated and requires extensive computational analysis of a large volume of data for each structure. Limitations to experimental structure determination of protein molecules necessitate the development of methods for protein structure prediction using computational modeling approaches.

Protein structure prediction has a long history marked by prediction contests, such as the Critical Assessment of protein Structure Prediction (CASP), which was first organized in 1994[9,10]. Structure prediction has achieved successes in many cases and is used in numerous applications[11]. Particularly, predicted structures can be combined with experimental data to comprehensively understand the structure and function of molecules[12–16]. In many cases, it is difficult to determine a high-quality structure based solely on experimental information. Hybrid methods that integrate structure prediction results and experimental data are promising for exploiting information from both experimental data and computational modeling or predictions[16–18]. For a structure prediction method to be successful, it must have two components: (1) an algorithm to generate a structure ensemble that includes good models, i.e., at least some models in the ensemble are similar to the correct structure (or the native structure); and (2) a scoring function that can rank the generated structures, so that the good models can be identified. Scoring functions either can guide the sampling of protein conformations to improve sampling efficiency, or can be used independently to assess model quality. The structure ensemble of a protein, also referred to as a decoy set, can be generated using several computational methods. The mainstream methods include homology modeling[19], structure threading[20,21], and segment assembly[22–24]. Advanced sampling algorithms can be applied to ensure the diversity of conformations in decoy sets to increase the chance of sampling the structure with the lowest energy[24]. In this study, we focused on the scoring function used to assess the quality and correctness of each generated model.

[1]Complex Systems Division, Beijing Computational Science Research Center, Haidian, Beijing, 100193, China. [2]School of Software Engineering, University of Science and Technology of China, Hefei, Anhui, 230026, China. [3]Physics Department, Beijing Normal University, Haidian, Beijing, 100875, China. *email: hgliu@csrc.ac.cn

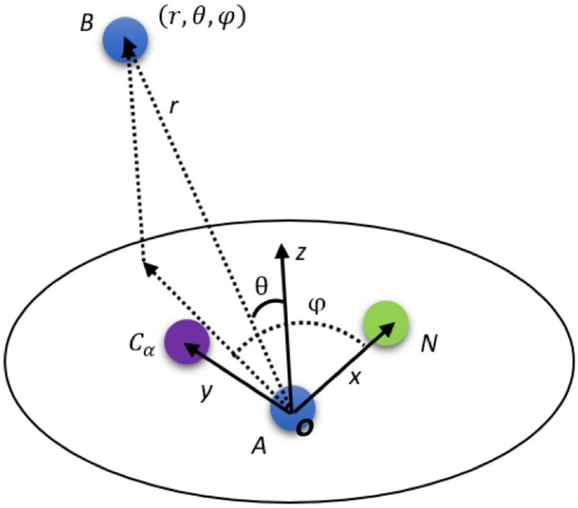

Figure 1. Schematic drawing of two neighboring amino acids (A and B). Location of amino acid B (represented using its geometry center) is shown in the coordinate system of amino acid A, defined with the positions of ***Cα*, *N*** atoms, and its geometric center (***o***).

There are two types of scoring functions. One is based on physiochemical principles—force fields in molecular modeling, such as Amber or Charmm for atomic models[25,26] and Martini or UNIRES for coarse-grained models[27–29]. The other type can be classified as empirical energy functions based on statistical knowledge of experimentally determined structures. There has been tremendous success in applying empirical energy functions to analyze protein structures. One famous example is the protein main chain dihedral angle distributions, known as the Ramachandran plot[30], which is widely used for protein structure validation[31–33]. Representative developments in empirical energy functions include PROSA, DFIRE, DOPE, RW, RWplus, and GOAP[34–38]. Orientation-dependent force fields were used to demonstrate the importance of incorporating the relative positions of amino acids[39,40]. Recent development in machine learning have also led to new frameworks for interaction energy development[41–46]. Inspired by these pioneering works, we developed a new energy function that describes amino acid neighborhood preferences. For each of the 20 natural amino acids, the neighboring amino acid was analyzed in detail with a focus on orientation preferences, described using the polar angle parameters. Distance-dependent energy functions have been well-studied and incorporated into existing methods, thus we focused on orientation-dependent energy functions in this study. Specifically, the preference was determined using 400 (20 × 20) matrices that describe the relative positioning (i.e., orientation) of any two amino acids. For any two amino acids, the probabilities of being neighbors and their relative positions were extracted from a high-resolution structure dataset. The probability distributions were converted to energy functions using the Boltzmann relation, and these energy functions were used to assess the quality of the decoy structures. Based on the results and the performance comparison with several other methods, we found that the neighborhood preference (NEPRE) program is effective for ranking decoy structures and quantifying the correctness of protein structures.

Methods

The native state structures of proteins are stabilized mostly by the interactions between atoms that are not covalently bonded, mainly including electrostatic and van der Waals interactions. Although these interactions are nonspecific, each amino acid is found to have preferences for its neighboring amino acid types, particularly its nearest neighbors. Furthermore, the relative positions of neighboring amino acids are critical for their packing in protein 3D structures. With this in mind, we carried out detailed statistical analysis on the neighborhood preference of each type of amino acid. First, a local coordinate system was established for each amino acid to describe its neighboring amino acid positions; the neighboring residues were mapped to the spherical coordinates defined around the amino acid of interest; these analyses were repeated for every amino acids in the protein structure to obtain statistics for the overall neighborhood preference. The final statistics were obtained from a non-redundant dataset composed of 14,647 PDB structures, with a sequence similarity cutoff at BLAST p-value of 10^{-7} (https://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html)[47]. All chains in the PDB were compared using the BLAST algorithm, followed by clustering using a single-linkage clustering procedure. At a p-value cutoff of 10^{-7}, structures with better sequence completeness and higher resolutions were selected to represent each group. The structures in this dataset were restricted to single-chain proteins to derive the intra-chain neighborhood preferences.

Local coordinate system for each amino acid. The local coordinate system was defined using the main chain atoms of each amino acid, as described previously[48]. This is the foundation of neighborhood analysis for each amino acid. Briefly, the geometry center of each amino acid was calculated as the average position of the associated atoms, then the X-Y plane was defined using the geometry center ($g_{c,}$ labeled as "***o***" in Fig. 1) of the amino acid, nitrogen atom (***N***), and carboxyl carbon atom (***Cα***). The geometry center, ***o***, is set to be the origin point of the local coordinate system for the amino acid (see Fig. 1). The positive x-direction is defined as $\boldsymbol{o} \rightarrow \boldsymbol{N}$, and then the positive y-direction can be defined in the X-Y plane such that the carboxyl atom ***Cα*** has a positive y coordinate. The z-direction is subsequently defined using the right-hand rule (Fig. 1).

The neighboring amino acids were selected based on the distances between the centers of the corresponding amino acid side chains. If the distance was within a given cutoff value (r_c), they were considered as neighbors. Once the neighborhood was defined, statistical analysis was carried out for the amino acids located within the cutoff distance. We used two approaches for the distance cutoff: a universal fixed cutoff for all amino acids or the cutoffs specific to the types of neighboring amino acids.

For the case of a universal fixed cutoff, algorithm performance was tested using various cutoff values, with r_c between 4 and 10 Å. For the type-dependent case, the cutoff was determined by summing the radii of two neighboring amino acids. The radii of 20 amino acids were obtained from the same non-redundant structure dataset.

Statistical model for amino acid contacts in protein molecules. The distribution function is related to energy via the quasi-Boltzmann's relation[49]; particularly, energy can be expressed as:

$$\mathrm{E} = -kTlog\frac{p_{obs}}{p_{exp}} \tag{1}$$

where p_{obs} and p_{exp} are the observed and expected probabilities in the subspace specified with parameters of interest. In NEPRE, p_{obs} and p_{exp} are specified with five parameters, $(i, j, r, \theta, \varphi)$, where (i, j) are the types of amino acids, and (r, θ, φ) represent the relative coordinate parameters of the latter (j) in the former (i) amino acid's local coordinate. To simplify the representations, the geometric center of each neighboring amino acid (***B*** in Fig. 1) was used to describe its location in the local coordinates of centered amino acid (***A*** in Fig. 1). From the structure database, the observation of amino acid type j in the neighborhood of amino acid type i is expressed as $P_{obs}(i, j, r, \theta, \varphi)$, with

$$P_{obs}(i, j, r, \theta, \varphi) = \frac{N_{ij}(r, \theta, \varphi)}{\sum_{i,j} N_{ij}} = \frac{N_{ij}}{\sum_{i,j} N_{ij}} * \frac{N_{ij}(r, \theta, \varphi)}{N_{ij}} = p_{ij} * p_{ij}(r, \theta, \varphi) \tag{2}$$

The expected values of the distribution of various amino acids are expressed as:

$$P_{exp}(i, j, r, \theta, \varphi) = p_i * p_j * r^2 sin\theta \Delta r \Delta\theta \Delta\varphi \tag{3}$$

According to the above derivation, we obtained:

$$E(i, j, r, \theta, \varphi) = -\ kTlog\frac{P_{obs}(i, j, r, \theta, \varphi)}{P_{exp}(i, j, r, \theta, \varphi)} = -\ kT\ \log\left(\frac{P_{ij}}{P_i P_j}\frac{P_{ij}(r, \theta, \varphi)}{r^2 sin\theta \Delta r \Delta\theta \Delta\varphi}\right) \tag{4}$$

where k is the Boltzmann constant, T is the temperature factor, and (r, θ, φ) is the spherical coordinate of amino acid j in the local coordinate system of amino acid i. Because absolute energy values are not required for ranking, we set kT = 1 in the program implementation. If necessary, the energy can be multiplied by kT to obtain physically meaningful values.

For a protein with M amino acids, the total energy ***E*** can be expressed as:

$$E = \sum_{m=1,M}\sum_{n\in\{n\}} E(t(m), t(n), r, \theta, \varphi) \tag{5}$$

where $E(\ldots)$ is the pairwise statistical energy described in Eq. (4), m is the index of the amino acid, $\{n\}$ is the indices of neighboring amino acid within the given distance cutoff of the amino acid m, and t(x) is the function that maps the amino acid indices to their types.

In the NEPRE implementation, the radial distance r was integrated from 0 to the distance cutoff r_c. Therefore, the statistics were simplified to the distributions in the sections specified by the angle parameters (θ, φ) within the contacting sphere. A regular grid system was used to divide the sphere into 20 × 20 regions (see Discussion section for other gridding schemes and the performance comparison in Figures S2 & S3), with angular intervals $\Delta\theta = \frac{\pi}{20}$ and $\Delta\varphi = \frac{2\pi}{20}$ (because the range for θ *is* $[0, \pi)$ and for φ is $[0, 2\pi)$). The unequal volume divisions were corrected by using the appropriate probability in the respective volume (see Eq. 3).

Testing decoy datasets. The performance of the algorithms was tested with publicly available decoy datasets. After a careful literature survey, we identified five published datasets (390 decoy sets in total): the I-Tasser dataset, denoted as I-Tasser(a), and four datasets generated using the 3DRobot programs, including I-Tasser(b), 3DRobot, Rosetta, and Modeller. Information about the datasets is summarized in Table 1. The I-Tasser(a) dataset was generated using the original I-Tasser protocol, where Monte Carlo simulation was used to assemble the structure scaffolds that can be aligned to models in the database[36]. The 3DRobot algorithm extended the I-Tasser method to allow structure sampling without restraints on the fragments, thus generating more diverse conformations for scoring function benchmarking[50].

The decoy structures were evaluated using the proposed NEPRE scoring function. Two metrics were used to characterize the ranking: (1) the success rates in identifying the native structures (or the most native-like structures); and (2) the Pearson correlation between the energy and root-mean-square-deviation (RMSD) with respect to the native structures.

Dataset Name	Protein size	No. of protein decoy sets	Number of structures in each decoy set	References
I-TASSER (a)	47–118 aa	56	400	36
I-TASSER (b)	47–118 aa	56	400	36,50
3DRobot	80–250 aa	200 (48 α-, 40 β-, and 112 α/β-single-domain proteins)	300	50
Rosetta	50–146 aa	58	100	50,55
Modeller	81–340 aa	20	200	50,56

Table 1. Summary of the five datasets.

Results

Type specific neighboring preferences for amino acids.

The 20 natural amino acids appear in protein molecules with different abundances. The probability of finding a specific type of amino acid in the non-redundant dataset is summarized in Fig. 2a, showing that hydrophobic amino acids, such as leucine, alanine, and valine, appear in protein molecules more frequently than the other amino acids. The probabilities of observing two spatially neighboring amino acids for the 20 × 20 pairs were shown in Fig. 2b for the case with a distance cutoff value of 6.0 Å. The neighborhood preference was quantified using the "observed to expected ratio", o/e, defined as $\frac{p(i,j)}{p(i)p(j)}$, shown in Fig. 2c. Certain amino acid types exhibited strong preferences for their neighbors, for example, cysteine strongly prefers another cysteine in its neighborhood (consistent with the observation of disulfide bonds). The type preferences, together with the orientation preferences described in the following, are useful for quantifying the packing of amino acids in protein structures.

Position preferences of amino acids in the neighborhood of each amino acid.

The amino acids interact with each other in their preferred positions, as revealed by the non-uniform distribution of one amino acid within the neighborhood of another amino acid (i.e., within the sphere centered at one amino acid). This provides additional information to the preferred types discussed in the previous section. In Fig. 3a, the pairwise energy function in the angle space (θ,φ) in the spherical coordinate are shown for all 20 × 20 amino acid pairs (see Eq. 4). Figure 3b shows the energy variances for each amino acid pair, corresponding to the distributions shown in Fig. 3a. Larger fluctuations (towards red colors) indicate stronger position preferences and vice versa. The probability distributions (p) and energies (E) of two examples (CYS around CYS, ALA around GLY) are shown in Fig. 3c. Cysteines showed distinguishable preferences for their neighboring cysteine residues, mainly concentrated in the red-colored region around (162°, 270°) in the left figure of Fig. 3c, corresponding to the conformation of disulfide bonds. For the case of alanine around glycine (two figures on the right-hand side of Fig. 3c), the probability distribution was less polarized, indicating that the interactions between alanine and glycine do not have particular strongly preferred positions. Next to the probability distributions, the corresponding pairwise interaction energy functions (Eq. 4) in the angle space are shown.

NEPRE performance of selecting native structures.

As described in the Methods section, the NEPRE program has two implementations depending on the choice of neighborhood cutoff values. One implementation utilizes a fixed cutoff value for all 20 types of amino acids, hereafter named as NEPRE-F (fixed cutoff); the second implementation uses cutoff values depending on the neighboring amino acid radii, named as NEPRE-R (radius-dependent cutoff).

The cutoff value is critical for the neighborhood boundary in the case of NEPRE-F; hence, we tested the scoring function at various cutoff values from 4 to 10 Å for the five datasets described in the Methods section (Table 1). The success rates for identifying the native structure from each decoy set are summarized in Table 2.

The overall performance of NEPRE-F was the best when the cutoff = 6 Å, where the native structures were scored as those with the lowest energy in 266 of 390 decoy sets. The second-best cutoff value was 7 Å, with 248 native structures identified, followed by the case with cutoff = 5 Å, identifying 225 native structures. Based on this criterion, a cutoff = 6 Å is a good choice for selecting native structures from decoys and was used as the default value for structure assessment using NEPRE-F. We also analyzed the correlation between the scoring function and structural difference with respect to the native state in each decoy (quantified using the RMSD with respect to the native structure, see Figure S4 for an example). Interestingly, we found that the correlation increased as the cutoff increased, with the cutoff = 10 Å giving the best Pearson correlation coefficients (Fig. 4). Considering that the ultimate goal of the scoring function is to select the native (or near native) structures from decoys, we used a cutoff = 6 Å as the default parameter in the following analysis.

For the case of NEPRE-R, the radii for each type of amino acid were extracted from the non-redundant dataset. The distributions of radii for 20 amino acids are shown in Figure S1 and the mean values are summarized in Table 3. These values were used to determine the cutoff values in the neighborhood statistics for specific amino acid types. For neighborhood analysis, the associated energy function was derived in the same manner as in the case of NEPRE-F.

Performance of native structure selection compared to other methods.

Using the five decoy sets, we evaluated the performance of several widely used methods with statistical potentials (DFIRE2, DOPE, GOAP, RW, and RWplus), whose executable programs can be obtained from internet. The success rates in identifying the native structure or narrowing down the native structures to a smaller number of candidates (5 or 10) were used to assess performance. As shown in Table 4, NEPRE-R and NEPRE-F showed advantages in recognizing the native

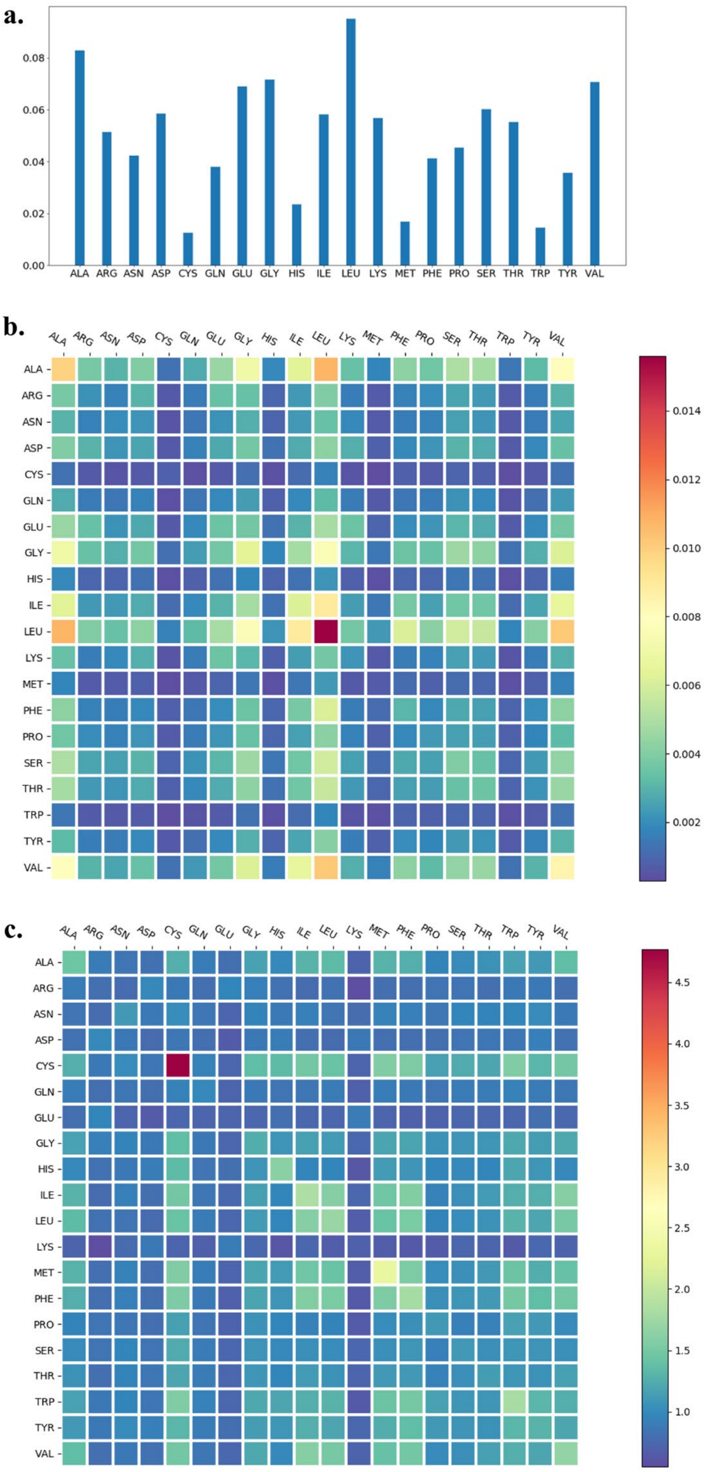

Figure 2. Probability of observing amino acids and amino acid neighbors. (**a**) Amino acid abundance (normalized) in the protein dataset; (**b**) Probability of amino acid pairs in the spatial neighborhood; (**c**) Observed to expected ratios for neighboring amino acids.

structures (represented as the number of TOP1 selected by each scoring function). If the success rates for the native structure included 5 or 10 structures with the lowest energies (indicated with TOP5 or TOP10 in Table 4), rather than the stringent requirement to be the lowest energy structure, we observed that the NEPRE algorithm performed well in all decoy sets, and NEPRE-F yielded better results than NEPRE-R. The scoring function with the best performance in each dataset is highlighted in bold font in Table 4.

Considering that native structures may not be among the decoy set sampled using the model generation programs, we tested the NEPRE's performance on selecting the decoy structures most similar to the corresponding

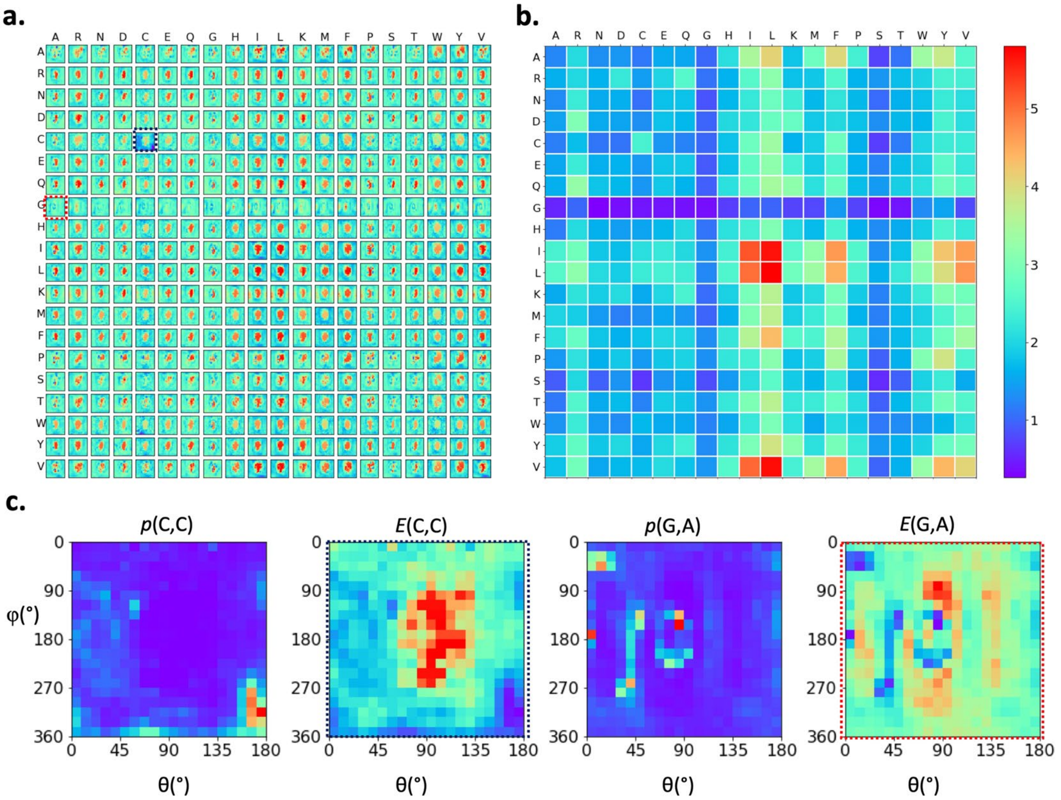

Figure 3. Distributions of amino acid in the neighborhood of another amino acid. (**a**) Position-dependent energy functions of all 20 × 20 pairs of amino acids. Each row shows the neighborhoods centered at a specific amino acid type; and each entry in a row is the interaction energy between a neighboring amino acid and the centered amino acid. (**b**) Energy variances for each amino acid pair shown in (**a**). (**c**) Two representative amino acid neighboring preferences in the forms of probability distributions and energy functions, showing position preferences: cysteine in the neighborhood of cysteine (p(C,C) and E(C,C) shown on the left-hand side) and alanine in the neighborhood of glycine (p(G,A) and E(G,A) on the right). The corresponding energy functions are enclosed with dashed lines in (**a**).

Cutoff	I-TASSER (a)	3DRobot	I-TASSER (b)	ROSETTA	Modeller
4 Å	9/56	40/200	5/56	7/58	9/20
5 Å	44/56	126/200	18/56	25/58	12/20
6 Å	**48/56**	**149/200**	**22/56**	**34/58**	**13/20**
7 Å	49/56	140/200	20/56	25/58	14/20
8 Å	49/56	119/200	19/56	18/58	13/20
9 Å	49/56	104/200	20/56	14/58	11/20
10 Å	49/56	89/200	20/56	9/58	11/20

Table 2. Number of success cases with different neighborhood distance cutoffs.

native structures (i.e., those with the smallest RMSD with respect to the native structures). Interestingly, we found that the performance was not as good in identifying native structures. For example, for the Modeller dataset, NEPRE-R and NEPRE-F identified the native structures in 14 and 13 of 20 cases, respectively, which was much better than for all other methods (see Table 4). In identification of the best decoy structure, the numbers were 6 and 5 (of 20 cases) for NEPRE-R and NEPRE-F, respectively. In contrast, DOPE performed the best in identifying the best decoy structures, succeeding in 8 cases for the Modeller dataset. The other scoring functions showed similar performance as the NEPRE methods. The detailed comparison results indicate that the NEPRE algorithm effectively identified the native (or near native) structures. If no structures are 'native-like' in a decoy set, then the NEPRE may not select the structure with the smallest RMSD. However, the positive correlation between the energy and RMSD is still valid in most cases (see Fig. 4).

Performance on CASP12 decoy datasets. Among the CASP12 targets, we selected a subset of proteins whose native structures have been determined by X-ray crystallography. This dataset contains 39 decoys composed of predicted models submitted by participants of the CASP12. The results of best decoy structure

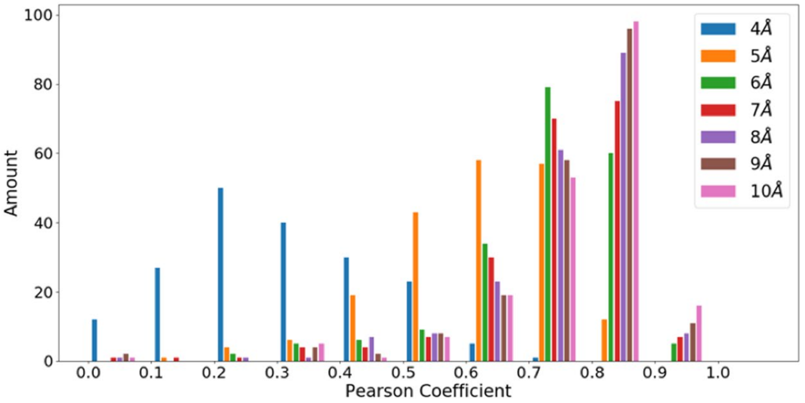

Figure 4. Overall ranking quality measured using Pearson correlation coefficients. The distributions of Pearson correlation coefficients are shown at distance cutoffs from 4 Å to 10 Å.

Amino acid type	Radius (Å)	Amino acid type	Radius (Å)
ALA	3.20	LEU	4.24
ARG	5.60	LYS	5.02
ASN	4.04	MET	4.47
ASP	4.04	PHE	4.99
CYS	3.65	PRO	3.61
GLN	4.64	SER	3.39
GLU	4.63	THR	3.56
GLY	1.72	TRP	5.38
HIS	4.73	TYR	5.36
ILE	3.94	VAL	3.55

Table 3. Amino acid radii extracted from the database.

		DFIRE2	DOPE	RW	RWplus	GOAP	NEPRE-R	NEPRE-F
I-TASSER(a)	Top1[$]	53	48	54	**56**[*]	3	50	48
(56)[#]	Top5	55	48	55	**56**	14	50	48
	Top10	55	49	55	**56**	17	53	50
I-TASSER(b)	Top1	0	11	0	0	3	20	**22**
(56)	Top5	2	25	1	2	3	27	**28**
	Top10	2	29	4	4	3	29	**33**
Rosetta	Top1	0	7	0	0	1	26	**34**
(58)	Top5	2	31	2	2	7	45	**49**
	Top10	5	43	8	8	7	51	**53**
Modeller	Top1	2	6	2	2	1	**14**	13
(20)	Top5	4	8	4	4	5	**16**	**16**
	Top10	6	10	6	7	6	17	**18**
3DRobot	Top1	38	63	0	0	4	129	**149**
(200)	Top5	49	141	5	8	13	160	**173**
	Top10	60	165	9	10	15	176	**183**

Table 4. Performance comparison of different potentials. [#]Number of proteins. [$]The number of cases whose native structures are in the *n* best models selected based on scoring functions. [*]Numbers in bold font indicate the best scoring function for that assessment.

identification are summarized in Fig. 5, in which the performance of NEPRE-F is compared to that of the DOPE, DFIRE, GOAP, and RW programs (the results of the RWplus were very similar to those of RW; see Figure S5). The DOPE program performed the best in identifying the decoy structure with the smallest RMSD with respect to the native structure, which is consistent with the testing results using the 390 generated decoy sets as summarized in the previous section. NEPRE-F outperformed DFIRE and RW in the CASP12 dataset. NEPRE-F identified decoy structures with RMSD < 3 Å in 18 of 39 cases. For the other methods, the number of successful cases was

18, 15, and 16 for DOPE, DFIRE, and RW, respectively. Furthermore, for decoy sets containing structures with RMSD < 5 Å, NEPRE-F successfully identified at least one of these good structures in 26 out of 31 targets (for all participants, eight targets were too difficult to predict any structure with RMSD < 5 Å; see Fig. 5). Thus, NEPRE-F made good predictions in 26 decoy sets by successfully identifying structures with RMSD < 5 Å and failed in five sets (31 decoy sets in total). In CASP performance evaluation, GDT_TS is often used to measure the similarity of the predicted model to the experimentally determined model[51]. The performance of NEPRE was compared to that of the other five potential energies; the results indicate that DOPE and GOAP performed best in selecting models with the best GDT_TS scores, and the performance of NEPRE was very similar to that of the other three methods (see Figure S7).

Discussion and Conclusion

In protein structures, amino acids exhibit preferences for their neighboring amino acids, in both amino acid types and relative positions. This property was systematically studied using experimentally determined structures. Based on the results of neighborhood preference, we developed a new algorithm, NEPRE, which is generally applicable for structure assessment of single-chain proteins. We tested this algorithm using five published decoy datasets (390 decoy sets in total) and a new dataset composed of 39 decoy sets formed the predicted models in CASP12. The performance of the NEPRE algorithm showed potentiality for structure prediction. The execution time was 3–4 s for proteins in the tested decoy sets, including PDB file parsing and energy calculation. Therefore, it is feasible to integrate the NEPRE algorithm in model generation programs to guide the sampling of desired structure ensembles. In addition, we have tried to compare the preferences of amino acids in single chains with that laying in the interfaces of protein complexes. Using a dataset extracted from the PISA database[52], we analyzed amino acid packing at the complex interfaces and found that the neighborhood preferences were similar to the results presented in this study in most cases (see Figure S8), with pronounced differences observed for alanine, cysteine, glycine, and valine. This indicates that the interface specificity of amino acid neighboring preferences are different. The application of NEPRE at the interfaces and detailed comparison of the preferences will be reported in a separate study.

There are two major considerations when discretizing the angle space into the regular grid: (1) the potential energy profile needs to be fine enough to accurately describe the neighborhood (i.e., amino acid packing preferences); and (2) the discretized region is large enough, so that the sampling extracted from the non-redundant database is sufficient for statistical significance. For the first consideration, we evaluated the performance of NEPRE using four discretization schemes, with 15, 20, 25, and 30 discrete sections. The results showed that finer discretization revealed more features, whereas a larger number of discretized regions was under-sampled. As a compromise, we chose N = 20 to balance the two considerations. Furthermore, using the decoys in the Modeller dataset, we compared the ranking performance and found that the performance in the scenario with N = 20 was nearly the same as that in finer discretization cases (see Supplementary materials). It is worthwhile to point out that discretization in angle parameter space can be improved to yield regions with nearly uniform volumes by using algorithms such as Fibonacci lattice or HEALPix[53,54].

The NEPRE algorithm was implemented in two forms depending on the cutoff value defining the neighborhood. The results showed that a cutoff = 6 Å is a good choice for all amino acids types. This performance of NEPRE-F was further validated using CASP12 decoy sets that were not used to determine the optimal distance cutoff values. The performance of NEPRE-F is better than that of NEPRE-R, whose distance cutoff values depend on the neighboring amino acid sizes. Intuitively, type dependent cutoff values should describe more precise interactions between amino acids, and therefore should lead to better performance. While the exact causes for NEPRE-R's inferior performance compared to its peer NEPRE-F (with cutoff = 6 Å) remain unclear, there are several possibilities. The radius values for each amino acid were obtained from the statistics in the protein structures, and the average values may not accurately reflect the neighboring interactions with other amino acids. For example, cysteine and serine each has two peak values, and using a single average value may result in misrepresentation of the neighborhood (see Supplementary materials). A universal fixed cutoff may define a neighborhood that is more suitable for the residue-level scoring function, as the distances between amino acids were measured using distances between their geometry centers.

The NEPRE algorithm performed well in recognizing the native structures in all five decoy datasets. In contrast, most other methods showed variations in their performances across the datasets. For example, the other five methods successfully recognized most of the native structures for dataset I-TASSER(a), but the success rates were lower for other four datasets. This reflects the challenges raised by the four datasets generated with the 3DRobot algorithm, which enhances the conformational diversity. If the native structure is absent (such as in the case of CASP competitions), the performance can be assessed by the success rate in identifying the best decoy structure. We found that the performance is worse in identifying the best decoy structures based on statistical energies. This is a common challenge for many scoring functions, due to the difficulty in accurate ranking near-native structures that have similar scores (energies). In some CASP12 cases, the predicted structures could be far from the native model (such as in decoys of T0932 and T0933, see Fig. 5). Under such circumstances, none of the predicted models are close enough to the native structures, making it very difficult to select the best decoy structures (out of bad predicted models).

The parameters of the NEPRE algorithm were not further fine-tuned, except for the cutoff distance in the NEPRE-F. The probability distributions of relative locations between amino acids were converted to statistical potentials using the Boltzmann relation. Therefore, the NEPRE algorithm is mainly built on the position preferences of neighboring amino acids in protein structures. The performance of the algorithm indicates that the orientation is critical for amino acid packing in protein structures. In this study, both NEPRE-F (with cutoff = 6 Å) and NEPRE-R considered only the nearest neighbors. The distance dependency of the statistical potential function is not explicitly described. In principle, by extending the distance cutoff to larger values, longer-range

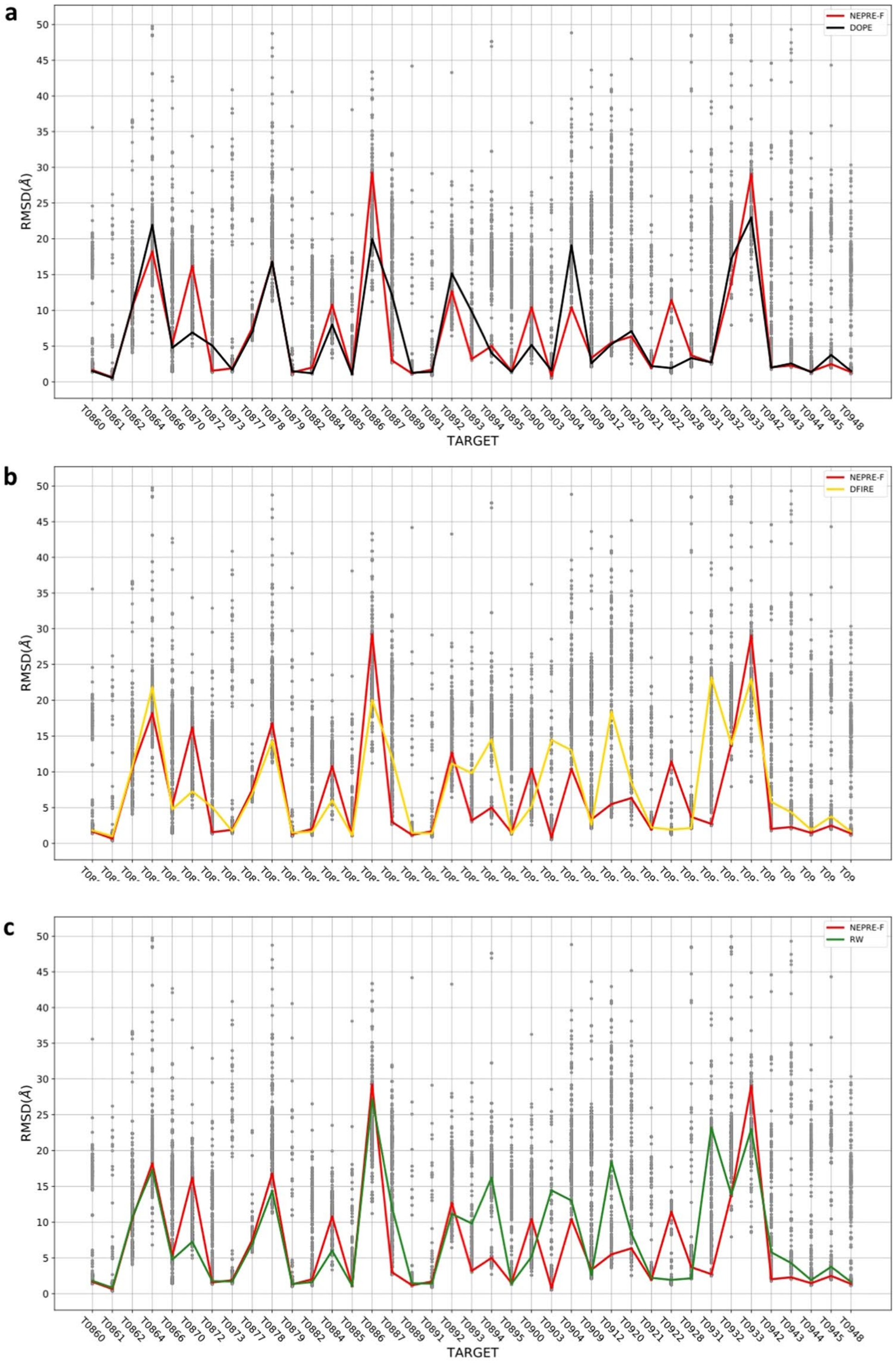

Figure 5. Performance of NEPRE-F in ranking the CASP12 decoy sets. The lines indicate the position of identified decoy structures with the lowest energies. The performances against three other scoring functions are compared: (**a**) DOPE; (**b**) DFIRE; (**c**) RW.

interactions can be described using a similar approach. An immediate extension is to develop a multi-layer neighborhood preference-based energy function by dividing the neighbors into layers. We tested this multi-layer potential energy and found the performance was comparable the the NEPRE-F (with $r_c = 6$) by including the first two layers ($r = 6$ and $r = 7$). The inclusion of more layers deteriorated the performance in recognizing the native structures from the decoys (see Figure S9 and Table S2). The multi-layer implementation may need a weighting scheme and distance-dependent angular space gridding to improve the performance.

In summary, the neighborhood of amino acids in protein structures was statistically analyzed, and the discovered preferences were quantified using the types and relative positions of the neighboring amino acids. Based on the neighborhood preference, the program NEPRE was implemented to assess protein structure quality. The results showed that the NEPRE program can identify native (or near native) structures from decoy sets, providing a foundation for extended applications in protein structure assessment and prediction studies.

Data availability

The source codes for NEPRE are available via https://github.com/LiuLab-CSRC/NePre.

Received: 19 June 2019; Accepted: 24 February 2020;
Published online: 09 March 2020

References

1. Bateman, A. *et al*. UniProt: A hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
2. Kim, M. S. *et al*. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
3. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nature Reviews Genetics* **14**, 35–48 (2013).
4. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology* **18**, 581–586 (2008).
5. Slabinski, L. *et al*. The challenge of protein structure determination-lessons from structural genomics. *Protein Sci.* **16**, 2472–2482 (2007).
6. Markwick, P. R. L., Malliavin, T. & Nilges, M. Structural biology by NMR: Structure, dynamics, and interactions. *PLoS Computational Biology* **4**, e1000168 (2008).
7. Billeter, M., Wagner, G. & Wüthrich, K. Solution NMR structure determination of proteins revisited. *J. Biomol. NMR* **42**, 155–158 (2008).
8. Cheng, Y. Single-particle cryo-EM—How did it get here and where will it go. *Science* **361**, 876–880 (2018).
9. Moult, J. A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology* **15**, 285–289 (2005).
10. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins Struct. Funct. Bioinforma.* **86**, 7–15 (2018).
11. Zhang, Y. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology* **18**, 342–348 (2008).
12. Nealon, J. O., Philomina, L. S. & McGuffin, L. J. Predictive and experimental approaches for elucidating protein-protein interactions and quaternary structures. *International Journal of Molecular Sciences* **18**, 2623 (2017).
13. Schneidman-Duhovny, D. *et al*. A method for integrative structure determination of protein-protein complexes. *Bioinformatics* **28**, 3282–3289 (2012).
14. Dos Reis, M. A., Aparicio, R. & Zhang, Y. Improving protein template recognition by using small-angle X-ray scattering profiles. *Biophys. J.* **101**, 2770–2781 (2011).
15. Latek, D., Ekonomiuk, D. & Kolinski, A. Protein structure prediction: Combining de novo modeling with sparse experimental data. *J. Comput. Chem.* **28**, 1668–1676 (2007).
16. Wang, H. & Liu, H. Determining Complex Structures using Docking Method with Single Particle Scattering Data. *Front. Mol. Biosci.* **4**, (2017).
17. Förster, F. *et al*. Integration of Small-Angle X-Ray Scattering Data into Structural Modeling of Proteins and Their Assemblies. *J. Mol. Biol.* **382**, 1089–1106 (2008).
18. Tuukkanen, A. T., Spilotros, A. & Svergun, D. I. Progress in small-angle scattering from biological solutions at high-brilliance synchrotrons. *IUCrJ* **4**, 518–528 (2017).
19. Martí-Renom, M. A. *et al*. Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
20. Lemer, C. M.-R., Rooman, M. J. & Wodak, S. J. Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins Struct. Funct. Bioinforma.* **23**, 337–355 (1995).
21. Xu, J., Jiao, F. & Yu, L. Protein structure prediction using threading. *Methods Mol. Biol.* **413**, 91–121 (2007).
22. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology* **383**, 66–93 (2004).
23. Lange, O. F. & Baker, D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins Struct. Funct. Bioinforma.* **80**, 884–895 (2012).
24. Lee, J. *et al*. De novo protein structure prediction by dynamic fragment assembly and conformational space annealing. *Proteins Struct. Funct. Bioinforma.* **79**, 2403–2417 (2011).
25. Case, D. A. *et al*. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **26**, 1668–1688 (2005).
26. Brooks, B. R. *et al*. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
27. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & De Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824 (2007).
28. Monticelli, L. *et al*. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **4**, 819–834 (2008).
29. Liwo, A. *et al*. Prediction of protein structure using a knowledge-based off-lattice united-residue force field and global optimization methods. *Theor. Chem. Acc.* **101**, 16–20 (1999).
30. Ramachandran, G. N. & Sasisekharan, V. Conformation of Polypeptides and Proteins. *Adv. Protein Chem.* **23**, 283–437 (1968).
31. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
32. Hooft, R. W. W., Sander, C. & Vriend, G. Objectively judging the quality of a protein structure from a Ramachandran plot. *Comput. Appl. Biosci. CABIOS* **13**, 425–430 (1997).
33. Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. MolProbity: Structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **32**, W615–W619 (2004).
34. Zhang, C. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Sci.* **13**, 391–399 (2004).
35. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).
36. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One* **5**, e15386 (2010).
37. Zhou, H. & Skolnick, J. GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* **101**, 2043–2052 (2011).
38. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins Struct. Funct. Bioinforma.* **17**, 355–362 (1993).
39. López-Blanco, J. R. & Chacón, P. KORP: Knowledge-based 6D potential for fast protein and loop modeling. *Bioinformatics* **35**, 3013–3019 (2019).

40. Karasikov, M., Pagès, G. & Grudinin, S. Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics* **35**, 2801–2808 (2019).
41. Ma, J., Wang, S., Wang, Z. & Xu, J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* **31**, 3506–3513 (2015).
42. Wang, J. *et al*. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **5**, 755–767 (2019).
43. Bhattacharya, D. & Valencia, A. RefineD: Improved protein structure refinement using machine learning based restrained relaxation. *Bioinformatics* **35**, 3320–3328 (2019).
44. Hanson, J., Paliwal, K. K., Litfin, T., Yang, Y. & Zhou, Y. Getting to Know Your Neighbor: Protein Structure Prediction Comes of Age with Contextual Machine Learning. *J. Comput. Biol.* cmb.2019.0193 (2019).
45. Long, S. & Tian, P. A simple neural network implementation of generalized solvation free energy for assessment of protein structural models. *RSC Adv.* **9**, 36227–36233 (2019).
46. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).
47. Gibrat, J. F., Madej, T. & Bryant, S. H. Surprising similarities in structure comparison. *Current Opinion in Structural Biology* **6**, 377–385 (1996).
48. Xiang, X. & Liu, H. IDPM: An online database for ion distribution in protein molecules. *BMC Bioinformatics* **19**, 102 (2018).
49. Finkelstein, A. V., Badretdinov, A. Y. & Ptitsyn, O. B. Physical reasons for secondary structure stability: α-Helices in short peptides. *Proteins Struct. Funct. Bioinforma.* **10**, 287–299 (1991).
50. Deng, H., Jia, Y. & Zhang, Y. 3DRobot: Automated generation of diverse and well-packed protein structure decoys. *Bioinformatics* **32**, 378–387 (2015).
51. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
52. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).
53. Svergun, D. I., IUCr. Solution scattering from biopolymers: advanced contrast-variation data analysis. *Acta Crystallogr. Sect. A Found. Crystallogr.* **50**, 391–402 (1994).
54. Gorski, K. M. *et al*. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *Astrophys. J.* **622**, 759–771 (2005).
55. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
56. John, B. & Sali, A. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* **31**, 3982–3992 (2003).

Acknowledgements

The authors are grateful to Prof. Yang Zhang and Dr. Chengxin Zhang from University of Michigan for the help with the decoys and the potential energy evaluations. This project is supported by National Natural Science Foundation of China (Nos. 11575021, U1530402, U1430237, 1181101028).

Author contributions

H.L. designed the work; S.L., X.X., and X.G. carried out the research, software implementation and analysis; all authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-61205-w.

Correspondence and requests for materials should be addressed to H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.