Improved protein structure refinement guided by deep learning based accuracy estimation

Naozumi Hiranuma[1,2,a], Hahnbeom Park[1,a], Minkyung Baek[1], Ivan Anishchanka[1], Justas Dauparas[1], and David Baker[1,3,*].

1 - Department of Biochemistry and Institute for Protein Design, University of Washington, WA, USA

2 - Paul G. Allen School of Computer Science & Engineering, University of Washington, WA, USA

3 - Howard Hughes Medical Institute, University of Washington, WA, USA

[a] *N.H and H.P contributed equally to this work.*

Correspondence to: *dabaker@uw.edu

Abstract

We develop a deep learning framework (DeepAccNet) that estimates per-residue accuracy and residue-residue distance signed error in protein models and uses these predictions to guide Rosetta protein structure refinement. The network uses 3D convolutions to evaluate local atomic environments followed by 2D convolutions to provide their global contexts and outperforms other methods that similarly predict the accuracy of protein structure models. Overall accuracy predictions for X-ray and cryoEM structures in the PDB correlate with their resolution, and the network should be broadly useful for assessing the accuracy of both predicted structure models and experimentally determined structures and identifying specific regions likely to be in error. Incorporation of the accuracy predictions at multiple stages in the Rosetta refinement protocol considerably increased the accuracy of the resulting protein structure models, illustrating how deep learning can improve search for global energy minima of biomolecules.

Introduction

Distance prediction through deep learning on amino acid co-evolution data has considerably advanced protein structure prediction [1–3]. However, in most cases, the predicted structures still deviate considerably from the actual structure [4]. The protein structure refinement challenge is to increase the accuracy of such starting models. To date, the most successful approaches have been with physically based methods that involve a large-scale search for low energy structures, for example with Rosetta [5] and/or molecular dynamics [6]. This is because any available homology and co-evolutionary information are typically already used in the generation of the starting models.

The major challenge in refinement is sampling; the space of possible structures that must be searched through even in the vicinity of a starting model is extremely large [5,7]. If it were possible to accurately identify what parts of an input protein model were most likely to be in error, and how these regions should be altered, it should be possible to considerably improve the search through structure space and hence the overall refinement process. Many methods

for estimation of model accuracy (EMA) have been described, including approaches based on deep-learning such as ProQ3D (based on per-residue Rosetta energy terms and multiple sequence alignments with multi-layer perceptrons [8]), and Ornate (based on 3D voxel atomic representations with 3D convolutional networks [9]). Non-deep-learning methods such as VoroMQA compare a Voronoi tessellation representation of atomic interactions against pre-collected statistics [10]. These methods focus on predicting per-residue accuracy. Few studies have sought to guide refinement using deep learning based accuracy predictions [11]; the most successful refinement protocols in the recent blind 13th Critical Assessment of Structure Prediction (CASP13) test either utilized very simple ensemble-based error estimations [5] or none at all [12]. This is likely because of the low specificity of most current accuracy prediction methods, which only predict which residues are likely to be inaccurately modeled, but not how they should be moved, and hence are less useful for guiding search.

Results

We set out to develop a deep learning based framework (DeepAccNet) that estimates the signed error in every residue-residue distance along with the local residue contact error, and we use this estimation to guide Rosetta based protein structure refinement. Our approach is schematically outlined in Figure 1.

Development of improved model accuracy predictor

We first sought to develop model accuracy predictors that provide both global and local information for guiding structure refinement. We developed network architectures that make the following three types of predictions given a protein structure model: local measures of structure accuracy measured by per residue C_β local distance difference test (l-DDT) scores [13], a native C_β contact map thresholded at 15 Å (referred to as *mask*), and per residue-pair distributions of signed C_β-C_β distance error against corresponding native structures (referred to as *estograms*; histogram of errors); $C\alpha$ is taken for GLY. Rather than predicting single error values for each pair of positions, we instead predict histograms of errors (analogous to the distance histograms employed in the structure prediction networks of [1–3]), which provide more detailed information about the distributions of possible structures and better represent the uncertainties inherent to error prediction. Networks were trained on alternative structures ("decoys") with model quality ranging from 50% to 90% in GDT-TS (global distance test - tertiary structure) [14] generated by homology modeling [15], trRosetta [1], and native structure perturbation (see Methods). ~150 decoy structures were generated for each of 7,314 X-ray crystal structures with resolution better than 2.5 Å lacking extensive crystal contacts and having sequence identity less than 40% to any of 73 refinement benchmark set proteins (see below). Of the approximately one million decoys, those for 280 and 285 of the 7,314 proteins were held out for validation and testing, respectively. More details of the training/test set and decoy structure generation can be found in Methods.

The predictions are based on 1D, 2D, and 3D features that reflect accuracy at different levels. Defects in high-resolution atomic packing are captured by 3D convolution operations performed on 3D atomic grids around each residue defined in a rotationally invariant local frame, similar to the Ornate method [9]. 2D features are defined for all residue pairs, and they include Rosetta inter-residue interaction terms, which further report on the details of the interatomic interactions, while residue-residue distance and angular orientation features provide lower resolution structural information. Multiple sequence alignment (MSA) information in the form of inter-residue distance prediction by the trRosetta[1] network and sequence embeddings from the ProtBert-BFD100 model [16] (or Bert, in short) are also optionally provided as 2D features. At the 1D per residue level, the features are the amino acid sequence, backbone torsion angles, and the Rosetta intra-residue energy terms (see Methods for details).

We implemented a deep neural network, DeepAccNet, that incorporates these 1D, 2D, and 3D features (Figure 1A). The networks first perform a series of 3D convolution operations on local atomic grids in coordinate frames centered on each residue. These convolutions generate features describing the local 3D environments of each of the N residues in the protein. These, together with additional residue level 1D input features (e.g. local torsional angles and individual residue energies), are combined with the 2D residue-residue input features by tiling (so that associated with each pair of residues there are both the input 2D features for that pair and the 1D features for both individual residues), and the resulting combined 2D feature description is input to a series of 2D convolutional layers using the ResNet architecture [17]. A notable advantage of our approach of tying together local 3D residue based atomic coordinate frames through a 2D distance map is the ability to integrate full atomic coordinate information in a rotationally invariant way; in contrast, a Cartesian representation of the full atomic coordinates would change upon rotation, substantially complicating network for both training and its use. Details of the network architecture, feature generation, and training processes are found in Methods.

Figure 2 shows examples of the predictions of DeepAccNet without MSA or Bert embeddings (referred to as *DeepAccNet-Standard*) on two randomly selected decoy structures for each of three target proteins (3lhnA, 4gmqA, and 3hixA) not included in training. In each case, the network generates different signed residue-residue distance error maps for the two decoys that qualitatively resemble the actual patterns of the structural errors (rows of Figure 2). The network also accurately predicts the variations in per residue model accuracy (l-DDT scores) for the different decoys. The left sample from 4gmqA (second row) is closer to the native structure than the other samples are, and the network correctly predicts the location of the smaller distance errors and l-DDT scores closer to 1. Overall, while the detailed predictions are not pixel-perfect, they provide considerable information on what parts of the structure need to move and in what ways to guide refinement. Predictions from the variants with the MSA (referred to as *DeepAccNet-MSA*) and Bert features (referred to as *DeepAccNet-Bert*) are visualized in Figure S1.

We compared the performance of the DeepAccNet networks to that of a baseline network trained only on residue-residue C_β distances. The performance of the DeepAccNet networks are considerably better on average for almost all the test set proteins (Figure S2A; Figure 3); they outperform the baseline C_β distance model in predicting estograms for residue pairs across different sequence separations and input distances (Figure S2B). The addition of the MSA or Bert information improves overall accuracy particularly for quite inaccurate models and residues (Figure S2CD). For all networks and the distance-only network, l-DDT score prediction does not decline substantially with increasing size (Spearman correlation coefficient, or Spearman-r, of -0.04 with p-value > 0.05 for protein size vs. DeepAccNet-Standard performance), but estogram prediction performance clearly declines for larger proteins (Spearman-r of 0.57 with p-value < 0.00001) (Figure S2E) -- for larger proteins with more interactions over long distances, estimating the direction and magnitude of errors is a much harder task while since l-DDT scores only consider local changes at short distances, they degrade less with increasing size.

In addition to distance map features, DeepAccNet networks take as input a) amino acid identities and properties, b) local atomic 3D environments for each residue, c) backbone torsion angles and residue-residue orientations, e) Rosetta energy terms, f) secondary structure information, g) MSA, and h) Bert information. To investigate the contributions of each of these features to network performance, we combined each with distance maps one at a time during training and evaluated performance through estogram cross-entropy loss and l-DDT score mean squared error on test sets (Figure 3A, Table S1). Apart from the MSA features, the largest contributions were from the 3D convolution-based features and the Bert embeddings (compare (v), (vi), and (vii)). There is a statistically significant difference between the network (ii) and (vii), suggesting that the features other than 3D-convolution and Bert help them glue together (p-value < 0.0001 with Wilcoxon signed-rank test for estogram loss between network (ii) and (vii)).

An effective accuracy prediction method should be useful for evaluating and identifying potential errors in experimentally determined structures as well as computational models. We investigated the performance of the network on experimental structures determined by X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and electron microscopy (EM) that were not included in the training set (details of the dataset can be found in Methods). The predicted C_β l-DDT values by the DeepAccNet variants are close to 1.0 for high-resolution crystal structures, as expected for nearly error free protein structures, and decreases for lower resolution structures (Figure 3C, left panel for DeepAccNet-Standard, Figure S3 for other variants). A similar correlation between predicted accuracy and resolution holds for X-ray structures of membrane proteins (Figure 3C, middle panel; Spearman-r 0.64 with p-value < 0.0001) and cryoEM structures (Figure 3C, right panel; Spearman-r 0.87 with p-value < 0.0001). Note that the good correlation found within the membrane proteins can be simply due to the difference in core packing; whether the network is aware of the membrane environment is unclear from the result. A list of X-ray structures with low predicted l-DDT despite their high experimental resolution is provided in

Table S2. Many of these are heme-proteins; as the network does not consider bound ligands, the regions surrounding them are detected as atypical for folded proteins, suggesting that the network may also be useful for predicting cofactor binding and other functional sites from apo-structures. NMR structures have lower predicted accuracies than high-resolution crystal structures (Figure 3D, right; Figure S3CD), which is not surprising given i) they were not included in the training set and ii) they represent solution averages rather than crystalline states. Despite their differences in structural aspects, it will be an interesting direction to train an accuracy network including NMR structures in the future.

We compared the DeepAccNet variants to other accuracy estimators (Figure 3B). As is clear from recent CASP experiments, co-evolution information derived from multiple sequence alignments provides detailed structure information; we include this as an optional input to our network (DeepAccNet-MSA) for two reasons: first, all available homology and co-evolutionary information is typically already used in generating the input models for protein structure refinement and second, in applications such as *de novo* protein design model evaluation, no evolutionary multiple sequence alignment information exists. DeepAccNet-Bert includes the Bert embeddings which are generated with a single sequence without any evolutionary alignments. We compared the performance of the DeepAccNet variants on the CASP13 EMA data (76 targets with approximately 150 decoy models each) to that of the methods that similarly estimate error from a single structure model. These are Ornate (group name 3DCNN) [9], a method from Lamoureux Lab [18], VoroMQA [10], ProQ3 [19], ProQ3D, ProQ3D-lDDT [8], and MODFOLD7 [20]; the former two use 3D convolutions similar to those used in our single residue environment feature calculations. We calculated (i) the Spearman-r of predicted and actual global l-DDT scores per target protein and (ii) area under receiver operator characteristic (ROC) curve for predicting mis-modeled residues per sample (C_β l-DDT< 0.6 [21]) which assesses global and local model accuracy estimation respectively. According to both metrics, DeepAccNet-Standard and DeepAccNet-Bert outperformed the other methods that do not use any evolutionary information; DeepAccNet-MSA also outperformed the other methods that use evolutionary multiple sequence alignment information (Figure 3B right). While this improved performance is very encouraging, it must be noted that our predictions are made after rather than before CASP13 data release so the comparison is not entirely fair: future blind accuracy prediction experiments will be necessary to compare methods on an absolutely even footing. As a step in this direction, we tested performance on structures released from the PDB after our network architecture was finalized that were collected in the CAMEO (Continuous Automated Model EvaluatiOn) [21] experiment between 2/22/2020 to 5/16/2020. We consistently observed that DeepAccNet-Standard and DeepAccNet-Bert improved on other methods that do not use evolutionary information, -- namely, VoroMQA [10], QMean3 [22], and Equant 2 [23] in both global (entire model) and local (per residue) accuracy prediction performance (Figure S4). DeepAccNet-MSA also showed the state of the art performance among the methods that use multiple sequence alignment. We could not compare signed residue-pair distance error predictions because this is not predicted by the other methods.

Guiding search in protein structure refinement using the accuracy predictor

We next experimented with incorporation of the network accuracy predictions into the Rosetta refinement protocol [5,24], which was one of the top methods tested in CASP13 [25]. Rosetta high-resolution refinement starts with a single model, and in a first diversification stage explores the energy landscape around it using a set of sampling operators, and then in a subsequent iterative intensification stage hones in on the lowest energy regions of the space. Search is controlled by an evolutionary algorithm which maintains a diverse but low energy pool through many iterations/generations. With improvements in the Rosetta energy function in the last several years [26,27], the bottleneck to improving refinement has largely become sampling close to the correct structure. The original protocol utilized model consensus-based accuracy estimations (i.e. regional accuracy estimated as inverse of fluctuation within an ensemble of structures sampled around the input model) to keep search focused in the relevant region of the space -- these have the obvious downside of limiting exploration in regions which need to change substantially from the input model but are located in deep false local energy minima.

To guide search, estograms and l-DDT scores were predicted and incorporated at every iteration in the Rosetta refinement protocol at three levels (details in Methods). First and most importantly, the estograms were converted to residue-residue interaction potentials with weight for each pair defined by a function of its estogram prediction confidence, and these potentials were added to the Rosetta energy function as restraints to guide sampling. Second, the per-residue l-DDT predictions were used to decide which regions to intensively sample or to recombine with other models. Third, global l-DDT prediction was used as the objective function during the selection stages of the evolutionary algorithm and to control the model diversity in the pool during iteration.

To benchmark the accuracy prediction guided refinement protocol, 73 protein refinement targets were collected from previous studies [5,24]. The starting structures were generally the best models available from automated structure prediction methods. A separate 7 targets from Park et al [5,24] were used to tune the restraint parameters and were excluded from the benchmarking in this study.

We found that network-based accuracy prediction consistently improves refinement across the benchmark examples. In Figure 4, refinement guided by the accuracy predictions from DeepAccNet-Standard is compared to our previous protocol in which simpler non-deep learning accuracy estimation was used. Refinement of many proteins in the benchmark set was previously quite challenging due to their size [24]; however, with the new protocol, consistent improvements are observed over the starting models regardless of protein size (Figure 4A, the l-DDT improve by 10% on average) and over the models produced with our previous unguided search (Figure 4B; the l-DDT improves by 4% on average). The number

of targets with l-DDT improvements of greater than 10% increases from 27% to 47% using DeepAccNet-Standard to guide refinement. These improvements are quite notable given how challenging the protein structure refinement problem is (comparison to other best predictors on the latest CASP targets is shown in Figure S8); for reference best improvements between successive biannual CASP challenges are typically < 2% [25]. Tracing back through the refinement trajectory reveals that the progress in both predicted and actual model quality occurs gradually through the stages and are quite well correlated to each other (Figure S5A). Predictions of more detailed per residue model quality also agree well with their actual values (Figure 4E).

We evaluated the practical impact of the improvement in refined model quality using the accuracy predictions by carrying out molecular replacement (MR) trials with experimental diffraction datasets (Fig 3C). On 41 X-ray data sets brought from the benchmark set, the fraction of cases for which robust MR hits were obtained was 0%, 20%, and 37% using pre-refined models, models refined by the non-deep learning protocol, and models refined using DeepAccNet-Standard, respectively.

Residue-pair restraints derived from the DeepAccNet estogram predictions were crucial for the successful refinement (Figure 4D and Figure S6A). When only residue-wise and global accuracy predictions (either from DeepAccNet or external EMA tool[10]) were utilized for the refinement calculations, performance did not statistically differ from our previous work ($P>0.1$). When Bert or MSA input was further provided to DeepAccNet (red bars in Figure 4D), significant increases in model quality was observed for a number of targets (Figure S6B). Final pool model quality analyses (Figure S7) suggest that sampling was improved by those extra inputs (i.e. overall model quality increases) while the single model selection was generally reasonable across the three different network-based-EMAs.

The model accuracy improvements occur across a broad range of protein sizes, starting model qualities, and types of errors. Refinement improved models across various secondary structures to similar extents and corrected secondary structures originally modeled incorrectly, increasing model secondary structure accuracy by almost 10% based on an 8-state definition [28] (Figure S5B, C). As shown in Figure 4F, improvements involve identification and modifications of erroneous regions when the overall structure is correct (TR776) as well as overall concerted movements when the core part of the model is somewhat inaccurate (5m1mA). The accuracy prediction network promotes this overall improvement in two ways: first, it provides a more accurate estimation of unreliable distance pairs and regions at every iteration of refinement for every model on which sampling can be focused, and second, it provides a means to effectively constrain the search space in the already accurately modeled regions through residue-residue pair restraints -- this is particularly important for refinement of large proteins. The network enables the refinement protocol to adjust how widely to search on a case by case basis; this is an advantage over most previous refinement approaches where search has generally been either too conservative or too aggressive [29].

DeepAccNet is available at https://github.com/hiranumn/DeepAccNet.

Discussion

Representations of the input data are critical for the success of deep learning approaches. In the case of proteins, the most complete description is the full Cartesian coordinates of all of the atoms, but these are transformed by rotation and hence not optimal for predicting rotationally invariant quantities such as error metrics. Hence most previous machine learning based accuracy prediction methods have not used the full atomic coordinates [8,10,19]. The previously described Ornate method does use atomic coordinates to predict accuracy, and solves the rotation dependence by setting up local reference frames for each residue. As in the Ornate method, DeepAccNet carries out 3D convolutions over atomic coordinates in residue centered frames, but we go beyond Ornate by integrating together this detailed residue information along with additional individual residue and residue-residue level geometric and energetic information by 2D convolutions over the full N x N residue-residue distance map. DeepAccNet-Bert further employs the sequence embeddings from the ProtBert language model [16], which provides a higher level representation of the amino acid sequence more directly relatable to 3D structures.

Evaluation of performance on CASP and CAMEO datasets shows that the DeepAccNet networks make state-of-the-art accuracy predictions, and they are the first to our knowledge to predict signed distance errors for protein structure refinement. Model quality estimations on X-ray crystal structures correlate with resolutions, and the network should also be useful in identifying errors in experimentally determined structures (Figure 3C). DeepAccNet performs well on both cryoEM and membrane protein structures, and it could be particularly useful for low-resolution structure determination and modeling of currently unsolved membrane proteins (Figure 3C). We also anticipate that the network will be useful in evaluating protein design models.

Guiding search using the network predictions improved Rosetta protein structure refinement over a wide range of protein sizes and starting model qualities (Figure 4). However, there is still considerable room for improvement in the combined method. To more effectively use the information in the accuracy predictions it will be useful to explore sampling strategies which can better utilize the network predictions and more frequent communication between Rosetta modeling and the accuracy prediction network -- the network is fast enough to evaluate the accuracy of many models more frequently. Also, we find that DeepAccNet often overestimates the quality of models when those are heavily optimized by the network through our refinement protocol (Figure S5A); adversarial training could help reduce this problem and allow more extensive refinement. It is clear that there is also considerably more to explore in using deep learning to guide refinement. For example, selection of which of the current sampling operators to use in a given situation, and the development of new sampling operators using generative models such as sampling missing regions by inpainting. More

generally, reinforcement learning approaches should help identify more sophisticated iterative search strategies.

Methods

Data preparation.

Training and test sets for protein model structures (often called decoys) are generated to most resemble starting models of real-case refinement problems. We reasoned that a relevant decoy structure should meet the following conditions: i) has template(s) not too far or close in sequence space; ii) does not have strong contacts to other protein chains, iii) should contain minimal fluctuating (i.e. missing density) regions. To this end, we picked a set of crystal structures from the PISCES server (deposited by May 1, 2018) containing 20,399 PDB entries with maximum sequence redundancy of 40% and minimum resolution of 2.5 Å. We further trimmed the list to 8,718 chains by limiting their size to 50-300 residues and requiring that proteins are either monomeric or have minimal interaction with other chains (weaker than 1 kcal/mol per residue in Rosetta energy). HHsearch [30] was used to search for templates; 50 templates with the highest HHsearch probability, sequence identity of at most 40% and sequence coverage of at least 50% are selected for model generation.

Decoy structures are generated using three methods: comparative modeling, native structure perturbation, and deep learning guided folding. Comparative modeling and native structure perturbation are done using RosettaCM [15]. For comparative modeling of each protein chain we repeated RosettaCM 500 times in total, every time randomly selecting a single template from the list. In order to increase the coverage of decoy structures at mid-to-high accuracy regime for targets lacking templates with GDT-TS > 50, 500 models are further generated providing a single template and 40% trimmed native structure as templates. Sampled decoy set for a protein chain is included in training/test data only if the total number of decoys at medium accuracy (GDT-TS to native ranging from 50 to 90) is larger than 50. Maximum 15 lowest scoring decoys at each GDT-TS bin (ranging from 50 to 90 with bin size 10) are collected, then the rest with lowest energy values are filled so as to make the set contain approximately 90 decoys. Native structures are perturbed to generate high-accuracy decoys. 30 models were generated by RosettaCM either by i) combining a partial model of a native structure with high-accuracy templates (GDT-TS > 90) or ii) inserting fragments at random positions of the native structure. Deep learning guided folding is done using trRosetta [1]. For each protein, 5 subsampled multiple sequence alignments (MSAs) are generated with various depths (i.e. number of sequences in MSA) ranging from 1 to maximum available. The standard trRosetta modeling is run 45 times for each of the subsampled MSAs. The final decoy set collected, consisting of about 150 structures (90 from comparative modeling, 30 from native perturbation, and 30 from deep learning guided folding) per each of 7,314 protein chains (6,749, 280, 285 for training, validation and test datasets), are thoroughly relaxed by Rosetta dual-relax [31] prior to the usage. The distribution of the starting l-DDT values of the test proteins are shown in Figure S9.

Model architectures and input features.

In our framework, convolution operations are performed in several dimensions, and different classes of features come in at different entry points of the network (Figure 1). Here, we briefly describe the network architecture as well as classes of features. More detailed descriptions about the features and model parameters are listed in Table S3 and S4.

The first set of input features to the network are voxelized Cartesian coordinates of atoms per residue, generated in a manner similar to Ornate [9]. Voxelization is performed individually for every residue in the corresponding local coordinate frame defined by backbone N, Cα, and C atoms. Such representation is translationally and rotationally invariant because projections onto local frames are independent of the global position of the protein structure in 3D space. The second set of inputs are per residue 1D features (e.g., amino acid sequence and properties, backbone angles, Rosetta intra-residue energy terms, and secondary structures) and per residue pair 2D features (e.g. residue-residue distances and orientations, Rosetta inter-residue energy terms, inter-residue distance predictions from the trRosetta network [1], and the ProtBert-BFD100 embeddings [16]).

In the first part of the neural network, the voxelized atomic coordinates go through a series of 3D convolution layers whose parameters are shared across residues. The resulting output tensor is flattened so that it becomes a 1D vector per residue, which is concatenated to other 1D features. The second part of the network matches the dimensionality of the features and performs a series of 2D convolution operations. Let us now denote that there are n residues, f_1 1D features, and f_2 2D features. Then, the input matrix of the 1D features M_1 has the shape of n by f_1, and the input matrix of the 2D features M_2 has the shape of n by n by f_2. We tile M_1 in the first and second axis of M_2, concatenating them to produce a feature matrix of size n by n by $2f_1+f_2$. The third axis of the resulting matrix represents vectors of size $2f_1+f_2$, which contain the 2d features and 1D features of i-th and j-th residues. This data representation allows us to convolve over both backbone chain and pairwise interactions.

The concatenated feature matrix goes through a residual network with 20 residual blocks, with cycling dilation rates of 1, 2, 4, and 8 (see Tables S4). Then, the network branches off to two arms of 4 residual blocks. These arms separately predict distributions of C_β distance errors for all pairs of residues (referred to as estograms) and whether a particular residue pair is within 15 Å in a corresponding native structure (referred to as masks). Estograms are defined over categorical distributions with 15 binned distance ranges; the boundary of bins are at -20.0Å, -15.0Å, -10.0Å, -4.0Å, -2.0Å, -1.0Å, -0.5Å, 0.5Å, 1.0Å, 2.0Å, 4.0Å, 10.0Å, 15.0Å, 20.0Å.

In the standard calculation of a C_β l-DDT score of *i*-th residue of a model structure, all pairs of C_β atoms that include the *i*-th residue and are less than 15Å in a reference structure are examined. 0.5Å, 1.0Å, 2.0Å, and 4.0Å cutoffs are used to determine the fractions of preserved C_β distances across the set of pairs. The final C_β l-DDT score is calculated by computing the arithmetic mean of all fractional values [13].

In our setup, we obviously do not have access to reference native structures. Instead, a C_β l-DDT score of *i*-th residue is predicted by combining the probabilistic predictions of estograms and masks as follows:

$$per\_residue\_LDDT = 0.25 * (\bar{p}_0 + \bar{p}_1 + \bar{p}_2 + \bar{p}_3)/\bar{p}_4$$

$\bar{p}_0$ is the mean of probability that the magnitude of C_β distance errors are less than 0.5Å, across all residue pairs that have *i*-th residue involved and predicted to be less than 15Å in its corresponding native structure. The former C_β distance errors are obtained from estogram predictions and the latter native distance information are directly obtained from mask predictions. $\bar{p}_1 ... \bar{p}_3$ are similar quantities with different cutoffs for errors; 1.0Å, 2.0Å, and 4.0Å, respectively. $\bar{p}_4$ is the mean probability that native distance is within 15Å and it is again directly obtained from mask predictions.

The network was trained to minimize categorical cross-entropy between true and predicted estograms and masks. Additionally, as noted, we calculated C_β l-DDT scores based on estograms and masks, and we used a small amount of mean squared loss between predicted and true scores as an auxiliary loss. The following weights on the three loss terms are used.

$$global\_loss = estogram\_loss + 10.0 * LDDT\_loss + 0.25 * mask\_loss$$

The weights are tuned so that the highest loss generally comes from $estogram\_loss$ since estograms are the richest source of information for the downstream refinement tasks. At each step of training, we selected a single decoy from decoy sets of a randomly chosen training protein without replacement. The decoy sets include native structures, in which case the target estograms ask networks to not modify any distance pairs. An epoch consists of a full cycle through training proteins, and the training processes usually converge after 100 epochs. Our predictions are generated by taking an ensemble of four models in the same training trajectory with best validation performance. We used an ADAM optimizer with a learning rate of 0.0005 and decay rate of 0.98 per epoch. Training and evaluation of the networks was performed on RTX2080 gpus.

Analyzing the importance of features.

Feature importance analysis was conducted to understand and quantify the contributions from different classes of features to accurately predicting accuracy of model structures. To do this, we combined each feature class with a distance map one at a time during training (or removed them in one particular case) and analyzed loss of predictions on a held-out test protein set. In addition to the DeepAccNet-Standard, -Bert, and -MSA, we trained 8 types of networks: i) distance map only, ii) distance with local atomic environments scanned with 3D convolution, iii) distance with Bert embeddings, iv) ii and iii combined, v) distance with

Rosetta energy terms, vi) distance with amino acid identities and their properties, vii) distance with secondary structure information, and iv) distance with backbone angles and residue-residue orientations. For each network, we took an ensemble of four models with best validation performance from the same trajectory in order to reduce noise.

We are aware that more sophisticated feature attribution methods for deep networks exist [32]; however, these methods attribute importance scores to features per output per sample. Since we have approximately a quarter million outputs and near million inputs with a typical 150 residue protein, these methods were not computationally feasible and tractable to analyze.

Comparing with other model accuracy estimation methods.
For the CASP13 datasets, we downloaded submissions of QA139_2 (ProQ3D), QA360_2 (ProQ3D-lDDT [8]), QA187_2 (ProQ3 [19]), QA067_2 (LamoureuxLab [18]), QA030_2 (VoroMQA-B [10]), QA275_2 (MODFOLD7), QA359_2 (Ornate, group name 3DCNN [9]) for the accuracy estimation category. The former five methods submitted their predictions for 76 common targets, whereas the last method, Ornate, only submitted for 55 targets. Thus, we decided to analyze predictions on the 76 common targets from all methods except for Ornate, which was only evaluated on 55 targets. An evaluation was performed in two metrics; i) Spearman-r of predicted quality scores across decoys of each target, and ii) area under ROC curve for predicting mis-modeled residues of each sample (C_β l-DDT< 0.6). The latter metric is one of the official CAMEO metrics for local accuracy evaluation. Samples whose residues are all below or above 0.6 C_β l-DDT are omitted. For assessing the performance of methods other than ours, their submitted estimations of global quality scores were evaluated against the true full-atom global l-DDT scores.

For the CAMEO datasets, we downloaded the QA datasets registered between 2/22/2020 to 5/16/2020. This corresponds to 206 targets with approximately 10 modeled structures on average. We downloaded submissions of "Baseline potential", EQuant 2, ModFOLD4, ModFOLD6, ModFOLD7_LDDT, ProQ2, ProQ3, ProQ3D, ProQ3D_LDDT, QMEAN3, QMEANDisco3, VoroMQA_sw5, and VoroMQA_v2. Some methods did not submit their predictions for all samples, and those missing predictions are omitted from the analysis.

Visualizing predictions.
Figure 2 visualizes true and predicted estograms per pair of residues. The images are generated by calculating the expected values of estograms by taking weighted sums of central error values from all bins. For the two bins that encode for errors larger than 20.0 Å and smaller than -20.0 Å, we define the central distance at their boundaries of 20.0 Å and -20.0 Å.

Native structure dataset
Native structures that were not used for model training and validation, monomeric, larger than 40 residues, and smaller than 300 residues for the X-ray and NMR structures, and smaller than 600 residues for EM structures were downloaded from the PDB. For Figure 3C, samples

with a resolution larger than 4Å and 5Å are ignored for the X-ray and EM structures, respectively. The histograms in Figure 3D are using all samples. In total, 23,672 X-ray structures, 88 EM structures, and 2,154 NMR structures are in the histograms. For NMR structures, regions highly varying across the models were trimmed. Structures were discarded if the number of remaining residues after trimming was less than 40 residues or half of the original chain length.

Dataset for refinement runs.

We took 73 proteins and their starting models from our previous work [5] with a few modifications as described below. Of the entire 84 targets used in our previous work, 7 small-sized targets (4zv5A, 5azxA, 5ghaE, 5i2qA, 5xgaA, TR569, T0743) are excluded from the benchmark set and were used for restraint parameter search. 8 additional targets (2n12A, 4idiA, 4z3uA, 5aozA, 5fidA, T0540, TR696, TR857) are excluded after more careful visual inspections as those had potential issues in their native structures (e.g. having contacts with ligands or other chains in crystal structures). 4 new targets were added from previous CASP refinement categories that were not included in the original set (TR747, TR750, TR776, TR884). Model accuracy is evaluated on a subset of ordered residues by trimming less confident residues according to the CASP standard evaluation criteria [25].

Refinement protocol.

Refinement protocol tested in this work inherits the framework from previous study [5]. The overall architecture consists of two stages (Figure 1B): first initial model diversification stage, followed by iterative model intensification stages where a pool of structures is maintained during optimization by an evolutionary algorithm. At the diversification stage, following accuracy estimation of the single starting model, two thousands of independent Rosetta modeling are attempted using RosettaCM [15]. In the iterative annealing stage, series of accuracy estimation, new structure generation, and pool selection steps are repeated iteratively. At each iteration, 10 model structures are selected from the current pool, then individual accuracy predictions are made for each of 10 structures in order to guide the generation of 12 new model structures starting from each (total 120). New pool with size of 50 is selected among 50 previous pool members plus 120 newly generated ones with criteria of i) the highest global l-DDT estimated and ii) model diversity within the pool. This process is repeated for 50 iterations. At every fifth iteration, a *recombination iteration* is called instead of a regular iteration where model structures are recombined with another member in the pool according to the residue l-DDT values predicted by the network (see below).

For modeling of a single structure at both diversification and intensification stages, first *unreliable regions* in the structure are estimated from accuracy prediction (see below). Structural information is removed in those regions and fully reconstructed from scratch. Fragment insertions are carried out in a coarse-grained broken-chain representation of the structure [15] focusing more on unreliable regions (5 times more frequently with respect to the rest part), followed by repeated side-chain rebuilding and minimization [31] in all-atom representation. Both coarse-grained and all-atom stage modeling are guided by *distance*

restraints derived from accuracy predictions in addition to Rosetta energy. Details of unreliable region predictions, recombination iteration, and restraints are reported in the following sections.

Unreliable region prediction

Accuracy values predicted from the network are used to identify unreliable regions. We noticed that the l-DDT metric has a preference for helical regions (as local contacts are almost always correct). To fix this systematic bias, we exclude short sequence separation contacts in the contact mask that are within sequence separation of 11 to get corrected residue l-DDT values. Then these values are smoothed through a 9-residue-window uniform weight kernel. The residues at the lowest accuracy are determined as unreliable regions. Two definitions of regions are made: in *static* definition, the accuracy threshold is varied until the fraction of unreliable regions lies between 10 to 20% of the entire structure. In *dynamic* definition, this range is defined as a function of predicted global accuracy (i.e. average residue-wise corrected accuracy): from f_{dyn} to f_{dyn} +10% with f_{dyn} = 20 + 20*(0.55 - Q)/30, where Q refers to predicted global accuracy. f_{dyn} is capped between 20 to 40%. In the diversification stage, one thousand models were generated for each definition of unreliable regions. Static definition is applied throughout the iterative stage.

Restraints

We classified residue pairs in three confidence levels: *high confidence*, *moderate confidence*, and *non-preserving*. Highly or moderately confident residue pairs stand for those whose distance should be fixed from the reference structure (i.e. starting structure) at different strengths; non-preserving pairs refer to the rest which can freely deviate.

Confident pairs are collected if C_β-C_β distance are not greater than 20Å and whose "probability with absolute estimated error ≤ 1Å", shortly P_{cen}, is above a certain threshold (e.g. 0.7). For those pairs, bounded functions are applied at coarse-grained modeling stage, and sum of sigmoid functions at all-atom modeling stage, minima centering at the original distance d_0 for both cases:

Bounded function:

$$
\begin{aligned}
f(d) &= \frac{(d-(d_0+tol+s))}{s} + 1 && \text{for } d > d_0 + tol + s \\
f(d) &= \left(\frac{d-(d_0+tol)}{s}\right)^2 && \text{for } d_0 + tol \le d \le d_0 + tol + s \\
f(d) &= 0 && \text{for } |d - d_0| < tol \\
f(d) &= \left(\frac{d-(d_0-tol)}{s}\right)^2 && \text{for } d_0 - tol - s \le d \le d_0 - tol \\
f(d) &= \frac{(d-(d_0-tol-s))}{s} + 1 && \text{for } d < d_0 - tol - s
\end{aligned}
$$

Sum of sigmoid function:

$$f(d) = w_{fa} * \left[\frac{-1}{1+exp(-5.0*(d-d_0+tol)/s)} + \frac{1}{(1+exp(-5.0*(d-d_0-tol)/s)} + 1 \right]$$

where *s* and *tol* stand for width and tolerance of the functions. Thresholds in P_{cen} values for highly confident pairs, P_{high}, and moderately confident pairs, $P_{moderate}$, are set at 0.8 and 0.7, with (s,tol) = (1.0,1.0) and (2.0,2.0), respectively, by analyzing the network test results shown in Figure S10**.** Restraint weight at all-atom stage modeling, w_{fa}, is set as 1.0. We noticed iterative refinement with these empirically determined parameters (w_{fa}, {P_{high}, $P_{moderate}$}) brought too conservative changes. We therefore ran another iterative refinement with a more aggressive parameter set (0.2,{0.8,0.9}) and chose the trajectory from whichever sampled a higher predicted global l-DDT.

For the rest non-preserving C_β-C_β pairs whose input distances are shorter than 40Å, error probability profiles (estograms) are converted into distance potentials by subtracting error bins from the original distances d_0 and taking log odds to convert probability into energy units. Instead of applying raw probabilities from the network, corrections are made against background probability collected from the statistics of the network's predictions over 20,000 decoy structures in the training set conditioning on sequence separation, original distances d_0, and predicted global model quality. The potential was applied in full form interpolated by spline function at the initial diversification stage, and was replaced by a simpler functional form in subsequent iterative process for efficiency:

$$\begin{aligned} f(d) &= (d-9)+1 && \text{for } d > 9\,\text{Å} \\ f(d) &= (d-8)^2 && \text{for } 8 \le d \le 9\,\text{Å} \\ f(d) &= 0 && \text{for } d < 8\,\text{Å} \end{aligned}$$

for those pairs predicted from estogram as contacting within 10Å. Contacts are predicted when $P_{contact}$ > 0.8, with $P_{contact}$ = sum(P_i) over i whose $d_0 + e_i$ < 10Å and P_i stands for probability in estogram at bin *i*.

Recombination Iteration

At the recombination iteration, instead of running RosettaCM as the sampling operator, model structures are directly generated by recombining the coordinates from two models according to the predicted residue l-DDT profiles by the network. For a “seed” member, 4 “partners” are identified among the remaining 49 members in the pool that have the most complementarity to the seed in the predicted residue l-DDT profiles. All the members in the pool are recombined individually with their 4 partners, resulting in a total 200 new structural models. For each seed-partner combination, first, “complementary regions” are identified where the seed is inferior to the partner in terms of predicted l-DDT, then coordinates at the regions are substituted to those from the partner. Multiple discontinuous regions are allowed but the total coverage is restricted to a range between 20 to 50% of total residues. Next, Rosetta FastRelax [31] is run by imposing residue-pair restraints from estograms brought from either the partner or the seed interpolated into pair potentials (see above). Restraints from

the partner are taken if any residue in the pair is included in complementary regions, and from the seed for the rest pairs. Recombination iterations are called at every 5 iterations to prevent over-convergence in the pool.

Final model selection

A model with the highest predicted global l-DDT is selected among 50 final pool members. Then a pool of structures similar to this structure (S-score [33] > 0.8) are collected from the entire iterative refinement trajectory, structurally averaged, and regularized in model geometry by running dual-relax [31] with strong backbone coordinate restraints with a harmonic constant of 10 kcal/mol Å^2, which was the identical post-processing procedure in our previous work [5]. The final model refers to this structurally averaged and subsequently regularized structure. Structural averaging adds 1% l-DDT gain on average.

Testing other EMA methods in the refinement protocol

To test the refinement coupled with an external non-DL geometrical EMA, VoroMQA[10] version 1.21 was downloaded and integrated into our refinement protocol script substituting DeepAccNet for global model ranking and unreliable region prediction. Because VoroMQA does not provide any residue-pair estimations, confidence in the distance between residue i,j (denoted as P_{ij}) was estimated by the logic used in our previous work[5,24]. Here, $P_{ij} = P_i * P_j$ where $P_i = \exp(-\lambda/L_i)$ and L_i is the residue-wise accuracy from VoroMQA; λ was set to 1.4 which gave the most similar distribution in weights as what was found in our previous work. Then P_{ij} was divided by the highest 70 percentile value capping the maximum value at 1.0. Residue pair restraints were applied at these per-pair weights with the identical functional forms described above. The same logic was applied to the refinement protocol using "DeepAccNet w/o 2D"; here λ=2.0 was used.

Molecular replacement (MR)

Of a total 50 targets native structures of which were determined by X-ray crystallography in the benchmark set, 41 are tested for MR. 9 targets are excluded as their crystal structures contained other proteins or domains with significant compositions (>50%). Phaser [34] in the Phenix suite version 1.18rc2-3793 is applied with MR_AUTO mode. Terminal residues are trimmed from model structures prior to MR if they do not directly interact with the rest of residues. B-factors are estimated by taking residue-wise DeepAccNet predictions: first, u_i, the position error at residue i (in Å), is estimated by using a formula: $u_i = 1.5*\exp[4*(0.7-\text{lddt}_{\text{predicted,i}})]$, where parameters were pre-fit to training set decoy structures. Then B-factor at residue i is calculated as $8\pi^2 u_i^2/3$.

Data Availability

Decoy structures generated for the training of the DeepAccNet models and their raw predictions on the held-out test, CASP13, and CAMEO set are available at the github repository https://github.com/hiranumn/DeepAccNet.

Code Availability

Code and accompanying scripts for the model accuracy predictors (DeepAccNet-Standard, DeepAccNet-MSA, and DeepAccNet-Bert) are implemented and made available at https://github.com/hiranumn/DeepAccNet.

Acknowledgements

We would like to thank Su-In Lee, Frank DiMaio, Sanaa Mansoor, Doug Tischer, and Alex Yuan for helpful discussions. This work is supported by Eric and Wendy Schmidt by recommendation of the Schmidt Futures program (N.H and H.P), Washington Research Foundation (M.B), NIAID Federal Contract # HHSN272201700059C (I.A), The Open Philanthropy Project Improving Protein Design Fund (J.D), and Howard Hughes Medical Institute and The Audacious Project at the Institute for Protein Design (D.B).

Author contributions

N.H, H.P, and D.B. designed research; N.H., I.A., M.B., and J.D. contributed in developing the deep learning networks; N.H. analyzed the network results; H.P. contributed to the application of the network on the refinement process; N.H., H.P., and D.B. wrote the manuscript.

References

1. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 1496–1503 (2020).
2. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
3. Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 16856–16865 (2019).
4. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics* vol. 87 1011–1020 (2019).
5. Park, H. *et al.* High-accuracy refinement using Rosetta in CASP13. *Proteins: Structure, Function, and Bioinformatics* vol. 87 1276–1282 (2019).
6. Heo, L. & Feig, M. Experimental accuracy in protein structure refinement via molecular dynamics simulations.

Proc. Natl. Acad. Sci. U. S. A. **115**, 13276–13281 (2018).

7. Feig, M. Computational protein structure refinement: almost there, yet still so far to go. *WIREs Comput Mol Sci* **7**, (2017).
8. Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* **33**, 1578–1580 (2017).
9. Pagès, G., Charmettant, B. & Grudinin, S. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics* **35**, 3313–3319 (2019).
10. Olechnovič, K. & Venclovas, Č. VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* **85**, 1131–1145 (2017).
11. Bhattacharya, D. refineD: improved protein structure refinement using machine learning based restrained relaxation. *Bioinformatics* **35**, 3320–3328 (2019).
12. Heo, L., Arbour, C. F. & Feig, M. Driven to near‐experimental accuracy by refinement via molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* vol. 87 1263–1275 (2019).
13. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
14. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
15. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).
16. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. doi:10.1101/2020.07.12.199554.
17. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) doi:10.1109/cvpr.2016.90.
18. Derevyanko, G., Grudinin, S., Bengio, Y. & Lamoureux, G. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* **34**, 4046–4053 (2018).
19. Uziela, K., Shu, N., Wallner, B. & Elofsson, A. ProQ3: Improved model quality assessments using Rosetta energy terms. *Scientific Reports* vol. 6 (2016).
20. Maghrabi, A. H. A. & McGuffin, L. J. Estimating the Quality of 3D Protein Models Using the ModFOLD7 Server. *Methods Mol. Biol.* **2165**, 69–81 (2020).
21. Haas, J. *et al.* Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of

structure prediction in CASP12. *Proteins* **86 Suppl 1**, 387–398 (2018).

22. Benkert, P., Tosatto, S. C. E. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics* vol. 71 261–277 (2008).
23. Bittrich, S., Heinke, F. & Labudde, D. eQuant - A Server for Fast Protein Model Quality Assessment by Integrating High-Dimensional Data and Machine Learning. *Communications in Computer and Information Science* 419–433 (2016) doi:10.1007/978-3-319-34099-9_32.
24. Park, H., Ovchinnikov, S., Kim, D. E., DiMaio, F. & Baker, D. Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3054–3059 (2018).
25. Read, R. J., Sammito, M. D., Kryshtafovych, A. & Croll, T. I. Evaluation of model refinement in CASP13. *Proteins: Structure, Function, and Bioinformatics* vol. 87 1249–1262 (2019).
26. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
27. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
28. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
29. Modi, V. & Dunbrack, R. L. Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function, and Bioinformatics* vol. 84 260–281 (2016).
30. Mariani, V., Kiefer, F., Schmidt, T., Haas, J. & Schwede, T. Assessment of template based protein structure predictions in CASP9. *Proteins* **79 Suppl 10**, 37–58 (2011).
31. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).
32. Sun, Y. & Sundararajan, M. Axiomatic attribution for multilinear functions. *Proceedings of the 12th ACM conference on Electronic commerce - EC '11* (2011) doi:10.1145/1993574.1993601.
33. Ray, A., Lindahl, E. & Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* **13**, 224 (2012).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
35. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

36. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).

37. Meiler, J., Zeidler, A., Schmaschke, F. & Muller, M. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Journal of Molecular Modeling* vol. 7 360–369 (2001).

Figures

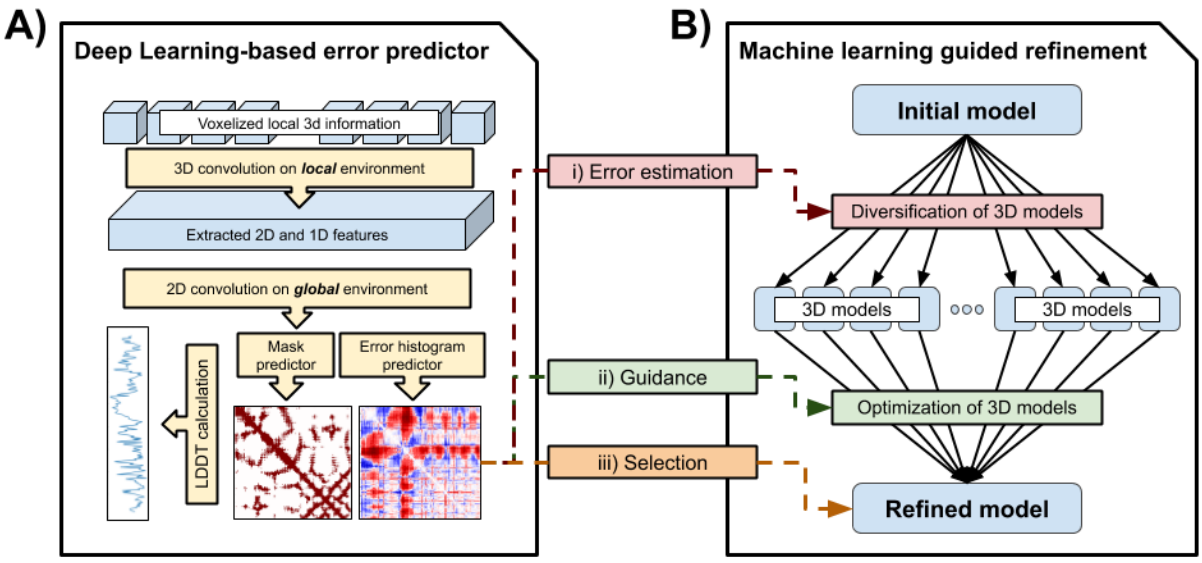

Figure 1: Approach overview. A) The deep learning network (DeepAccNet) consists of a series of 3D and 2D convolution operations. The networks are trained to predict i) the signed C_β-C_β distance error distribution for each residue pair (error histogram or *estogram* in short), ii) the native C_β contact map with a threshold of 15Å (referred to as mask), iii) the C_β l-DDT score per residue; C_α is taken for GLY. Input features to the network include: a) distance maps, b) amino acid identities and properties, c) local atomic environments scanned with 3D convolutions, d) backbone angles, e) residue angular orientations, f) Rosetta energy terms, and g) secondary structure information. Multiple sequence alignment (MSA) information in the form of inter-residue distance prediction by the trRosetta network and sequence embeddings from the ProtBert-BFD100 model (or Bert, in short) are also optionally provided as 2D features. Details of the network architecture and features are provided in Methods. B) The machine learning guided refinement protocol uses the trained neural networks in three ways; the estimated l-DDT scores are used to identify regions for more intensive sampling and model recombination, the estimated pairwise error distributions are used to guide diversification and optimization of structure(s), and finally the estimated global l-DDT score, which is mean of per-residue values, to select models during and at the end of the iterative refinement process.

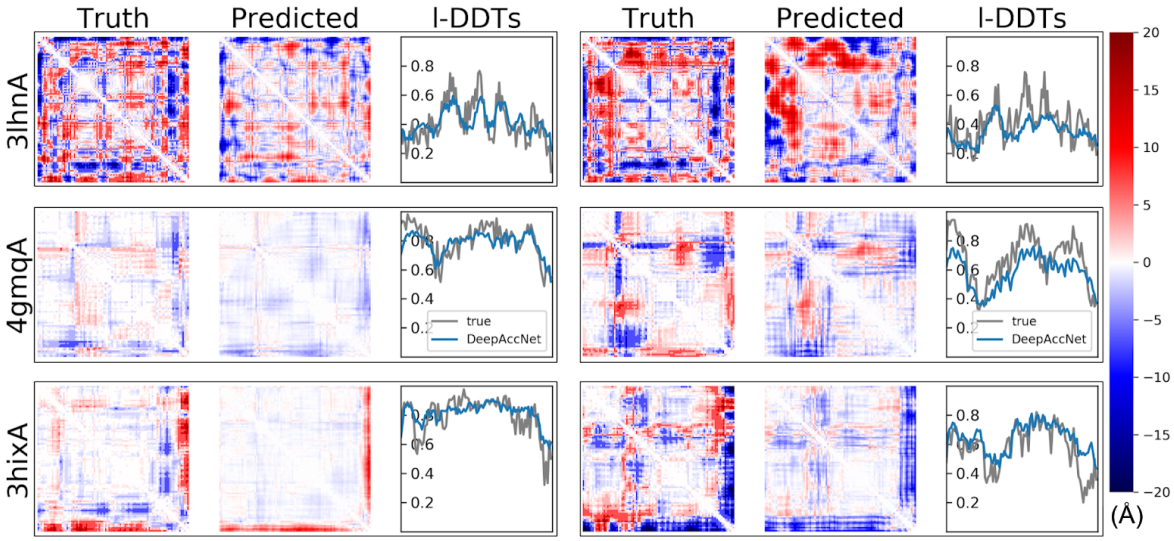

Figure 2: Example estograms and l-DDT score prediction. Model predictions for two randomly selected decoys for three test proteins were randomly selected (3lhnA, 4gmqA, 3hixA; size 108, 92, and 94 respectively; black rectangular boxes delineate results for single decoy). The first and fourth columns show true maps of errors, the second and fifth columns show predicted maps of errors, and the third and sixth columns show predicted and true l-DDT scores. The i, j element of the error map is the expectation of actual or predicted estograms between residues i and j in the model and native structure. Red and blue indicate that the pair of residues are too far apart and too close, respectively. The color density shows the magnitude of expected errors.

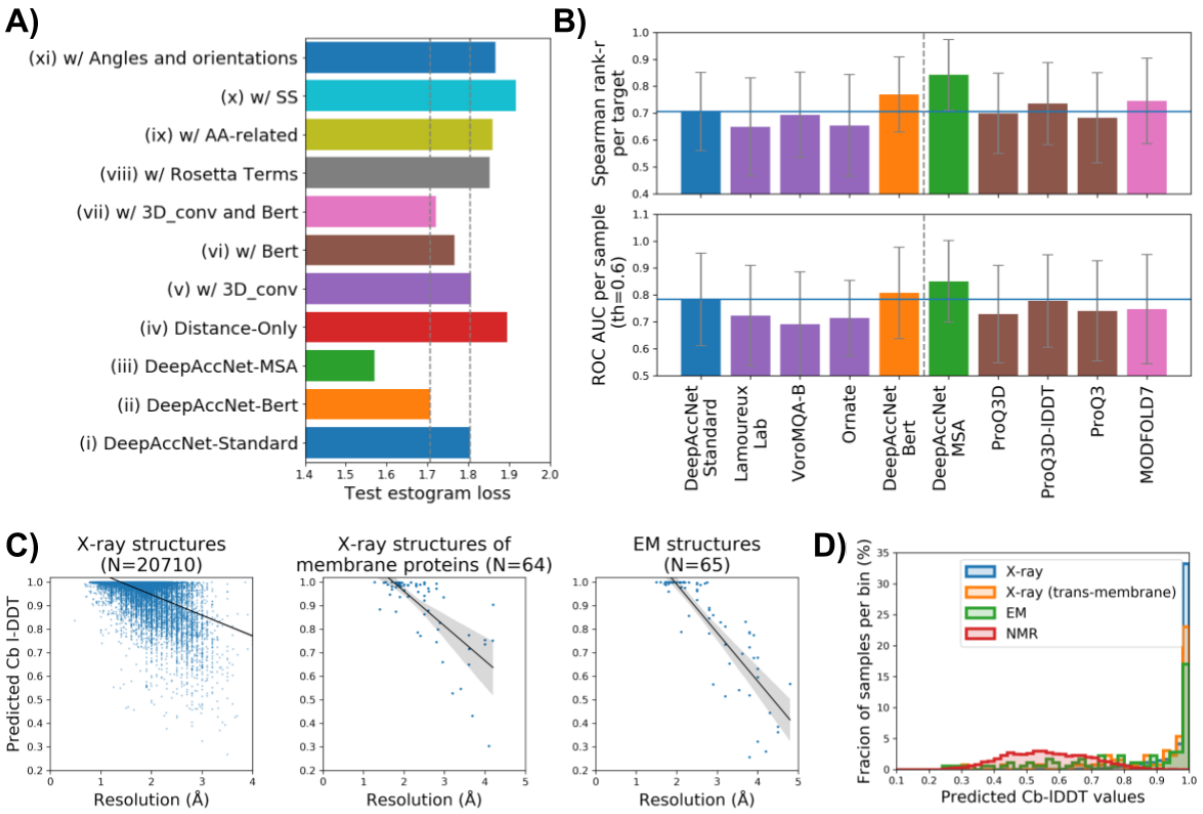

Figure 3: DeepAccNet performance. A) Contribution of individual features to network performance; all models include the distance matrix features. Overall, the largest contribution is from the features generated by 3D convolutions on local environments, Bert embeddings, and MSA information. Estogram (A, cross-entropy) loss values averaged over all decoys for each test protein are shown as one data point. The grey dotted line shows the values from predictors (i) and (ii). B) Comparison of the performance of single model accuracy estimation (EMA) methods on CASP13 data. (top) Performance of global accuracy estimation measured by the Spearman correlation coefficient (*r*-value) of predicted and actual global l-DDT scores per target protein. (bottom) Performance of local accuracy estimation measured by area under receiver operator characteristic (ROC) curve for predicting mis-modeled residues per sample (C_β l-DDT< 0.6). The blue horizontal lines show the value of DeepAccNet-Standard. The methods to the left of the dotted line do not use coevolutionary information. Quasi-single EMA method is shown in pink. Error bars show standard deviation. C) Predicted C_β l-DDT by DeepAccNet-Standard correlates with resolution for X-ray structures (left; Spearman-r 0.48 with p-value < 0.0001), X-ray structures of transmembrane proteins (middle; Spearman-r 0.64 with p-value < 0.0001), and cryoEM structures (right; Spearman-r 0.87 with p-value < 0.0001). D) X-ray structures have higher predicted l-DDT values by DeepAccNet-Standard than NMR structures.

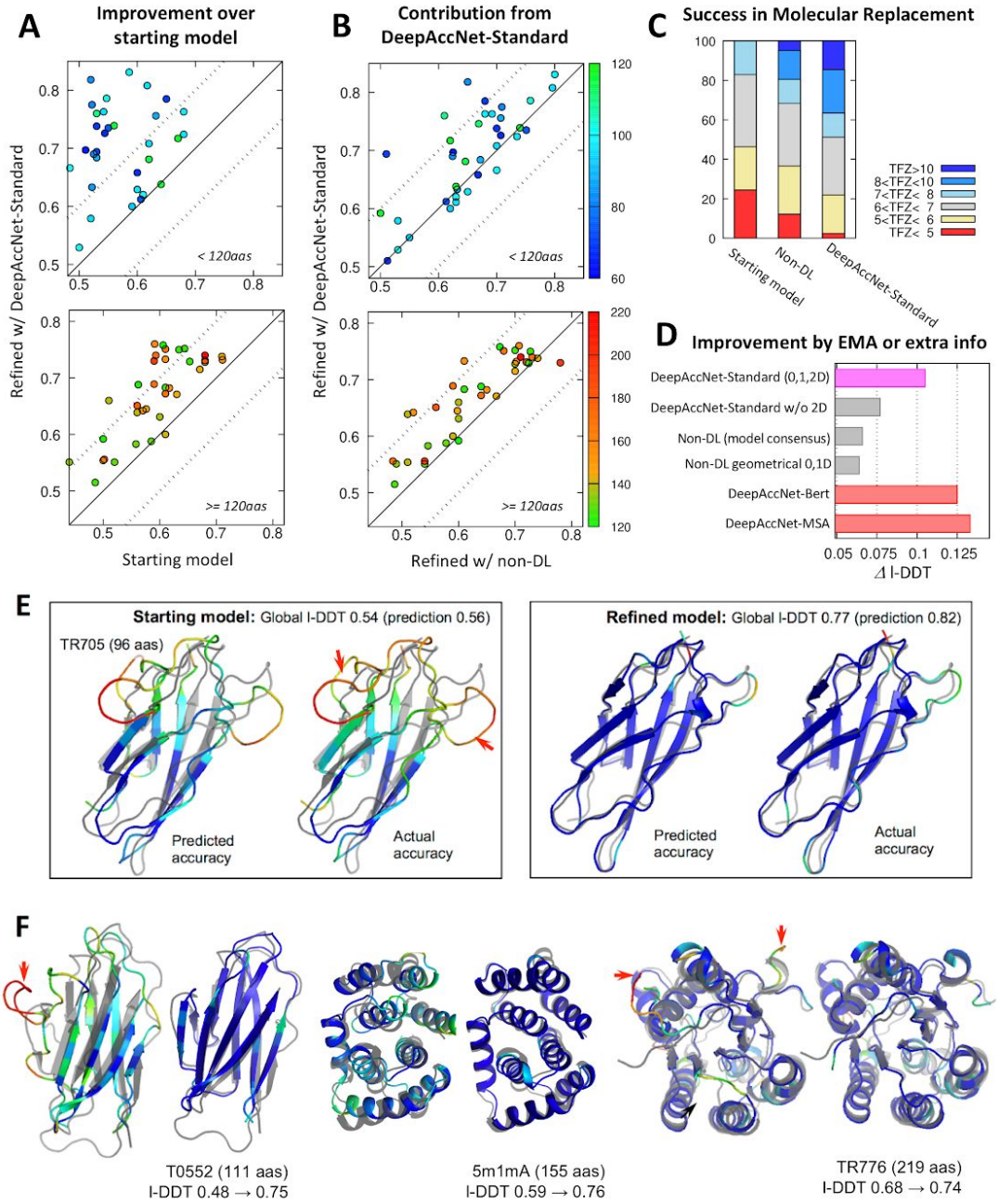

Figure 4. Consistent improvement in model structures from refinement runs guided by deep learning based accuracy predictions. Refinement calculations guided and not guided by network accuracy predictions were carried out on a 73 protein target set [5,24] (see Methods for details). A) Network guided refinement consistently improves starting model. B) Network guided refinement trajectories produce larger improvements than unguided refinement trajectories. The accuracy of the refined structure (l-DDT; y-axis) is compared with that of the starting structure in A, and with the final refined structure using non-DL-based model consensus accuracy predictions in B [5]. Top and bottom panels show results for proteins less than 120 residues in length and 120 or more residues in length, respectively. Each point represents a protein target with color indicating the protein size (scale shown at the right side of panel B). C) Molecular replacement experiments on 41 benchmark cases using three different sets of models: i) starting models, ii) refined models from the non-deep learning protocol, and iii)

guided by DeepAccNet-Standard. Distributions of TFZ (translation function Z-score) values obtained from Phaser software [34] are reported; TFZ values greater than 8 are considered robust MR solutions. D) Model improvements brought about by utilizing DeepAccNet-Standard (magenta), different EMA methods (gray bars), and other DeepAccNet variants trained with Bert or MSA features (red bars). Average improvements tested on the 73 target set are shown. For the "DeepAccNet-Standard w/o 2D" and "geometrical EMA"[10], residue pair distance confidences are estimated by the multiplication of residue-wise accuracy following the scheme in our previous work[5,24] (details can be found in Methods). E) Example of predicted versus actual per-residue accuracy prediction. Predicted and actual l-DDT values are shown before (left) and after refinement (right) with a color scheme representing local l-DDT from 0.0 (red) to 0.7 (blue). Native structure is overlaid in gray color. Red arrows in the panels highlight major regions that have been improved. F) Examples of improvements in refined model structures. For each target, starting structures are shown on the left and the refined model on the right. Color scheme is the same as E, showing the actual accuracy.

Supplementary Figures

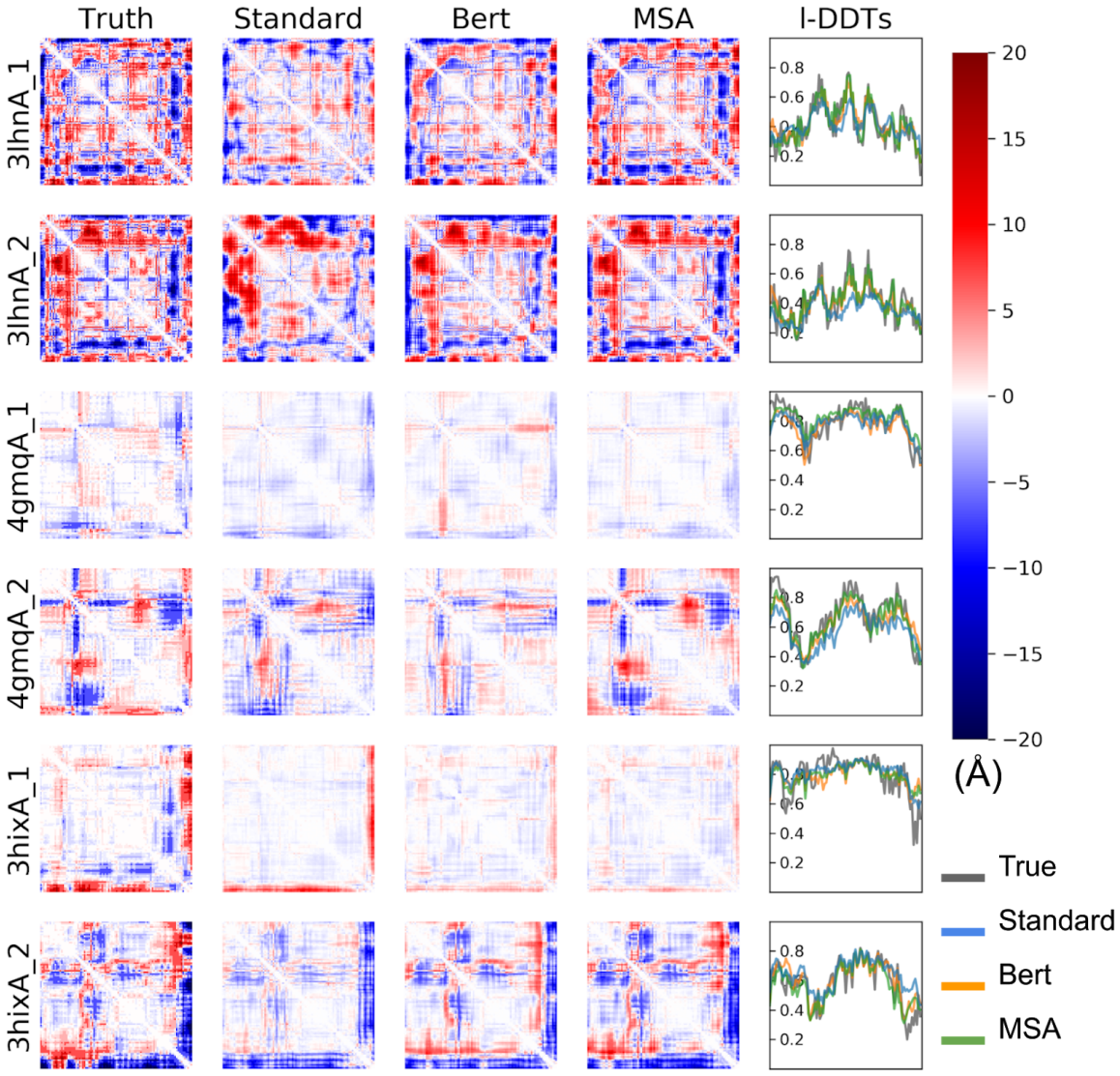

Figure S1: Example estograms and l-DDT score prediction from DeepAccNet standard, Bert and MSA. Model predictions for the same set of decoys from Figure 2 (3lhnA, 4gmqA, 3hixA; size 108, 92 and 94 respectively). The first column shows true maps of errors, the second to fourth columns show predicted maps of errors, and the last column shows predicted and true l-DDT scores. The i, j element of the error map is the expectation of actual or predicted estograms between residues i and j in the model and native structure. Red and blue indicate that the pair of residues are too far apart and too close, respectively. The color density shows the magnitude of expected errors.

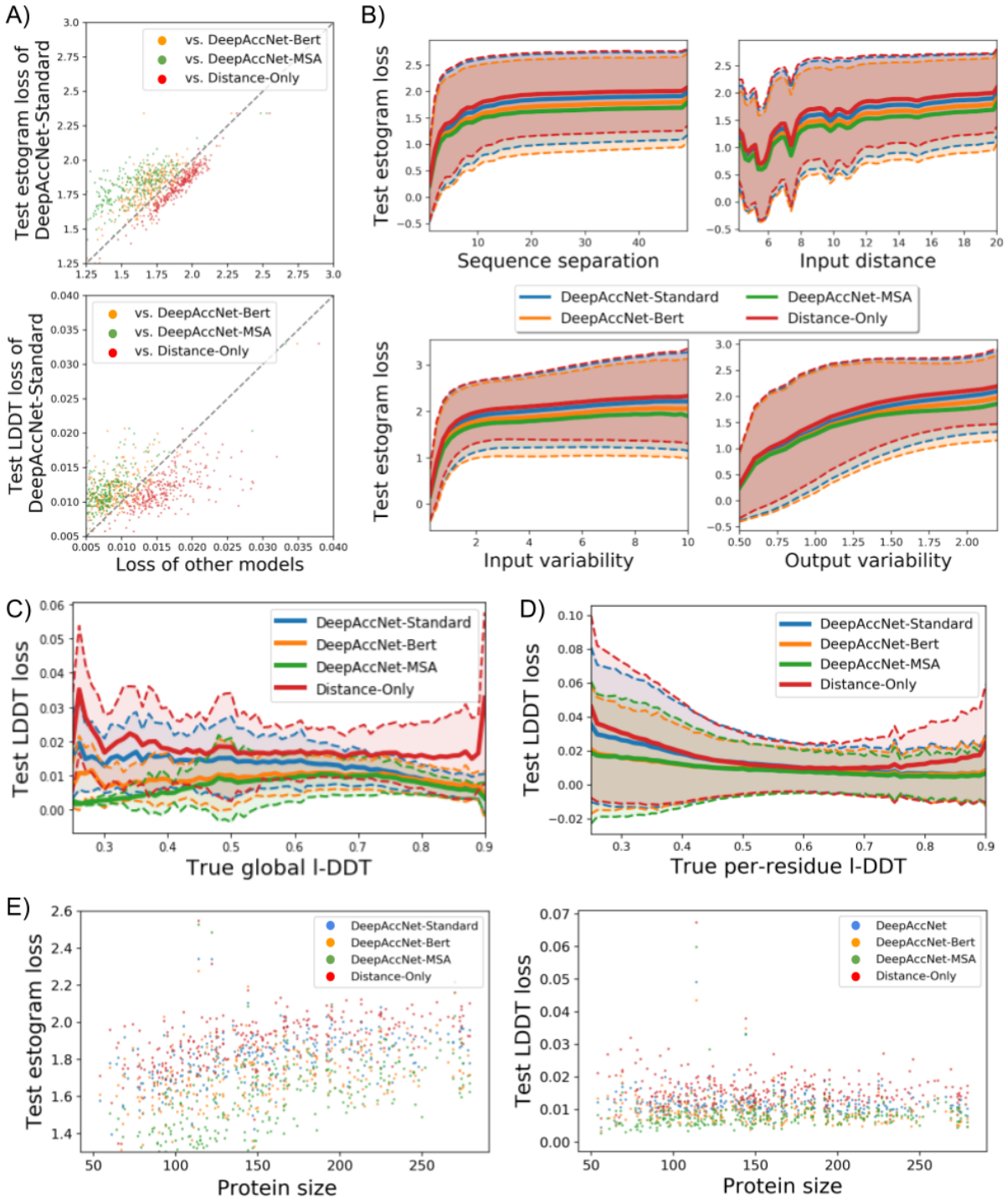

Figure S2. A) Comparison of the variants of DeepAccNet and distance-only network on predicted estograms (top) and l-DDT scores (bottom). Each dot represents the loss for a single protein averaged over all decoys. Lower loss values indicate better performance. Estograms are evaluated by cross-entropy loss, and per residue l-DDT scores are evaluated by mean-squared error. B) Test estogram loss plotted against four conditions; sequence separation, input distance, input variability (standard deviation of input distance across decoys from the same target), and output variability (entropy of true estogram across decoys from the same target). The loss values are binned in terms of x-axis properties. The mean value at each bin is shown on the y-axis, and the range of one z-score is

shown with the shaded area. CD) Dependence of l-DDT score loss on true l-DDT per-model (C) and per residue (D). Loss values are binned in terms of the true l-DDT scores. The mean of loss values at each bin is shown on the y-axis as a solid line, and the range of one Z-score is shown with the shaded area. E) Dependence of estogram (left) and l-DDT score per residue (right) loss on protein size. Each dot is an average loss value for a single target protein over all decoys.

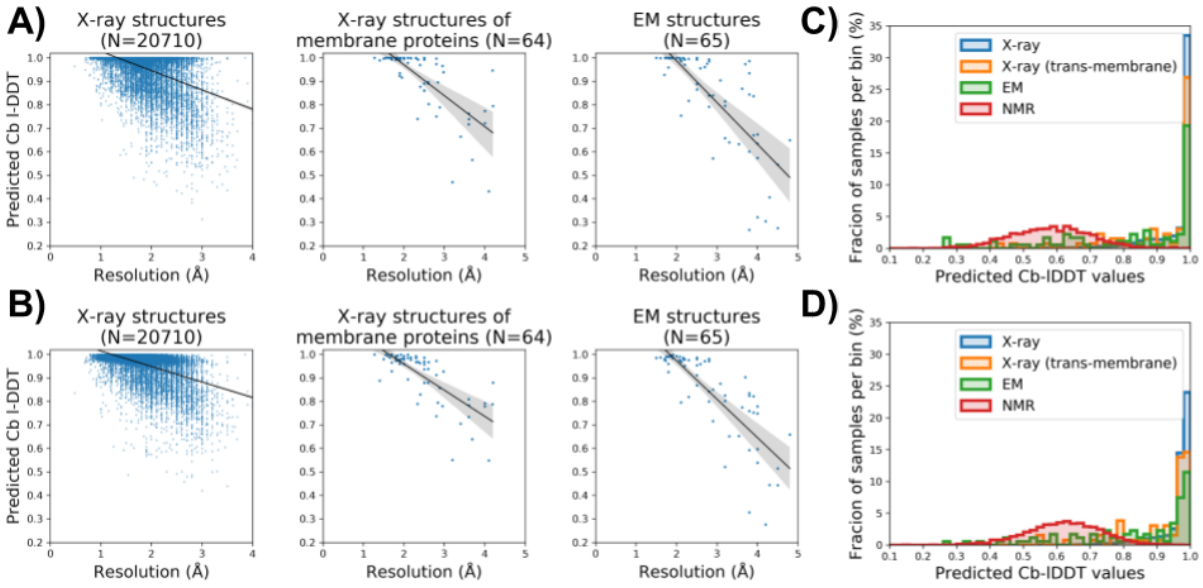

Figure S3. AB) Predicted C_β l-DDT by DeepAccNet-Bert (A) and DeepAccNet-MSA (B) correlates with resolutions for X-ray structures (left; Spearman-r 0.43 and 0.44 with p-value < 0.0001 for the Bert and MSA variants, respectively), X-ray structures of transmembrane proteins (middle; Spearman-r 0.73 and 0.74 with p-value < 0.0001 for the Bert and MSA variants, respectively), and cryoEM structures (right; Spearman-r 0.82 and 0.84 with p-value < 0.0001 for the Bert and MSA variants, respectively). CD) X-ray structures have higher predicted l-DDT values by DeepAccNet-Bert and -MSA than NMR structures.

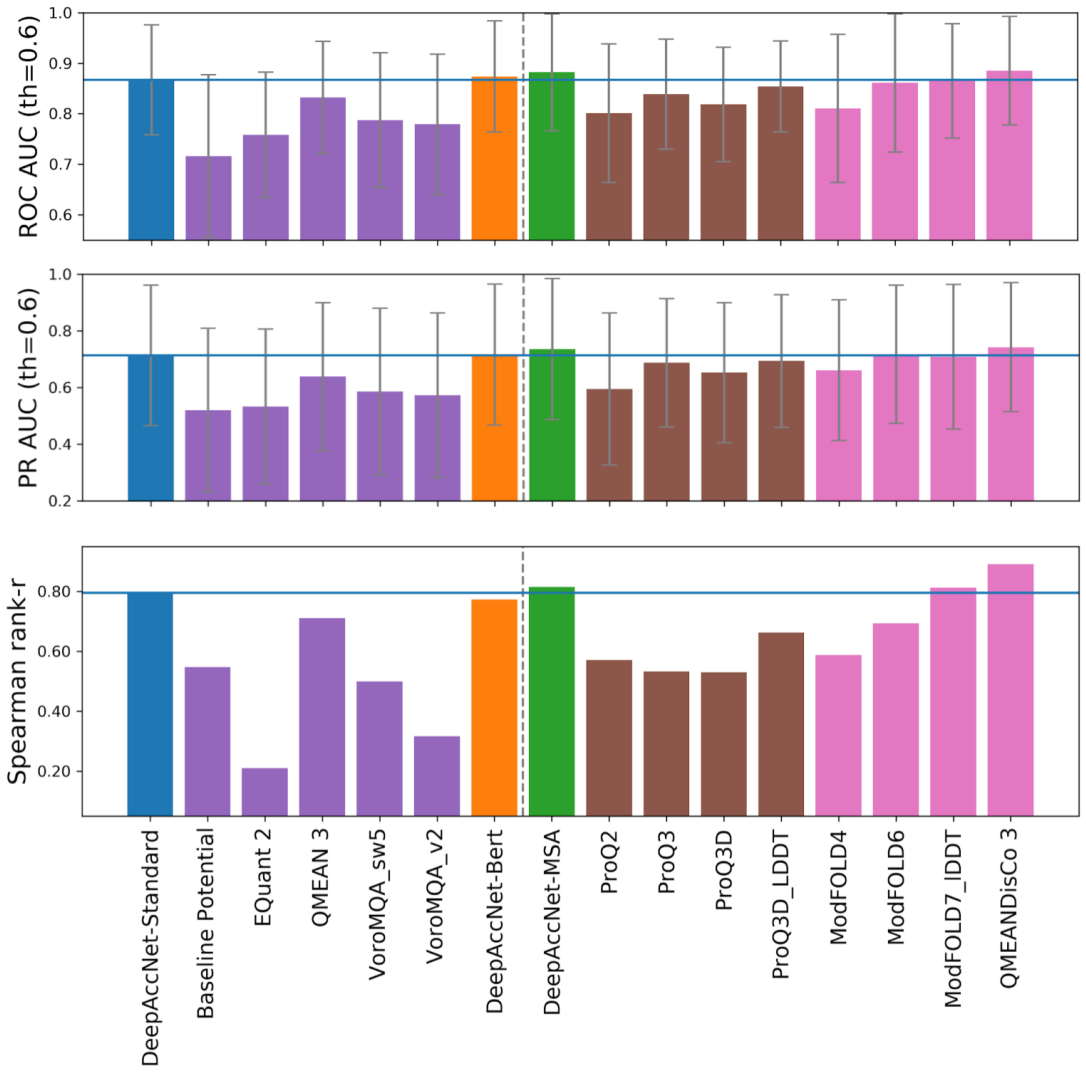

Figure S4. Comparison of the performance of single model accuracy estimation (EMA) methods on CAMEO data. (Top, middle) Performance of local accuracy estimation measured by area under receiver operator characteristic (ROC, top) curve and precision-recall curve (PR, middle) for predicting mis-modeled residues per sample (C_β l-DDT< 0.6). Error bars show standard deviation. (Bottom) Performance of global accuracy estimation measured by the Spearman correlation coefficient (*r*-value) of predicted and actual global l-DDT scores. Since the number of models per target was small, correlation was measured globally across all targets. The blue horizontal lines show the value of DeepAccNet-Standard. The methods to the left of the dotted line do not use coevolutionary information. Quasi-single models are shown in pink.

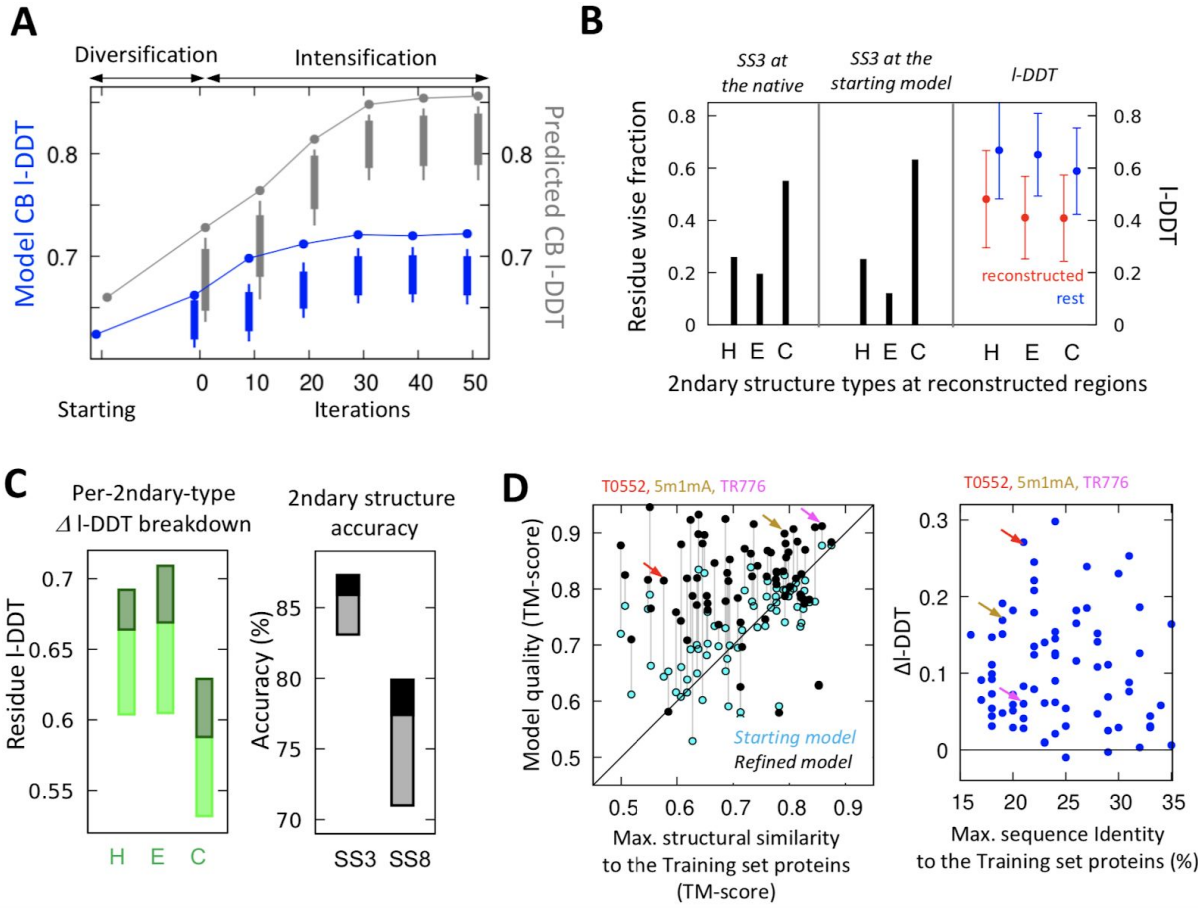

Figure S5. Detailed analyses of refinement results. A) Actual and predicted model accuracy improvements throughout the refinement trajectory. Model quality (actual in blue and predicted in gray, CB l-DDT is used for direct comparison), averaged over 73 benchmark cases, is shown through the refinement process. Points and bars show the model1 quality and the quality range of 50 models in the pool, respectively. **B)** 3-state secondary structure type at the reconstructed regions (H:helix, E:extended, C:coil). Residue-wise fractions of each type are plotted according to the native structure (left) and to the starting model structure (middle), respectively. (right) Pre-refinement l-DDT values at reconstructed regions and the rest preserved regions, shown in red and blue colors, respectively (average by circles; standard deviations by error bars). **C)** Breakdown of accuracy improvements by secondary structure types. In upper panels, light colored boxes represent improvements without DeepAccNet-Standard, while darker regions of the boxes represent additional improvements gained with DeepAccNet-Standard; these are calculated over the complete benchmark set. (left panel) Similar improvements are observed across secondary structure types. (right panel) Improvements in model secondary structure accuracy are evaluated on 3- or 8-states following DSSP annotations [28]; improvements are evident in both 3 state and 8 state local structure prediction. (bottom panel) **D)** Correlation between refinement performance and highest structural/sequence similarity of the target to the training set proteins. (left panel). Correlation between the maximum structural similarity (x-axis) versus the starting/refined model quality (y-axis) shown in TM-score [35]. (right panel) Correlation between the maximum sequence identity (%) versus the refinement performance (in l-DDT change). In both panels, targets highlighted in Figure 4 are shown in colored arrows.

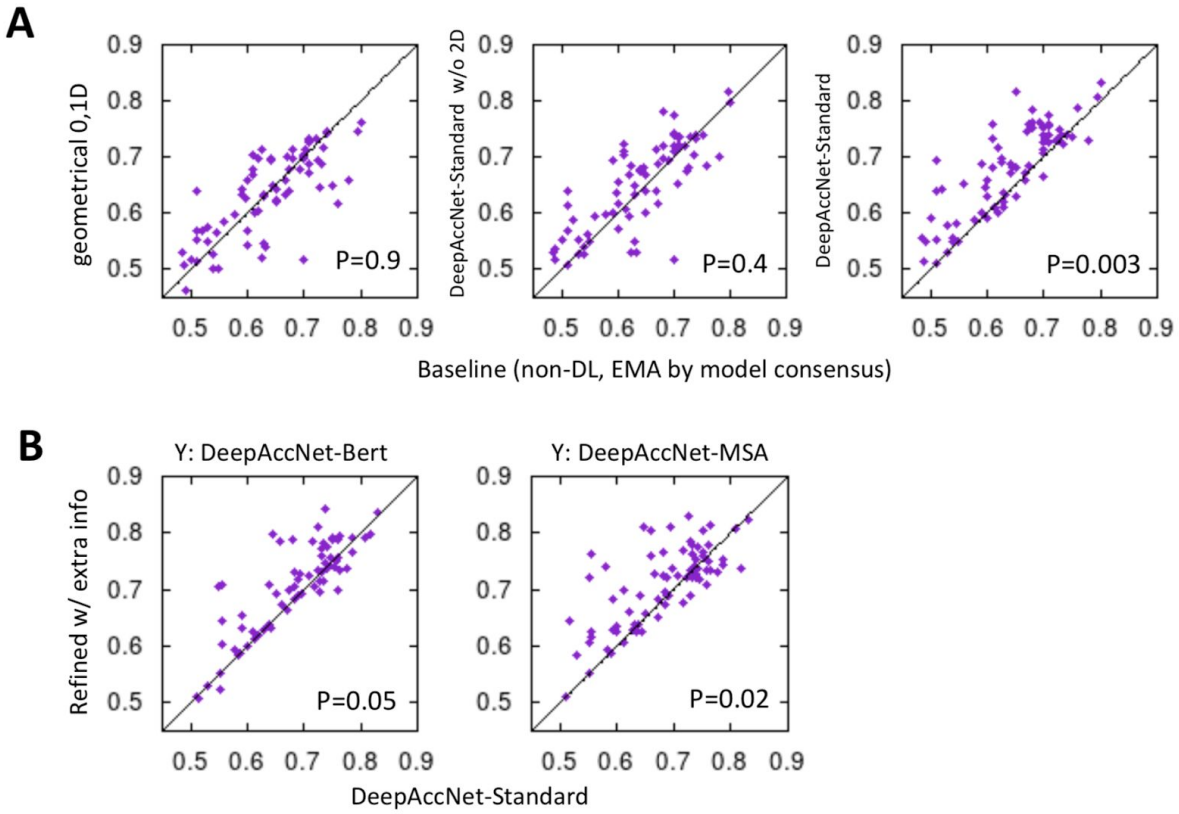

Figure S6. Breakdown of Figure 4D: Comparison of refinement performances by EMA methods or extra information utilized. A) Refinement performance with different EMA methods taken during refinement, compared to that of our baseline approach (x-axis) [5,24] using model consensus for 1D (region detection) and 2D (residue pair confidence) and Rosetta energy for 0D (global ranking). B) Refinement performance gained by providing extra input from Bert and MSA features, compared to DeepAccNet without such extra input features (x-axis).

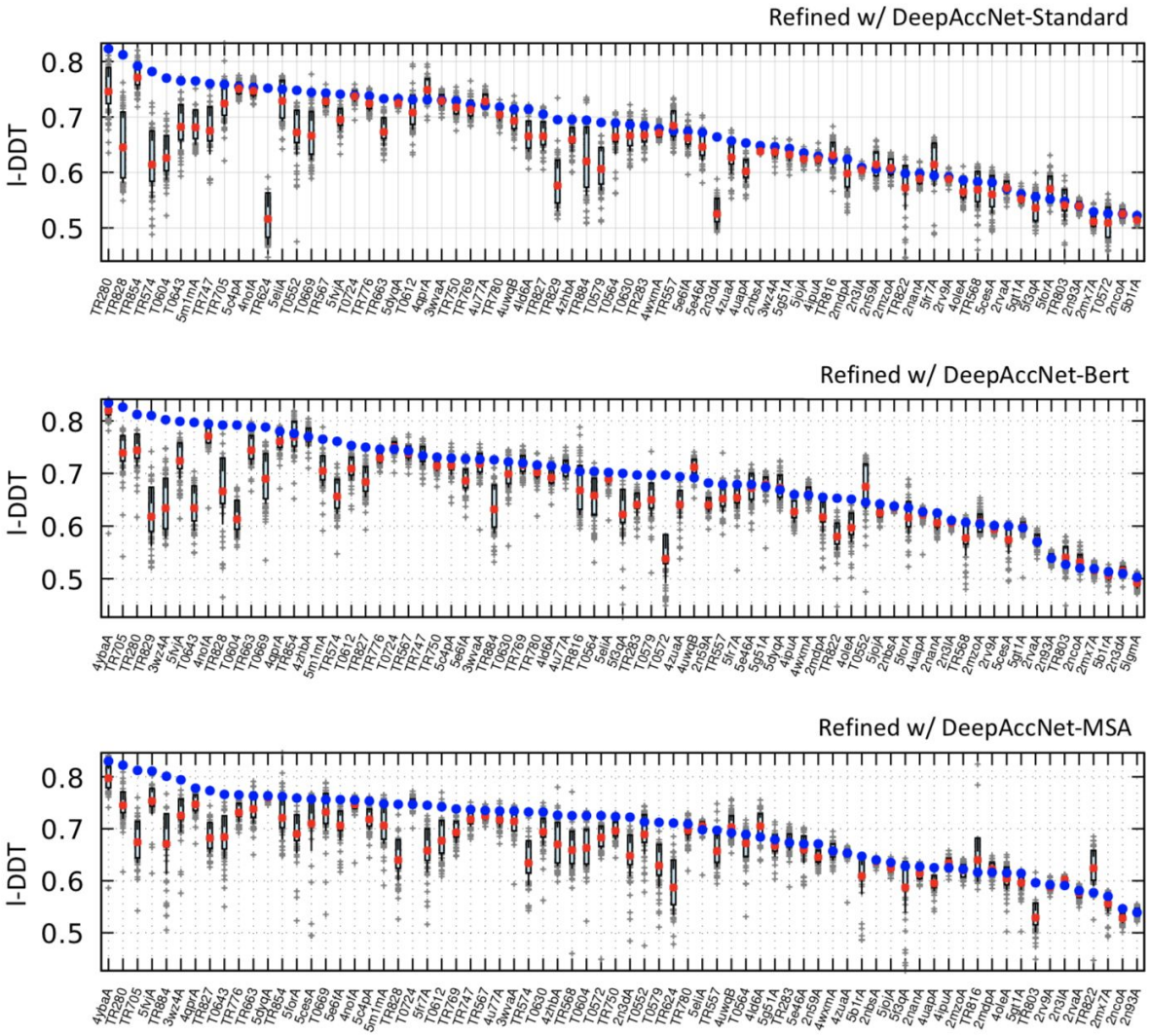

Figure S7. The model quality of the final iteration structural pool and the selected one from the refinement runs using DeepAccNet-Standard, -Bert, and -MSA. 1st and 3rd quartile of the model qualities in the final iteration models shown in cyan bars, their mean in red dots, selected by DeepAccNet (without structural averaging) in blue dots, and individual values in gray crosses.

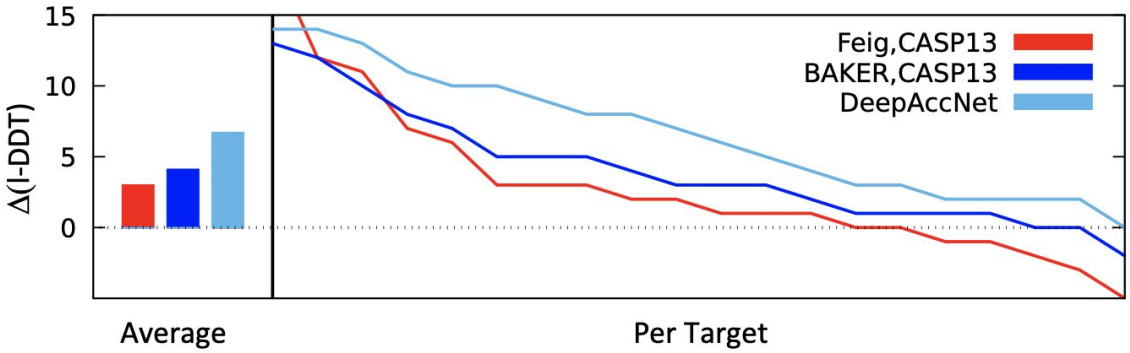

Figure S8. Performances of the methods on CASP13 refinement category targets. Improvements in l-DDT scores over starting models are shown. Two leading groups in CASP13, Feig and Baker, are brought in for the comparison against refinement with DeepAccNet; Feig group ran long MD simulations, while BAKER group ran the non-DL refinement method presented in the main text with subsequent short MD simulations. Net l-DDT changes for both of these groups range within 3~4%, compared to 7% by DeepAccNet-guided refinement. 9 targets from the CASP13 refinement category are removed from the analysis for which the native structures contain heavy oligomeric contacts or are determined at low resolutions (>3 Å).

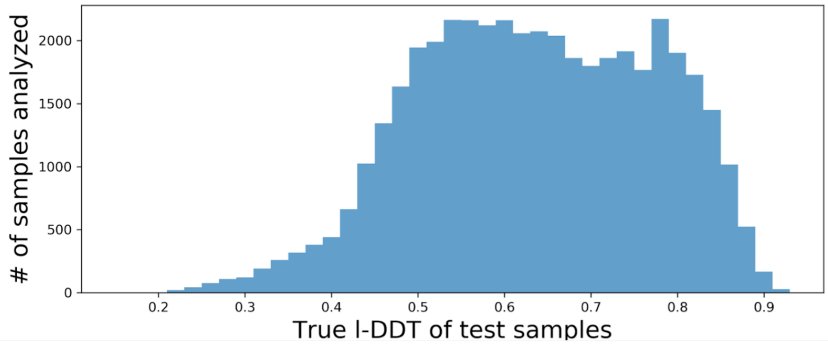

Figure S9. Numbers of samples that participated in loss analysis based on starting l-DDT scores.

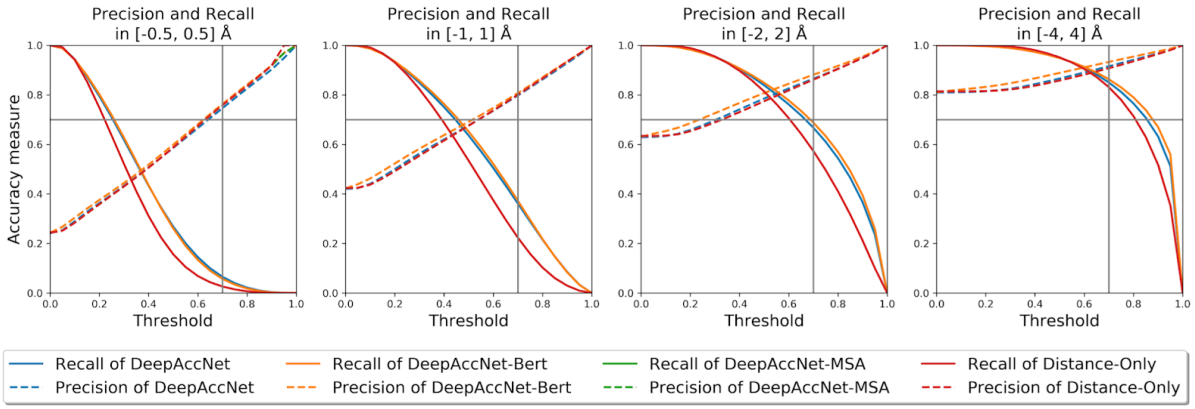

Figure S10. Assessment of binary correct/incorrect predictions. Actual error values were grouped into correct and incorrect bins. In each panel, a distance is counted as correct if the actual distance error (from that of the native structure) is within a certain range, while a prediction is counted correct if the sum of probability over the given range in the estogram is above the threshold value (x-axis). Error range definitions are [-0.5, 0.5], [-1, 1] , [-2, 2], and [-4, 4] Å from the left to the right panel. The dotted lines show recall values and solid lines show precision values. The grey lines visualize the thresholding of 0.7 used in the downstream refinement process.

Supplementary Tables

Models	Held-out proteins (# proteins=285)			True global l-DDT < 0.7			True global l-DDT > 0.7		
	Esto	Mask	l-DDT	Esto	Mask	l-DDT	Esto	Mask	l-DDT
(i) DAN-Standard	1.805	0.200	0.012	1.939	0.250	0.014	1.567	0.110	0.009
(ii) DAN-Bert	1.697	0.171	0.009	1.781	0.208	0.010	1.548	0.106	0.009
(iii) DAN-MSA	1.557	0.135	0.008	1.594	0.158	0.009	1.489	0.094	0.008
(iv) C_β distance	1.901	0.217	0.017	2.022	0.270	0.017	1.685	0.123	0.016
(v) 3D conv	1.808	0.200	0.012	1.936	0.250	0.013	1.581	0.111	0.010
(vi) Bert	1.761	0.181	0.012	1.836	0.217	0.012	1.628	0.115	0.012
(vii) 3D+Bert	1.714	0.175	0.010	1.794	0.211	0.010	1.570	0.110	0.010
(viii) Rosetta	1.854	0.209	0.013	1.986	0.262	0.015	1.617	0.115	0.011
(ix) AA-related	1.863	0.208	0.014	1.977	0.258	0.014	1.659	0.119	0.014
(x) Sec struct	1.922	0.222	0.017	2.049	0.275	0.018	1.695	0.127	0.015
(xi) Angles and orientations	1.870	0.212	0.015	2.006	0.266	0.017	1.627	0.117	0.012

Table S1: Performance of the variants of distance-based networks trained with and without a certain class of features. Performance is measured by cross-entropy for estograms and masks and mean squared error for l-DDT scores. For each setting, we ensembled the prediction from four models with the best validation performance from the same training trajectory (see Methods). Columns 2-4 report the quality of the three predictions averaged over all held-out decoy structures. Columns 5-7 report the quality of the predictions on decoys with low true quality (global l-DDT < 0.7). Columns 8-10 report the quality of the predictions on decoys with high true quality (global l-DDT > 0.7).

6B17, 3URO, 3TWG, 5DYR, 6HR0, 1P9G, 4G4L, 6EWN, 4HB6, 5JQF, 4U2W, 4HB8, 1MBN, 4HAJ, 1CYC, 1VXB, 3H4N, 2SBT, 1NXB, 4HBF, 1G7V, 2EWI, 1J0O, 2SNS, 4HDL, 3SJ4, 3H34, 4D5M, 1MBS, 1OS6, 2EWU, 1LWK, 1LYZ, 3TRV, 3SJ0, 4Z0W, 1ACX, 1PMK, 3TJW, 1HH5, 1M1R, 6DK5, 2ZVS, 3D6T, 2AOA, 3SEL, 6FM8, 5YP8, 4EFX, 1TGL, 3SJ1, 1TIA, 2EWK, 2XJI, 5HDD, 6CDX, 5VBD, 4HC3, 3NIR, 2YYX, 1HGU

Table S2: List of X-ray native structures with low C_β-lddt despite their high experimental resolution.

Distance-based	i) C_β to C_β distance map, Cα is taken for GLY, ii) Cα to Tip-atom distance map and its transpose, iii) Tip-atom to Tip-atom distance map, and iv) sequence separation map. The distance maps (i~iv) go through a variance reduction process with *arcsinh(x)*. See Table S5 for the definition of tip atoms.
Amino acid properties	i) One-hot encoded amino acids. ii) Blosum62 scores [36]. iii) Per amino-acid feature sets from Meiler et al [37].
Rosetta energy terms	i) Two-body energy terms: fa_atr, fa_rep, fa_sol, lk_ball_wtd, fa_elec, hbond_bb_sc, and hbond_sc. ii) One-body energy terms: p_aa_pp, rama_prepro, omega, fa_dun. iii) Presence of backbone-to-backbone hydrogen bonds.
Backbone angles and lengths	i) Phi, Psi, and Omega angles. ii) Standardized length between backbone atoms.
residue-residue orientations	i) Full 6 degrees of freedom of translation and rotation. ii) cosine and sine of Dihedral and planar angles defined by Yang et al [1].
Secondary structures	1-hot encoded representation of three state secondary structures given by DSSP solver.
Local atomic environments	24 by 24 by 24 voxels of size 0.8Å. In total, it covers an area of size 19.2Å by 19.2Å by 19.2Å. There are 20 channels for 20 atom types defined by Rosetta (See Table S3). The coordinate frame is fixed based on backbone N,Ca,C atoms [9].
Multiple sequence alignment	Inter-residue distance (30 by N by N, where N is protein size) predictions from trRosetta [1] gives indirect access to evolutionary multiple sequence alignments
Bert embeddings	Attention heads from the last attention layer of the ProtBert-BFD100 model [16] (16 by N by N, where N is protein size)

Table S3: Generated features for all 9 major feature classes. Some features are scaled and normalized to a reasonable range. Please refer to the code available at github for further details on the normalization scheme.

Layers groups	Descriptions
3D convolution layers	This group has four layers of 3D convolution operations with 20, 20, 30, and 20 filters with sizes of 1, 3, 4, 4, respectively. Elu activation is used. Mean pooling of filter size 4 with stride 4 was performed at the end.
Feature merging	This operation merges flattened 3D conv outputs, 2D, and 1D features (**see Methods**). One layer of 2D convolution with 32 filters of size 1 and instance normalization are applied. Elu activation is then used. Finally, the output is upsampled to 256 channels for the following ResNet operations.
Residual blocks 1	Each residual block consists of (i) elu activation, (ii) projection down to 128 channels, (iii) elu activation layer (iv) 3 by 3 convolution, (V) elu activation, (vi) projection up to 256 channels. Instance normalization operations are applied. Residual connection adds inputs to (i) with outputs of (vi). 20 residual blocks are stacked. Dilation is applied to (iv) with a cycling dilation size of 1,2,4,8.
Residual blocks 2 for estograms and masks	Two arms of four residual blocks are applied to predict estograms and masks. The same numbers of channels (256-->128-->256) are used.
l-DDT calculation layers	l-DDT values are calculated within gpu memory based on predicted estograms and masks (**see Methods)**.
Loss	(i) Estograms are evaluated with categorical cross-entropy loss. (ii) Masks are evaluated with binary cross-entropy loss. (iii) l-DDT values are evaluated with mean squared loss. Global loss is defined and shown in Method.

Table S4: Model architectures for the DeepAccNet. Please refer to the code available at github for further details on the implementation.

amino acid	**ALA**	**CYS**	**ASP**	**ASN**	**GLU**	**GLN**	**PHE**	**HIS**	**ILE**	**GLY**
tip atom	CB	SG	CG	CG	CD	CD	CZ	NE2	CD1	CA
amino acid	**LEU**	**MET**	**ARG**	**LYS**	**PRO**	**VAL**	**TYR**	**TRP**	**SER**	**THR**
tip atom	CG	SD	CZ	NZ	CG	CB	OH	CH2	OG	OG1

Table S5: Definitions of tip atoms for each residue.