Deep Ranking in Template-free Protein Structure Prediction

Xiao Chen
Dept of Electrical Engineering and Computer Science
University of Missouri, Columbia

Nasrin Akhter
Dept of Computer Science
George Mason University

Zhiye Guo
Dept of Electrical Engineering and Computer Science
University of Missouri, Columbia

Tianqi Wu
Dept of Electrical Engineering and Computer Science
University of Missouri, Columbia

Jie Hou
Dept of Computer Science
Saint Louis University

Amarda Shehu
Dept of Computer Science
George Mason University

Jianlin Cheng*
Dept of Electrical Engineering and Computer Science
University of Missouri, Columbia

ABSTRACT

The road to the discovery of the biological activities of a protein molecule in the cell goes through knowledge of its three-dimensional, biologically-active structure(s). Current evidence suggests significant regions of the protein universe are inaccessible by either wet-laboratory structure determination or homology modeling. While great progress has been made by computational approaches in elucidating dark regions of the proteome, inherent challenges remain. In this paper, we advance research on addressing one such a challenge known as model (quality) assessment. In essence, the task involves discriminating relevant structure(s) among many computed for a protein of interest. We propose a method based on deep learning and evaluate it on tertiary structures computed by a popular de-novo platform on benchmark datasets. The method uses novel protein residue-residue distance features, improved residue-residue contacts, together with other features, such as energies and model topology similarity, to estimate the quality of protein models. A detailed evaluation shows that the proposed method outperforms related ones and advances the state of the art in model assessment.

CCS CONCEPTS

• **Computing methodologies** → **Bio-inspired approaches**; • **Applied computing** → **Molecular structural biology**; **Bioinformatics**.

*Corresponding Author: chengji@missouri.edu

ACM-BCB '20, August 30 – September 02, 2020, Atlanta, GA, USA

ACM ISBN yyy…$15.00
DOI: 10.1145/3388440.3412469

KEYWORDS

protein structure prediction; model quality assessment; deep learning; distance prediction; contact prediction.

ACM Reference Format:
Xiao Chen, Nasrin Akhter, Zhiye Guo, Tianqi Wu, Jie Hou, Amarda Shehu, and Jianlin Cheng. 2020. Deep Ranking in Template-free Protein Structure Prediction. In *11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '20), August 30 – September 02, 2020, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

The three-dimensional/tertiary structure of a protein molecule determines to a great extent the molecular mechanisms in which it is involved in the cell [7]. While knowledge of the biologically-active/native tertiary structure of a protein holds great promise for decoding its biological function(s), determining this structure poses varying challenges in wet and dry laboratories [32]. Advancements made in wet-laboratory structure determination lag behind the increasingly faster and cheaper high-throughput gene sequencing technologies that have yielded millions of protein sequences [6]. In contrast, the number of known biologically-active/native protein structures is an order of magnitude less. For instance, as of June 2020, the number of experimentally-known structures deposited in the Protein Data Bank (PDB) [5] is around 165, 000.

Furthermore, current analysis suggests that significant regions of the protein universe are never observed by experimental structure determination and are inaccessible to homology modeling [42]; the latter refers to the setting where a protein structure with a sufficiently-similar amino-acid sequence to a protein of interest exists, and can thus be used as a reliable structural template. Work in [42] shows that for 546, 000 Swiss-Prot proteins, 44–54% of the proteome in eukaryotes and viruses is dark, compared with only $\tilde{1}4$ % in archaea and bacteria.

The dark proteome is a great motivation for computational approaches to tackle protein structure prediction (PSP). In particular, the setting where the only direct information about a target protein is its amino-acid sequence is known as de-novo or template-free

PSP. The history of de-novo approaches is rich [28]. While a review of such approaches is beyond the scope of this paper, we highlight here two key challenges.

The first challenge *in silico* is an instance of the curse of dimensionality in modeling. The space of possible tertiary structures of a given amino-acid sequence is vast and high-dimensional; protein molecules are inherently plastic and undergo both continuous and discrete motions in three-dimensional space [47]. Significant advances have been made in this regard and have resulted in many software packages and methods, such as AlphaFold [46], Rosetta [27], Quark [53], and many others [37–39, 54]. A key common characteristic of these approaches is the discretization of the search space by relying on the concept of structural fragments distilled from experimentally-determined native structures and utilized to assemble novel structures of a given target.

The second challenge *in silico* and the direct focus of our enquiry in this paper has to do with our current ability (or, rather, inability) to accurately evaluate the "nativeness" of a computed tertiary structure. The predominant approach is to generate tertiary structures that minimize a designed scoring function. The latter models and accounts for the interaction energy among the atoms in a given tertiary structure. It is now well-known that scoring/energy functions are inherently inaccurate [2, 11, 35]. They often drive the exploration of a protein structure space to local minima that contain structures very different from a known native structure.

In a blind setting, it is unknown which computed structures are sufficiently close to the sought native structure to be deemed near-native. Doing so is known as model accuracy/quality assessment (QA), model selection, or decoy selection; the terms "model" and "decoy" refer to a computed/generated tertiary structure in this context. While there are technical differences between assessment and selection (selection involving further algorithms that act upon the assessed structures), in this paper, we focus on model assessment and specifically evaluate the ability of a data-driven approach that leverages deep learning to assess tertiary structures generated by Rosetta for their nativeness.

In this paper, we propose a deep learning approach for model assessment. The approach builds over many years of research in the Cheng laboratory. The proposed method uses novel protein residue-residue distance features, improved residue-residue contacts, as well as other features, such as energies and model topology similarity, to estimate the quality of tertiary structures of a given target protein. A detailed evaluation shows that the method outperforms related ones and advances the state of the art in model assessment.

Before we relate further details, we place our contributions in context through a brief summary of related work in Section 1.1. The rest of this paper is organized as follows. Section 2 describes the proposed method in greater detail and relates the experimental setup. Section 3 relates the detailed evaluation on diverse benchmark datasets. Section 4 then concludes the paper with a summary and discussion of future work.

1.1 Related Work

Early work revealed that interaction energy was a poor indicator of nativeness [49]. This motivated many researchers to design more accurate scoring functions via which to assess models on an individual basis [12, 29, 48]. These so-known single-model methods are varied in the scoring functions they utilize to assess and rank tertiary structures (the given models). Some use physics-based functions based on physical properties of atomic interactions [26]. Others use knowledge-based/statistical scoring functions that rely on statistical analysis of known native structures [33].

The early challenges with scoring functions steered the community attention towards approaches that ignored energy. Clustering-based methods soon became prominent and dominated the QA category in the Critical Assessment of protein Structure Prediction (CASP) competition [13]. Until recently, clustering-based methods outperformed single-model methods [24]. However, single-model methods have progressed considerably and can now compete with clustering-based methods [25]. Clustering-based methods pose several concerns. For instance, since they are based on consensus, they cannot identify good decoys in sparse, low-quality decoy datasets, where near-native structures are significantly under-sampled.

Currently, quasi-single model methods and supervised learning methods have taken hold in the model assessment community and are shown to outperform clustering-based methods. Quasi-single model methods combine concepts of single- and multi-model methods [17, 41]. Their main approach is to compare decoys to selected, high-quality reference structures [22]. Methods based on supervised learning continue to grow in popularity and diversity. They leverage Support Vector Machines [9, 30], Random Forest [31], Neural Networks [8, 36], ensemble learning [34], and more. Feature sets used are diverse, derived from terms of statistical scoring functions [18, 56] and/or expert-constructed structural features [43, 44].

In particular, machine learning methods started to gain in popularity in early 2000. These methods were developed to predict the quality of protein structures (also referred to as models) using the energies, statistical potentials, or some other features of given models as input [4]. The majority of the early methods produced relative quality scores (e.g. energy) to rank models, which could not measure the absolute quality of the models, e.g., the similarity between a model and the given native structure. A significant advance to that allowed addressing this problem was a data-driven machine learning approach [50] that directly trained an SVM on structural features (e.g. secondary structure, solvent accessibility, residue-residue contact probabilities) of protein models to predict their absolute quality score (e.g., the GDT-TS score [55] that measures distance with respect to unknown native structures).

An increasingly powerful set of supervised learning methods for determining the absolute quality score leverages deep learning [8, 10, 19, 20]. For instance, work in [8] proposes deepQA, a single-model decoy selection method that utilizes energy, structural, and physio-chemical characteristics of a decoy for quality prediction. Improved performance has been observed with models based on convolutional neural networks (CNNs). Work in [19] uses a deep one-dimensional CNN (1DCNN) to build a single-model decoy selection method. The authors make use of two 1DCNNs to predict the local and global quality of a decoy. In [40], the authors propose Ornate, a single-model method that applies a deep three dimensional CNN (3DCNN) for model quality estimation. 3DCNN has also been used successfully in [45]. In particular, the deep learning method DeepRank [10] that uses deep neural networks to integrate various complementary features, such as model

topology similarity, statistical, secondary structure, solvent accessibility, and novel residue-residue contact scores has achieved the best performance in the 13th CASP competition, demonstrating the potential of applying deep learning in this area.

2 METHODS

In this work, we develop DeepRank2, which improves upon DeepRank to better predict the quality of protein models. DeepRank2 utilizes a similar deep architecture of DeepRank, but leverages more features. Below, we relate via a schematic in Figure 1, the shared architecture between DeepRank and DeepRank2. Briefly, the architecture is the multi dense layer neural net. Compare to DeepRank, we increase the hidden layer numbers and hidden units numbers. We also reduce from 10-fold cross-validation to 5-fold cross-validation. With the larger dataset in DeepRank2 training, we add the Batch Normalization layer [21] to speed up the training progress. Also, we utilize the Kaiming initialization [16] for all dense layer in our model.

DeepRank2 is different from DeepRank in three aspects. At an architectural level, the number of layers of deep networks used in DeepRank2 is increased. DeepRank2 is also trained on a larger dataset of protein models than DeepRank. More importantly, DeepRank2 makes use of feature selection to selected the top features from DeepRank feature pool. Specifically, in DeepRank2, a novel feature, the correlation between the residue-residue distances in a protein model and the protein distance map predicted by DeepDist [52], is added to the existing feature set of DeepRank. In addition, the contact score features are generated by comparing the contacts of a structural model with the contact map predicted by our latest residue-residue contact and distance prediction method, DNCON4 [51](unpublished) which training data is 6463 targets from training list used in DMPfold[14]. Feature selection is also applied to select most informative features used by DeepRank.

2.1 Correlation Feature

We introduce a new, correlation feature which is a novel distance-based feature. For each server prediction model of a CASP target, we build two distance maps, A and B. A is an $L \times L$ distance map directly transferred from the server prediction model. B is an $L \times L$ generated by DeepDist; L denotes sequence length in terms of the number of amino acids.

To create the correlation feature, we make use of two thresholds, an index threshold and a value threshold. These thresholds help filter the elements in the A and B matrices. For both matrices, we only consider values above the diagonal, as the matrices are symmetrical. To filter the matrices, we use the following three conditions for any element $M_{i,j}$ in the matrix: (a) $|i - j| \geq index\_threshold$; (b) $M_{i,j} \leq value\_threshold$; and (c) for any index (i, j), $A_{i,j}$ and $B_{i,j}$ should both satisfy conditions (a) and (b).

The remaining elements in each matrix that satisfy the above conditions are then stored respectively in two empty arrays based on the index order (row-major). In this manner, we obtain two equal-length lists. We then calculate the Pearson correlation coefficient between the lists. In our experiments in this paper, we set 6 as the index threshold, because the element might contain limited information if $|i - j| < 6$, for 16 as the value threshold, we consider

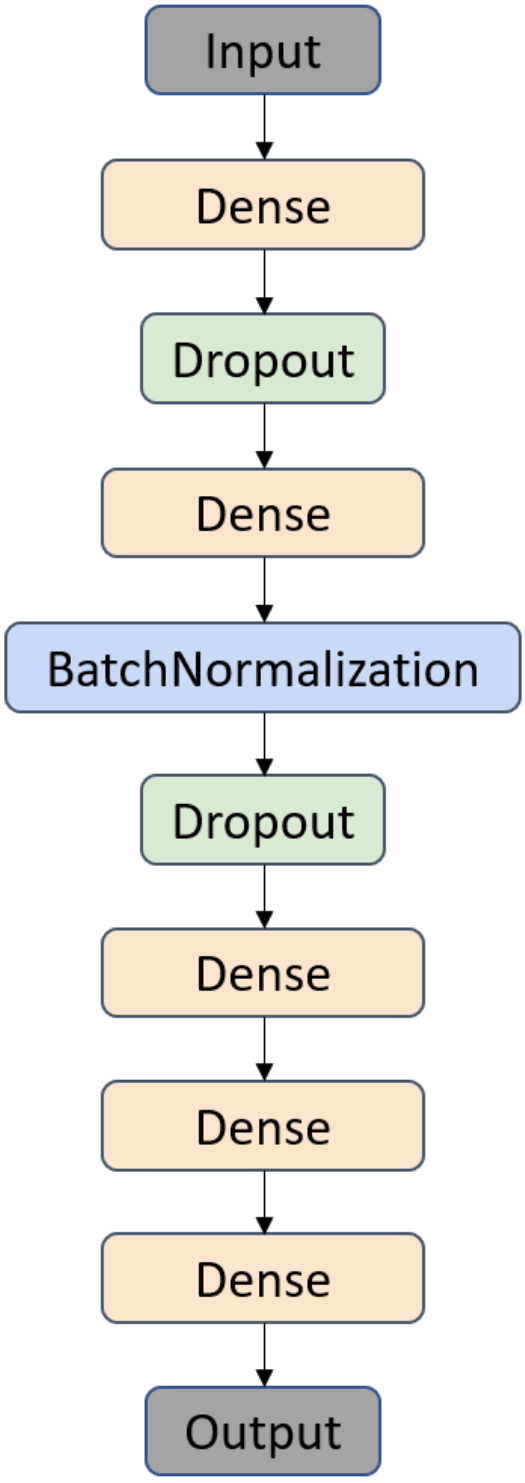

Figure 1: The schematic shows the architecture of DeepRank2. Dropout layer and Batch-Normalization layer help us avoid over-fitting problem.For We set 0.2 dropout rate for each Dropout layer. For each dense layer, we set 3 as maximum norm for the incoming weights

value larger than 16 might be noise for training since current version DeepDist show a higher accuracy for predicting value less than 16.

Figure 2 demonstrates this process. We utilize a specific protein target from our benchmark set, T0949, as an example. Figure 2 shows only the first 10 rows and columns of each of the matrices (A and B); the full size of each matrix is 183×183.

2.2 Feature Selection

The other features employed in DeepRank2 are those employed by our group in DeepRank debuted in CASP13 [19]. In total, the number of features is 23. In this paper, we conduct further analysis to select the top of these 23 features. We use the Light Gradient Boosting Machine (LightGBM) [15] for this purpose. LightGBM is a well-developed software that is popular and state-of-the-art for feature selection. It calculates feature importance based on two options. One option allows to employ the number of times the feature has been used in modeling. Another option is to employ the total gain of the feature in modeling. We choose the first option here. Before selecting the features, we update the contact-based features predicted from DNCON2 [1] to those predicted from DNCON4; we also add the new correlation-based feature (described above) into the feature pool.

We build a regression model based on the resulting 24 features by LightGBM. For this model, we utilize all models from CASP8 to CASP12 as our training data and model from CASP13 for testing. Figure 3 shows the importance of each feature, which ranges from 178 to 1456, with a mean value of 604. We sequentially remove the last feature based on the descending order, and then use the remaining features to retrain the ranking model. We repeat this process until we reach 19 features, which are then used for training DeepRank2. It is worth noting that the correlation-based feature we introduce in this paper is ranked 4th in the feature pool. Specifically, updated features are feature_dncon4_long_range, feature_dncon4_medium_range, and feature_dncon4_short_range, wich are ranked 7th, 8th, and 10th, respectively.

We train two models with the same structure and the same parameters setting on CASP8 to CASP12 then evaluate the performance on CASP13. Model_1 applies 23 features and achieves 0.0492 loss, model_2 get 0.0461 loss with 19 features. Generally, based on table 1, selected 19 features shows a better performance than 23 features on the same experiment environment.

Table 1: Two models with same structure and parameters setting, One model is training with 23 features, another one with 19 features. The lower loss is highlighted by bold font.

#	Model_1(23 features)	Model_2(19 features)
Fold1	0.0471	**0.0462**
Fold2	0.0469	**0.0468**
Fold3	0.0497	**0.0443**
Fold4	**0.0467**	0.0495
Fold5	0.0496	**0.0468**
Ensemble	0.0492	**0.0461**

2.3 Training

To train DeepRank2, we use all server models from CASP8 to CASP12 as our training dataset. Based on our feature selection work, we generate 19 features for each server model (we note that 'each server model' refers to a tertiary structure predicted from a server in CASP). We apply five-fold cross-validation for training; that is, we split the training dataset into five equal-size subsets. Every time, we applied four subsets as training data, using the fifth for parameter tuning. We then select the best-performing model based on performance over the CASP13 dataset. This process is repeated five times. After we obtain five models, we use the average value of the five models as the final prediction.

2.4 Experimental Setup

We present results on two separate testing datasets. The first consists of Rosetta-generated structures for 8 protein targets selected from the CASP12 and CASP13 free-modeling category. These targets are listed in Table 2. We note that this dataset of structures have no overlap with the structures/predictions made by the various servers in CASP8, CASP12, and CASP13 used to train DeepRank2. Specifically, for each of the 8 targets, we use the amino-acid sequence (in FASTA format) and 3-, 9-residue fragments to generate between $36,000$ and $55,000$ tertiary structures via the Rosetta AbInitio protocol [27].

The set of structures generated for each target is further reduced for computational expediency. We do so by down-sampling 600 structures from each dataset as follows. We construct a nearest-neighbor graph embedding of the tertiary structures for each target and then identifies basins (Rosetta all-atom energies are available for each structure); the procedure is described in detail in [3]. A dataset is divided into basins/groups of structures, and we select the $n = 15$ largest basins, selecting from each of them the $600/n$ lowest-energy structures in a basin. In this manner, the dataset Rosetta-generated structures for each target is reduced to 600 low-energy structures.

Table 2: Testing datasets. The target IDs are shown in Column 2. The length of each sequence (number of amino acids) is shown in Column 3.

#	Target ID	Length
1	T0898-D2	55
2	T0953s1-D1	67
3	T0886-D1	69
4	T1008-D1	77
5	T0953s2-D3	77
6	T0960-D2	84
7	T0892-D2	110
8	T0859-D1	113

The second dataset employed to evaluate DeepRank2 consists of all tertiary structures submitted for a given protein target by the various servers participating in CASP13. CASP13 released 90 total targets and canceled 10 of them. In this experiment, we evaluate performance on a total of 80 targets. For each target, we have about 150 server-predicted models/structures. This dataset allows

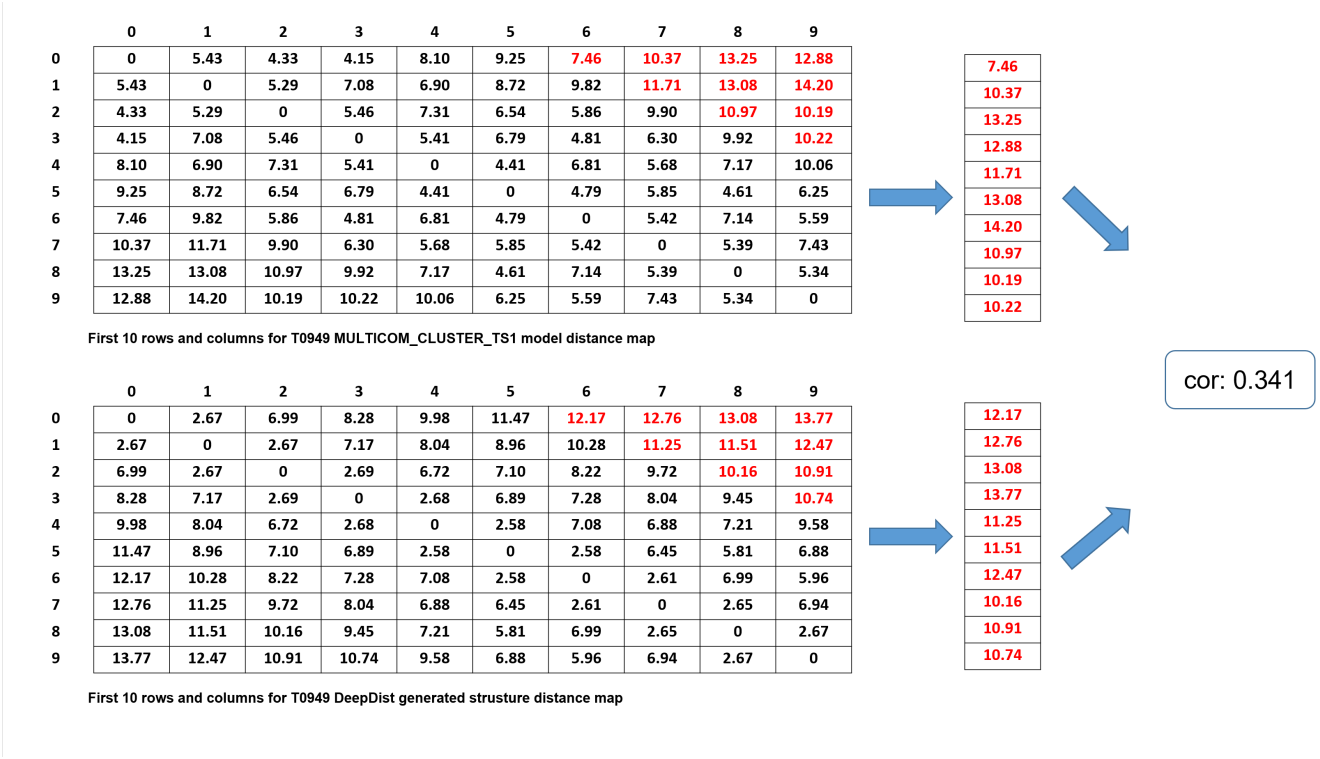

Figure 2: The top matrix is the A distance matrix we only show the first 10 rows and 10 columns on a specific protein target for T0949. The matrix calculated from the tertiary structure predicted for this CASP13 target from the MULTICOM_CLUSTER_TS1 server. The bottom matrix is the B matrix (again we show only the first 10 rows and columns) generated by DeepDist for T0949. Values in red are those that satisfy the conditions detailed above. In this demonstration, the correlation feature value is 0.341.

us to obtain an aggregate view of the performance of DeepRank2, whereas the previous dataset allows us to expose in greater detail the method's performance.

2.4.1 Performance Metrics. The performance metric we employ here as a proxy for the global quality of a structure/model is its GDT-TS similarity score from the native structure. GDT-TS is popular with CASP participants and the protein structure modeling community in comparing two tertiary structures. Specifically, we utilize GDT-TS to compare a model/structure predicted as the top/best model by a method over a dataset of structures available for a protein target and the corresponding native structure for that target.

GDT-TS stands for global distance test and is a measure of similarity between two protein structures with known amino acid correspondences [55]. Its popularity is related to the evidence that it is a more sensitive and accurate metric than the root-mean-square-deviation (RMSD) metric. GDT-TS measurements are used as assessment criteria in CASP. The metric measures the largest set (as a fraction of the whole length) of amino-acid's alpha carbon atoms in structure A falling within a defined distance cutoff of their position in structure B, after superimposing the two structures. GDT-TS is the average over four such scores calculated at 4 consecutive distance cutoffs (1, 2, 4, and 8Å). GDT-TS varies between 0 and 1, with higher values indicating higher structural similarity.

3 RESULTS

We evaluate the performance of DeepRank2 in comparison to DeepRank, as well as SBROD[23]. SBROD stands for "Smooth orientation-dependent scoring function." This scoring function evaluates tertiary structures at backbone resolution; that is, only interactions among backbone atoms are considered. SBROD now denotes more generally a method of ranking tertiary structures by this scoring function to assess them. The top structure according to this function is offered as prediction. Each method is applied to tertiary structures available for a target to predict a top structure/model per target.

3.1 Evaluation on Rosetta-generated Tertiary Structures

We first report on the performance of each method on the (down-sampled) Rosetta-generated dataset of tertiary structures for each of the 8 CASP12 and CASP13 targets listed in Section 2. We report in Table 3 the respective loss of each method by measuring the difference between the GDT-TS value of the top model predicted by each method and the actual native structure.

Specifically, Row 2 ('Best model') in Table 3 shows the GDT-TS of the best model in a dataset. This is the smallest GDT-TS distance over all structures in a dataset from the known native structure for the target. This value serves as a reference, as one

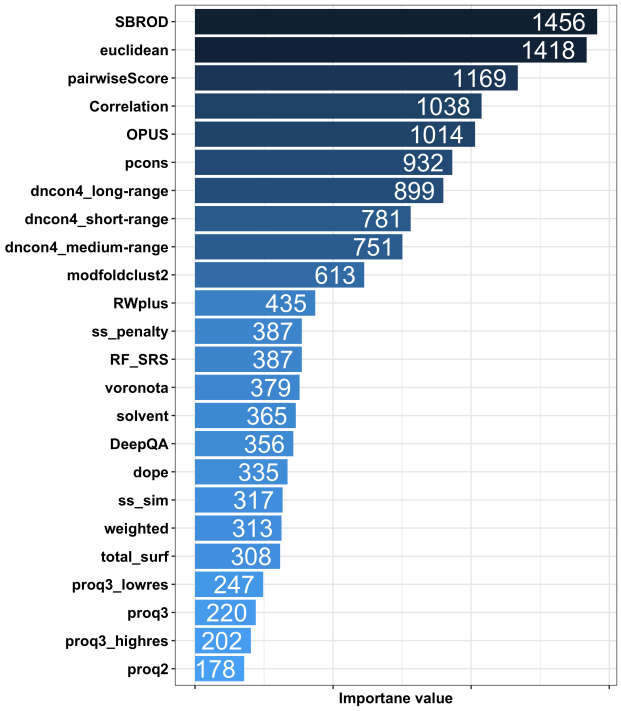

Figure 3: The bar plot shows the feature importance.

cannot outperform it. In the best case, a model assessment method can identify the lowest GDT-TS model. The next two rows in Table 3 provide more information regarding the distribution of GDT-TS distances of structures in a dataset by relating the median and 75th percentile values, respectively. The last three rows relate the GDT-TS values of the top model selected by DeepRank2, the top model selected by DeepRank, and the top model selected by SBROD, respectively.

Table 3 highlights the highest GDT-TS on each target over the three methods under comparison. To obtain an aggregate view of the performance, we calculate a loss value with each prediction (by a method on a target) by subtracting the highest GDT-TS (corresponding to the best model) in the dataset of structures available for a target from the GDT-TS value of the model predicted as the best by a method on the corresponding target. These values are averaged over all targets to obtain an average loss for each method. Considering all the targets, DeepRank2 achieves an average loss of 0.1024, whereas DeepRank's average loss is 0.1107, and SBROD's achieves an average loss of 0.1165. DeepRank2 achieves a lower average loss than DeepRank and SBROD. Overall, DeepRank2 shows the best performance on this dataset.

Figure 4 provides information beyond the top predicted model and shows the GDT-TS distribution over all structures for each target; that is, each down-sampled structure (for each target) has been evaluated based on its GDT-TS distance from a given native structure (for the same target). A red vertical line is drawn to show the median value of the distribution. The green vertical line shows the GDT-TS value of the model predicted as top by DeepRank2. The orange vertical line shows the GDT-TS value of the model predicted as top by DeepRank. The blue vertical line shows the GDT-TS value of the model predicted as top by SBROD. We note that for targets T0886-D1 and T1008-D1 the blue and green lines overlap, because DeepRank2 and SBROD select the same model.

Figure 4 shows that DeepRank2 achieves a better performance for targets T0859-D1, T0886-D1, and T0953s1-D1. For these three targets, the DeepRank2-selected model is significantly better (in quality as measured by GDT-TS) than DeepRank and SBROD, as its GDT-TS value is over the 75th percentile of the distribution. On the rest of the targets, DeepRank and SBROD show better performance than DeepRank2; however, the differences in the GDT-TS values are small.

Figure 5 shows the DeepRank2-selected model for four targets, T0859-D1, T0892-D2, T0953s2-D3, and T1008-D1. In each case, the top-ranked model, drawn in light blue, is superimposed over the corresponding native structure, which is drawn in light yellow. Figure 5 relates the quality of the DeepRank2-selected models by showing that these models are structurally similar to the native structure.

3.2 Evaluation on CASP13 Dataset of Server-Predicted Structures

DeepRank2, DeepRank, and SBROD are now evaluated over tertiary structures submitted by servers in CASP13 for each of the 80 CASP13 targets. DeepRank2 achieves an average loss of 0.0461 on the CASP13 dataset, compared to DeepRank's loss of 0.051 and SBROD's loss of 0.0873. DeepRank2 significantly improves upon DeepRank and SBROD. In addition, DeepRank2 selects the best structural model in 10 targets.

We additionally measure the correlation between the best structural models' GDT-TS values and DeepRank2-selected models' GDT-TS values (over all 80 targets). Figure 6 shows these values and the regression line. The correlation coefficient is 0.96, which indicates a strong performance by DeepRank2. The larger the distance between a data point and the regression line, the bigger the loss (represented by the size of the disk drawing each data point). Most data points are close to the regression line, with only a few disks away from it. The ground truth values are normal distributed (p-value of 0.1525). This is also true of the (DeepRank2-)predictive values (p-value of 0.118).

More detail is provided in Figure 7. The histogram of GDT-TS values values has a long tail, and the distribution's skewness value, measured as shown in Equation (1), is 1.0918. Equation 1, where x_i is a data point, and $\bar{x}$ is the mean in a distribution of n data points, shows that the more data and the smaller the mean value in a distribution, the higher the skewness value. In this distribution, the median value of 0.02850 is much smaller than the mean value of 0.04613, which indicates that most target's loss is close to 0.

$$skewness = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} \quad (1)$$

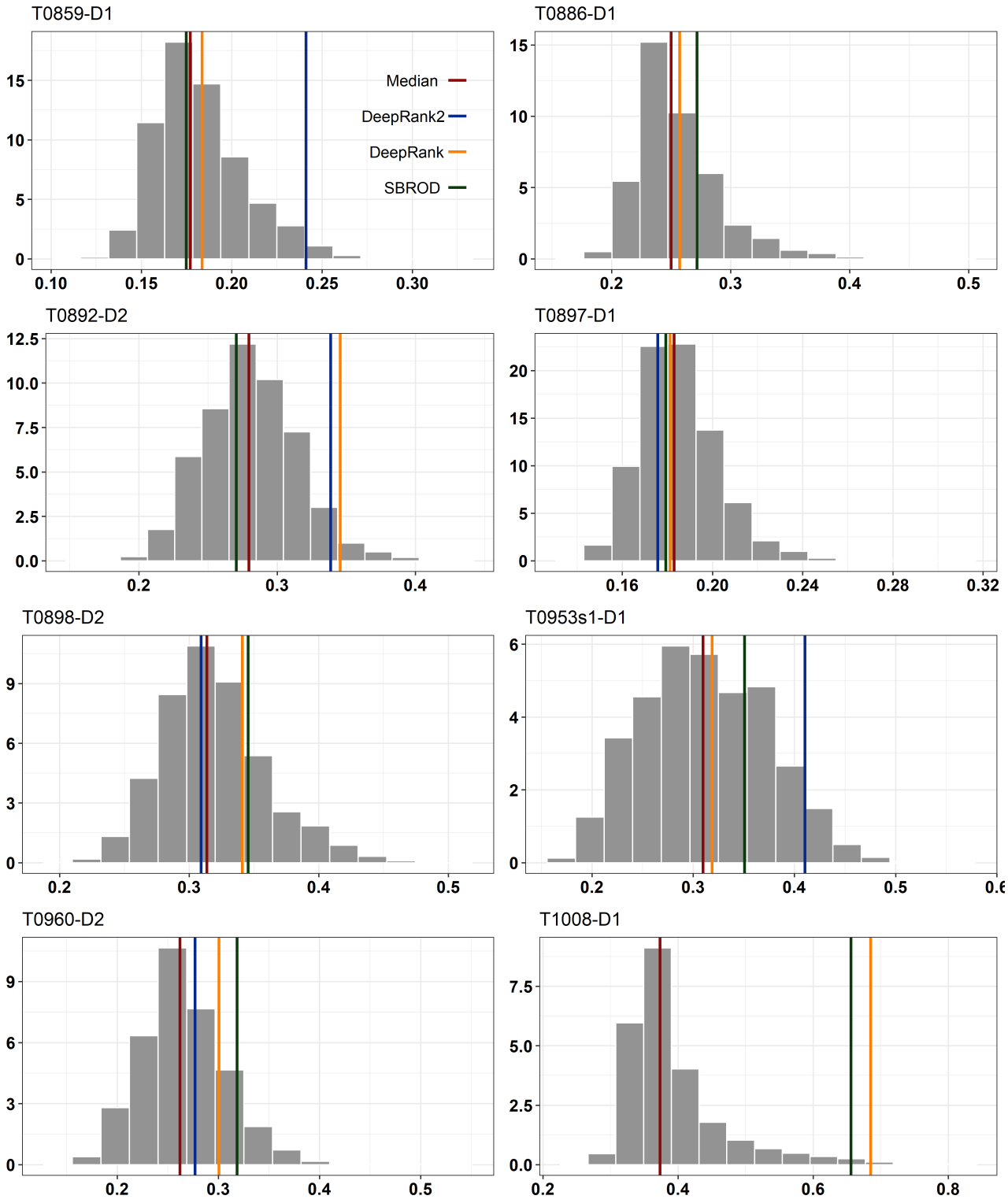

Figure 4: The GDT-TS distribution is shown over the dataset of structures for each target. The green vertical line shows the GDT-TS value of the model predicted as top by DeepRank2. The orange vertical line shows the GDT-TS value of the model predicted as top by DeepRank. The blue vertical line shows the GDT-TS value of the model predicted as top by SBROD.

Table 3: The top three rows relate the GDT-TS distance from a known native structure of the best model (the lowest GDT-TS distance over all structures in a dataset), the median GDT-TS distance over the distribution of distances corresponding to a given dataset of tertiary structures per target, and the 75th percentile value. The last three rows show the GDT-TS distance from the known native structure for the top model predicted by DeepRank2 DeepRank, and SBROD, respectively. If one of these methods achieves a higher value on a target than the other method, this is indicated by highlighting the value in bold font.

#	T0898-D2	T0886-D1	T0892-D2	T0897-D1	T0960-D2	T0953s1-D1	T0859-D1	T1008-D1
Best model	0.4682	0.4312	0.4000	0.2536	0.4256	0.4963	0.2832	0.7403
Median	0.3182	0.2500	0.2818	0.1830	0.2679	0.3172	0.1792	0.3799
75th percentile	0.3455	0.2717	0.3045	0.1938	0.2946	0.3582	0.1969	0.4261
DeepRank2	0.3091	**0.2717**	0.3386	0.1757	0.2768	**0.4104**	**0.2412**	0.6558
DeepRank	0.3409	0.2572	**0.3455**	**0.1812**	0.3006	0.3185	0.1836	**0.6851**
SBROD	**0.3455**	**0.2717**	0.2705	0.1793	**0.3185**	0.3507	0.1748	0.6558

4 CONCLUSION

In this paper, we present a new model assessment method, DeepRank2. The method builds over a deep network architecture, employing a variety of features and feature selection to utilize the most informative features. DeepRank2 is evaluated over diverse CASP targets to identify and predict the top tertiary structure/model from datasets generated in a de-novo setting via the Rosetta AbInitio protocol or via CASP servers. The evaluations show that DeepRank2 performs very well across datasets, indicating its potential and warranting further work in advancing the state of the art in model assessment in protein structure prediction.

5 ACKNOWLEDGMENTS

This work is supported in part by NSF IIS Grants (No. 1763233 and 1763246) and an NSF DBI grant (No. 1759934). Computations were run in part on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA (URL: http://orc.gmu.edu). This material is additionally based upon work supported by (while serving at) the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Cheng J. Adhikari B, Hou J. 2017. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* 34, 9 (2017), 1466–1472.

[2] N. Akhter, W. Qiao, and A. Shehu. 2018. An Energy Landscape Treatment of Decoy Selection in Template-free Protein Structure Prediction. *Computation* 6, 2 (2018), 39.

[3] N. Akhter and A. Shehu. 2018. From Extraction of Local Structures of Protein Energy Landscapes to Improved Decoy Selection in Template-free Protein Structure Prediction. *Molecules* 23, 1 (2018), 216.

[4] Wallner B and Elofsson A. 2003. Can correct protein models be identified? *Protein Sci* 12 (2003), 1073–1086.

[5] H. M. Berman, K. Henrick, and H. Nakamura. 2003. Announcing the worldwide Protein Data Bank. *Nature Structural Biology* 10, 12 (2003), 980–980.

[6] C. E. Blaby-Haas and V. de Crécy-Lagard. 2013. Mining high-throughput experimental data to link gene and function. *Trends Biotechnol* 29, 4 (2013), 174–182.

[7] D. D. Boehr and P. E. Wright. 2008. How do proteins interact? *science* 320, 5882 (2008), 1429–1430.

[8] Renzhi Cao, Debswapna Bhattacharya, Jie Hou, and Jianlin Cheng. 2016. DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinf* 17, 1 (2016), 495.

[9] S Chatterjee, S Ghosh, and S Vishveshwara. 2013. Network properties of decoys and CASP predicted models: a comparison with native protein structures. *Molecular BioSystems* 9, 7 (2013), 1774–1788.

[10] Jianlin Cheng, Myong-Ho Choe, Arne Elofsson, Kun-Sop Han, Jie Hou, Ali Maghrabi, Liam McGuffin, David Menéndez-Hurtado, Kliment Olechnovič, Torsten Schwede, Gabriel Studer, Karolis Uziela, Česlovas Venclovas, and Björn Wallner. 2019. Estimation of model accuracy in CASP13. *Proteins: Structure, Function, and Bioinformatics* 87 (07 2019). https://doi.org/10.1002/prot.25767

[11] R. Das. 2011. Four small puzzles that Rosetta doesn't solve. *PLoS ONE* 6, 5 (2011), e20044.

[12] B.N. Dominy and C.L. Brooks. 2002. Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem* 23 (2002), 147–160.

[13] Arne Elofsson, Keehyoung Joo, Chen Keasar, Jooyoung Lee, Ali HA Maghrabi, Balachandran Manavalan, Liam J McGuffin, David Ménendez Hurtado, Claudio Mirabello, Robert Pilstål, et al. 2018. Methods for estimation of model accuracy in CASP12. *Proteins: Struct, Funct, and Bioinf* 86 (2018), 361–373.

[14] Kandathil Shaun M. Jones David T. Greener, Joe G. 2019. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature Communications* (2019).

[15] Thomas Finley Taifeng Wang Wei Chen Weidong Ma Qiwei Ye Tie-Yan Liu Guolin Ke, Qi Meng. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS, 1466–1472.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *The IEEE International Conference on Computer Vision (ICCV)*.

[17] Zhiquan He, Meshari Alazmi, Jingfen Zhang, and Dong Xu. 2013. Protein structural model selection by combining consensus and single scoring methods. *PLoS ONE* 8, 9 (2013), e74006.

[18] Zhiquan He, Yi Shang, Dong Xu, Yang Xu, and Jingfen Zhang. 2012. Protein structural model selection based on protein-dependent scoring function. *Statistics & Interface* 5, 1 (2012), 109–115.

[19] J. Hou, R. Cao, and J. Cheng. 2019. Deep convolutional neural networks for predicting the quality of single protein structural models. *bioRxiv* (2019), 590620.

[20] J. Hou, T. Wu, R. Cao, and J. Cheng. 2019. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins* 87 (2019), 1165–1178.

[21] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 448–456.

[22] Xiaoyang Jing, Kai Wang, Ruqian Lu, and Qiwen Dong. 2016. Sorting protein decoys by machine-learning-to-rank. *Sci Reports* 6 (2016), 31571.

[23] Mikhail Karasikov, Guillaume Pagès, and Sergei Grudinin. 2018. Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics* 35, 16 (12 2018), 2801–2808. https://doi.org/10.1093/bioinformatics/bty1037 arXiv:https://academic.oup.com/bioinformatics/article-pdf/35/16/2801/29154701/bty1037.pdf

[24] Andriy Kryshtafovych, Alessandro Barbato, Krzysztof Fidelis, Bohdan Monastyrskyy, Torsten Schwede, and Anna Tramontano. 2014. Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins: Struct, Funct, and Bioinf* 82 (2014), 112–126.

[25] Andriy Kryshtafovych, Bohdan Monastyrskyy, Krzysztof Fidelis, Torsten Schwede, and Anna Tramontano. 2018. Assessment of model accuracy estimations in CASP12. *Proteins: Struct, Funct, and Bioinf* 86 (2018), 345–360.

[26] Themis Lazaridis and Martin Karplus. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* 288, 3 (1999), 477–487.

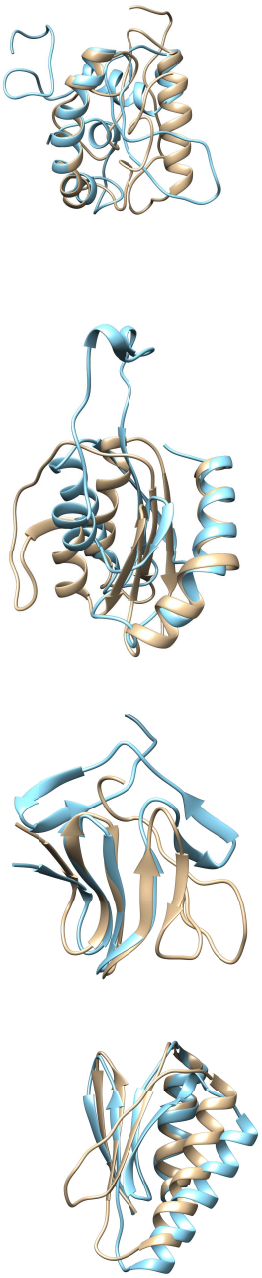

Figure 5: The DeepRank2-selected model, drawn in light blue, is superimposed over the native structure, which is drawn in light yellow. The top panel shows target T0859-D1. The next panel shows T0892-D2, followed by T0953s1-D3, and T1008-D1.

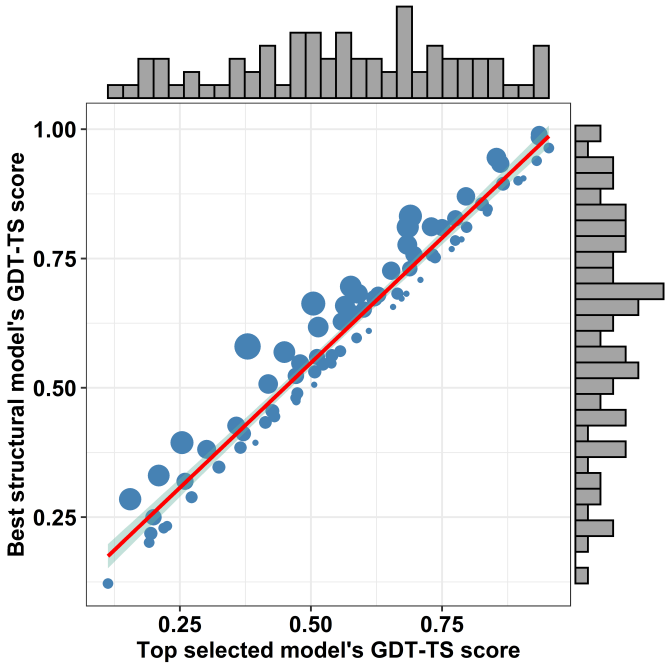

Figure 6: Best structural model's GDT-TS value versus the DeepRank2-selected model's GDT-TS value for each of the 80 CASP13 protein targets.

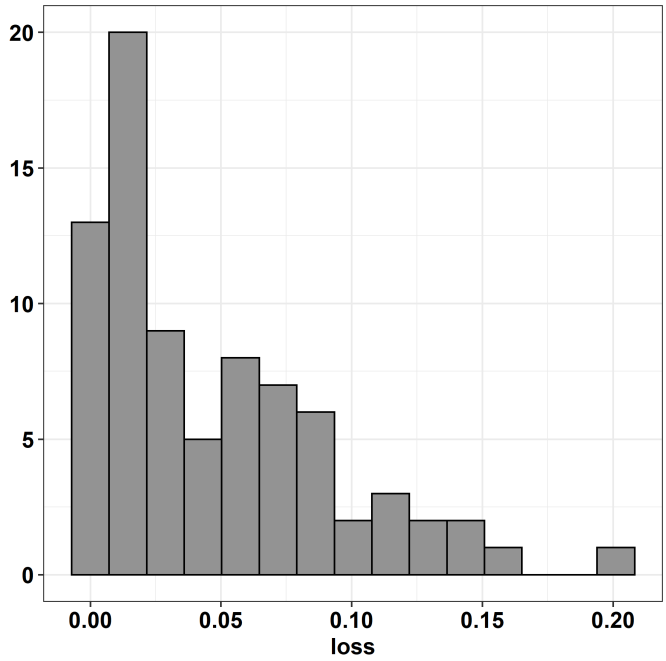

Figure 7: Histogram shows the distribution of loss by DeepRank2 over the 80 CASP13 targets.

[27] A. Leaver-Fay et al. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487 (2011), 545–574.

[28] J. Lee, P. Freddolino, and Y. Zhang. 2017. Ab initio protein structure prediction. In *From Protein Structure to Function with Bioinformatics* (2 ed.), D. J. Rigden (Ed.). Springer London, Chapter 1, 3–35.

[29] H. Lu and J. Skolnick. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* 44 (2001), 223–232.

[30] Balachandran Manavalan and Jooyoung Lee. 2017. SVMQA: support–vector-machine-based protein single-model quality assessment. *Bioinformatics* 33, 16 (2017), 2496–2503.

[31] Balachandran Manavalan, Juyong Lee, and Jooyoung Lee. 2014. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PloS one* 9, 9 (2014), e106542.

[32] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu. 2016. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comp. Biol.* 12, 4 (2016), e1004619.

[33] Brendan J McConkey, Vladimir Sobolev, and Marvin Edelman. 2003. Discrimination of native protein structures using atom–atom contact scoring. *Proc Natl Acad Sci USA* 100, 6 (2003), 3215–3220.

[34] Shokoufeh Mirzaei, Tomer Sidi, Chen Keasar, and Silvia Crivelli. 2016. Purely structural protein scoring functions using support vector machine and ensemble learning. *IEEE/ACM Trans Comp Biol & Bioinf* (2016).

[35] K. Molloy, S. Saleh, and A. Shehu. 2013. Probabilistic Search and Energy Guidance for Biased Decoy Sampling in Ab-initio Protein Structure Prediction. *IEEE/ACM Trans Comput Biol and Bioinf* 10, 5 (2013), 1162–1175.

[36] Son P Nguyen, Yi Shang, and Dong Xu. 2014. DL-PRO: A novel deep learning method for protein model quality assessment. In *Int Conf Neural Networks (IJCNN)*. IEEE, 2071–2078.

[37] B. Olson, K. A. De Jong, and A. Shehu. 2013. Off-Lattice Protein Structure Prediction with Homologous Crossover. In *Conf on Genetic and Evolutionary Computation (GECCO)*. ACM, New York, NY, 287–294.

[38] B. Olson and A. Shehu. 2013. Multi-Objective Stochastic Search for Sampling Local Minima in the Protein Energy Surface. In *ACM Conf on Bioinf and Comp Biol (BCB)*. Washington, D. C., 430–439.

[39] B. Olson and A. Shehu. 2014. Multi-Objective Optimization Techniques for Conformational Sampling in Template-Free Protein Structure Prediction. In *Intl Conf on Bioinf and Comp Biol (BICoB)*. Las Vegas, NV, 143–148.

[40] Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. 2018. Protein model quality assessment using 3D oriented convolutional neural networks. *bioRxiv* (2018), 432146.

[41] Marcin Pawlowski, Lukasz Kozlowski, and Andrzej Kloczkowski. 2016. MQAPsingle: A quasi single-model approach for estimation of the quality of individual protein structure models. *Proteins: Struct, Funct, and Bioinf* 84, 8 (2016), 1021–1028.

[42] N. Perdigao, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, and S. I. O'Donoghue. 2015. Unexpected features of the dark proteome. *Proc Natl Acad Sci USA* 112, 52 (2015), 15898–1590.

[43] Jian Qiu, Will Sheffler, David Baker, and William Stafford Noble. 2008. Ranking predicted protein structures with support vector regression. *Proteins: Struct, Funct, and Bioinf* 71, 3 (2008), 1175–1182.

[44] Arjun Ray, Erik Lindahl, and Björn Wallner. 2012. Improved model quality assessment using ProQ2. *BMC Bioinf* 13, 1 (2012), 224.

[45] Rin Sato and Takashi Ishida. 2019. Protein model accuracy estimation based on local structure quality assessment using 3D convolutional neural network. *PloS one* 14, 9 (2019), e0221347.

[46] Andrew Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577 (01 2020), 706–710. https://doi.org/10.1038/s41586-019-1923-7

[47] A. Shehu. 2013. Probabilistic Search and Optimization for Protein Energy Landscapes. In *Handbook of Computational Molecular Biology*, S. Aluru and A. Singh (Eds.). Chapman & Hall/CRC Computer & Information Science Series.

[48] Karolis Uziela and Björn Wallner. 2016. ProQ2: estimation of model accuracy implemented in Rosetta. *Bioinformatics* 32, 9 (2016), 1411–1413.

[49] Yury N Vorobjev and Jan Hermans. 2001. Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci* 10, 12 (2001), 2498–2506.

[50] Z. Wang, A. N. Tegge, and J. Cheng. 2009. Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins* 75 (2009), 638–647.

[51] T. Wu, Z Guo, and J Cheng. 2019. DNCON4 V1.0. https://github.com/jianlin-cheng/DNCON4_system.

[52] Tianqi Wu, Zhiye Guo, Jie Hou, and Jianlin Cheng. 2020. DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *bioRxiv* (2020). https://doi.org/10.1101/2020.03.17.995910 arXiv:https://www.biorxiv.org/content/early/2020/03/18/2020.03.17.995910.full.pdf

[53] D. Xu and Y. Zhang. 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. 80, 7 (2012), 1715–1735.

[54] A. Zaman and A. Shehu. 2019. Balancing multiple objectives in conformation sampling to control decoy diversity in template-free protein structure prediction. *BMC Bioinformatics* 20, 1 (2019), 211. https://doi.org/10.1186/s12859-019-2794-5

[55] A. Zemla. 2003. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research* 31 (2003), 3370–3374.

[56] Hongyi Zhou and Jeffrey Skolnick. 2011. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J* 101, 8 (2011), 2043–2052.