11-17-2020

New Methods for Deep Learning based Real-valued Inter-residue Distance Prediction

Jacob Barger
University of Missouri-St. Louis, jsbp67@umsystem.edu

New Methods for Deep Learning based Real-valued Inter-residue Distance Prediction

by

Jacob S. Barger

A Thesis

Submitted to The Graduate School of the

University of Missouri-St. Louis
in partial fulfillment of the requirements for the degree

Master of Science

In

Computer Science

December 2020

Advisory Committee

Badri Adhikari, Ph.D.
(Chairperson)

Sharlee Climer, Ph.D.

Mark Hauschild, Ph.D.

Abstract

Background: Much of the recent success in protein structure prediction has been a result of accurate protein contact prediction–a binary classification problem. Dozens of methods, built from various types of machine learning and deep learning algorithms, have been published over the last two decades for predicting contacts. Recently, many groups, including Google DeepMind, have demonstrated that reformulating the problem as a multi-class classification problem is a more promising direction to pursue. As an alternative approach, we recently proposed real-valued distance predictions, formulating the problem as a regression problem. The nuances of protein 3D structures make this formulation appropriate, allowing predictions to reflect inter-residue distances in nature. Despite these promises, the accurate prediction of real-valued distances remains relatively unexplored; possibly due to classification being better suited to machine and deep learning algorithms.

Methods: Can regression methods be designed to predict real-valued distances as precise as binary contacts? To investigate this, we propose multiple novel methods of input label engineering, which is different from feature engineering, with the goal of optimizing the distribution of distances to cater to the loss function of the deep-learning model. Since an important utility of predicted contacts or distances is to build three-dimensional models, we also tested if predicted distances can reconstruct more accurate models than contacts.

Results: Our results demonstrate, for the first time, that deep learning methods for real-valued protein distance prediction can deliver distances as precise as binary classification methods. When using an optimal distance transformation function on the standard PSICOV dataset consisting of 150 representative proteins, the precision of top-NC long-range contacts improves from 60.9% to 61.4% when predicting real-valued distances instead of contacts. When building three-dimensional models, we observed an average TM-score increase from 0.61 to 0.72,

highlighting the advantage of predicting real-valued distances.

Acknowledgements

This year has been one of tremendous growth and education, of which a substantial amount was driven by the tasks required for the work below. I am tremendously grateful for the constant mentoring and drive provided by Dr. Badri Adhikari, without whom none of this would have been possible. The time with which Dr. Adhikari spent helping to develop this work has been and will continue to be invaluable to me. I'd also like to thank my parents, Steve and Gina, for working hard to provide the opportunities necessary to be where I am, and my partner, Lindsay, for her constant, immeasurable support.

For their diligent reading, corrections, and revisions, I'd like to thank Alex McClelland, Jill Wright, Dr. Sharlee Climer, and Dr. Mark Hauschild.

For their suggestions in the research for this manuscript, I'd also like like to recognize Dr. Adrian Clingher and Dr. Uday Chakraborty.

Contents

List of Figures

List of Tables

Chapter 1

Introduction & Background

One of the most complex problems in biology, how an amino acid sequence folds into a three-dimensional shape, i.e., the protein folding problem, has challenged researchers since the 1960s [1]. Despite the fact that the problem is of substantial medical [2] and various biological [3] significance, there are still many barriers in researchers' ability to generate protein models with reliable accuracy. Expensive laboratory methods such as Nuclear Magnetic Resonance (NMR) and X-ray crystallography provide high-resolution three-dimensional (3D) structural information [4, 5], but often fail when applied on difficult proteins [6]. As more informative data and capable computing resources became available, *in-silico* methods for predicting models were introduced to compensate for some of the disadvantages and limitations of these laboratory methods. Early on, these *in-silico* methods demonstrated accuracy behind that of the standard laboratory techniques, likely due to the hardware limitations and the computational complexity of the problem [7]. One milestone in the narrowing of this performance gap was the proposal of using inter-residue contacts, or utilizing distances $d < 8$ Å (Angstrom) as binary indicators of protein active sites in the 1970's [8, 9]. This is particularly useful because the analysis of a protein's active sites is important in determining the overall functionality of the protein [10]. As one of the first methods to define

and utilize contacts, in [9], authors generate a contact map by pairing the carbon alpha atoms within 8 Å of each other. This pairing process provided a template with which the amino acids could be arranged into a 3D model, and also allowed for computationally generated protein models to yield models similar in accuracy to medium resolution NMR and X-ray crystallography generated structures [11]. More recently, several innovations such as the integration of co-evolution signals and machine learning techniques have significantly improved both the precision of contact prediction methods [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] and the accuracy of *in-silico* protein 3D models [24, 25, 26].

From the origin of constraint-guided protein structure prediction, a question still remains—can we computationally predict real-valued distances? Inter-residue contacts have undoubtedly improved the accuracy of 3D models, and they have served as a viable substitute when real-valued distance information is not available [27]. However, inter-atomic forces within proteins naturally occur as continuous real-values. While contact predictions are usually accompanied by corresponding probabilities, they lack the granularity required for accurate 3D model reconstruction. Protein structures also contain far fewer contacts than non-contacts [28]. This makes 3D model building based solely on contact information more dependent on the conformational algorithms which actually generate the 3D models. This disconnect in the field has begun to be addressed in recent years, with newer methods [29] such as AlphaFold [30], trRosetta [31], and RaptorX [28] adopting binned (multi-classification) methods. In parallel to these efforts to continue the multi-class classification formulation, real-valued distance prediction is emerging as an alternative approach of substantial potential [27, 29, 32, 33].

One issue that arises with real-valued distance prediction formulated as a regression problem, instead of binned multi-classification or binary classification, is the tendency of the model to optimize itself to predict larger inter-residue physical distances over smaller ones. This is due to the fact that there is typically more

larger distances than shorter, contact-range, distances for a given protein, and thus the loss function will prioritize correcting the prediction of these larger distances first. However, smaller inter-residue distances are more useful for various biological and physiological applications [34], and for distance-guided modeling. As one solution, in our recent work, we proposed a real-valued distance prediction method to address this by reciprocating the distances such that a small physical distance translates into a large loss and vice versa [32]. Similarly, as another solution, flooring distances to a fixed threshold such as 16 Å was proposed in the DeepDist method [29]. Despite attempts to predict real-valued distances, these methods remain inferior, in terms of contact prediction precision, to the binary classification methods. If any of these forms of input label engineering for input real-valued distances–flooring or transforming–can be extended to perform with accuracy competitive to that of binary classification methods, it will open many new possibilities to predict distances as they naturally occur.

This work explores various label engineering strategies implemented for real-valued distance regression, their accuracy when compared to contacts, and the quality of the 3D models yielded. As one solution, we propose and explore real-valued distance prediction methods which focus on small distances by reciprocating the distances such that a small physical distance translates into a large loss and vice versa. We further examined the design of an optimal transformation function, the impact of the function the distribution of actual distances, and the performance of transformation-based predictions in the generation of 3D models. Similarly, as another solution, we rigorously test the flooring of input distances set to various fixed thresholds paired with different loss functions to gauge their impact on performance. We then generate models with the predictions yielded by this method and compare them with both transformation and contact generated models. As each of these proposed methods have demonstrated the capability to predict with competitive contact precision, we then combine the distance flooring and transformation strategies to see if they can complement each other in a way which yields

a higher accuracy. We show that each of these methods predict and generate contacts with the same or better accuracy than models trained on binary data, and that the granularity implicit to real-valued distances offers a number of benefits for improving the accuracy of generated 3D models.

Chapter 2

Methods

2.1 Dataset

We use the standard development set consisting of 3,456 representative protein chains used by the DEEPCOV [35], DEEPCON [36], and PDNET [32] methods. As test sets, we use 150 proteins in the PSICOV dataset [35] and 131 hard proteins from the Continuous Automated Model Evaluation (CAMEO) dataset, which were used to benchmark the trRosetta method [31]. After building multiple sequence alignments (MSA) from the 'fasta' sequences of these protein structures, as the input features, we utilize co-evolution features, secondary structures, position-specific scoring matrix derived features, statistical potentials, alignment statistics, and Atchley factors. The PSICOV test set is relatively easier than the CAMEO set due to the availability of high quality MSAs [31].

2.2 ResNet Architecture

We develop two-dimensional (2D) deep residual neural network (ResNet) based methods for contact and distance prediction. Each residual block consists of a batch normalization layer, followed by a rectified linear units (ReLU) activation, 64 convolutional filters of 3 x 3 kernel size, another convolutional layer with ReLU activation, and a dropout layer with a dropout rate set to 0.3. The second convolutional layer in each residual block has alternating dilation rates of 1, 2, and 4. Alternating dilation rates have been found to slightly improve the precision [29, 31, 30]. An additional convolutional layer with a single filter at the end of the network generates a single channel 2D contact or distance map. For contact prediction, we set the last activation to 'sigmoid' and for real-valued distance prediction we leave it to ReLU. For our experiments, we build a deep ResNet consisting of 64 residual blocks having around 4,747,941 network parameters. The loss function for each function is set to logarithmic hyperbolic cosine (LOGCOSH) with 'rmsprop' as the optimizer. The time required for each epoch on these parameters averages to be approximately 17.5 minutes when trained on a GTX 1080 Ti. The models were trained with: a crop size of 128, 128 epochs, 64 blocks, and 64 filters per layer. Training and generating predictions for the test sets requires approximately 30 hours.

2.3 Distance and Contact Evaluation

We evaluate our ResNet methods trained with various transformation functions using the standard precision metrics [37, 38] for evaluating predicted contacts—precision of top L/5, top L, and top NC contacts. Here, L is the number of valid residues in the corresponding native structure and NC is the total number of contacts in the native structure. Also, as defined by the Critical Assessment

of protein Structure Prediction (CASP) organizers, we define a residue pair as a contact if their carbon-beta atoms (alpha in case of glycine) are less than 8 Å apart. The method with which distances are converted to contact probabilities for evaluation is to take real-valued distance predictions d and apply the function $p = 4/d$, where p denotes the contact probability. This allows a predicted distance $d = 8.0$ Å to generate a contact probability of $p = 0.5$, and any distances $d < 4$ Å are set to $p = 1.0$, or a definite contact. For the evaluation of distances, we use the mean absolute error, root mean squared error, and local distance difference test (LDDT) score [39] using DISTEVAL available at http://deep.cs.umsl.edu/disteval/.

2.4 Label Engineering

Our first set of experiments examined the effects of transforming the real-valued inter-residue distances using various novel rational functions. Since the goal behind real-valued distance prediction is to optimize the deep learning model for predicting smaller, more useful [10] inter-residue distances, a transformation function $f(d)$ is applied to the true distance before it is passed to the model as transformed labels to compute loss. These transformation functions, in general, reciprocate the distribution of distances such that a distance larger than a threshold r Å transforms to a smaller value, and a distance less than r transforms to a larger value. For example, in the PDNET method, where the transformation function is $100/d$, and the associated threshold is 10 Å. Ideally, these transformation functions should change the distribution such that distances larger than r are compressed into a smaller range and distances smaller than r are stretched over a bigger range **(see Figure 2.1B)**. During training, a deep ResNet model only receives the transformed distance values as output labels, and hence predicts transformed distances as well. To obtain actual distance values (in Å), an inverse of the transformation function, must be carried out on the predicted distance during the evaluation of

the model's predictions. If a model predicts a transformed distance d', the inverse function $f^{-1}(d')$ is applied to obtain the actual predicted distance d Å. For example, if we take the transformation function $100/d$, and the input distance $d = 8$ Å, then we get the transformed distance $d' = 12.5$, which can then be converted back to the original distance by performing inverse function $100/d'$. It is important to note that this may be considered label engineering but not feature engineering. To study the effect of transformation visually, we also plotted the distance distributions as density plots before and after transformation, highlighting the regions around r.

In our second set of experiments, we floor the distances larger than a certain threshold t, i.e., $d[d > t] = t$, as it allows the model to focus on the prediction of shorter physical distances. This focus is due to the model quickly learning to predict the threshold t for the entire distance map **(see Figure 2.1C)**. We tested thresholds $t = 9, 10, 11, ..., 22$ for models with loss functions set to mean squared error (MSE) and LOGCOSH. We also tested the effects of combining this approach with the previously mentioned approach of distance transformation.

2.5 Distance Evaluation via 3D Model Reconstruction

The ultimate assessment of predicted contacts and distances is their power to guide 3D modeling. To apply this assessment, a series of experiments were conducted converting the distances and contacts into model-ready constraints between carbon beta atoms. This generation of models yields a direct visual comparison between distance-generated, contact-generated, and true structure 3D models. This serves to illustrate the point that, intuitively, the increased granularity and regression based nature of real-valued distance predictions may allow for more accurate

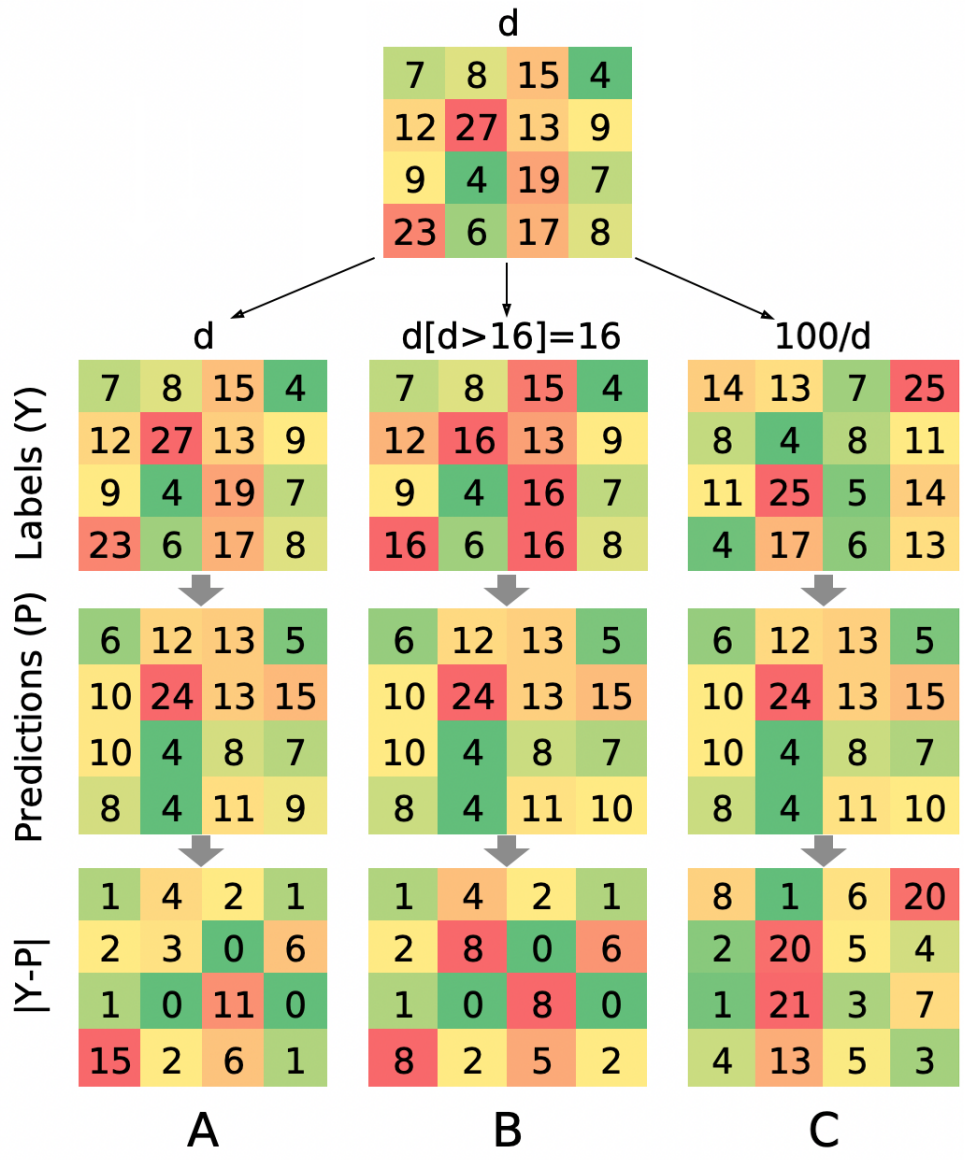

Figure 2.1: Illustration of how loss is affected by (A) no label engineering, (B) flooring, and (C) transformation. The three matrices in the first row represent the distance labels Y, the matrices in the second row represent the predictions P, and the last row shows the absolute difference $|Y - P|$. Without label engineering, loss is higher for larger distances but shorter distances are important to predict correctly. Flooring the labels resolves this to an extent but transformation inverses the distances so the loss is inversely proportional to the true distance values.

3D protein structures when compared to the same experiments utilizing binary contact-based information. The initial round of model building was carried out on the PSICOV 150 set, using an upgraded version of the CONFOLD method, [40] where it accepts a real-valued distance map as an input, instead of a contact 'RR' file. All distances predicted below 12 Å were used to build 20 models with non-relaxed distance constraints, i.e. with predicted distance itself as the upper and lower bound. For each protein, the model with minimum energy is selected as the top model for evaluation. We also validated the CONFOLD findings with a light round of Rosetta [41] model building via the Static AbinitioRelax tool of the Rosetta platform, where distances or contacts were passed in as weighted constraints via the BOUNDED function. More detail on our use of the Rosetta model building process can be found in **S1**.

Chapter 3

Results

3.1 Optimizing Transformation Functions for Real-valued Distance Predictions

The real-valued distance prediction method using the transformation function $100/d$ in the PDNET method [32] performed slightly worse than the contact prediction method on both test datasets. The precision of top L long-range contacts was 67.1% for the distance prediction method and 68.4% for the contact prediction method on the PSICOV dataset, and 46.2% vs 47.2% on the CAMEO set. To develop a real-valued distance prediction method that can surpass the contact precision benchmark, the first logical step was to generalize this transformation function in the form s/d and search for values of s which yield a high precision. We tested values of s much higher than 100, such as 300, and observed decrease in precision. However, smaller values such as 6 and 10 showed improved performance. Specifically, on the PSICOV set, the precision of top L long-range contacts for $s = 6, 10, 100$, and 300 were 66.9%, 67.5%, 67.1%, and 66.6% respectively. In **Figure 3.1**, we graphed these four transformation functions where the plot shows that the region above the $100/d$ transformation yields poor precision compared

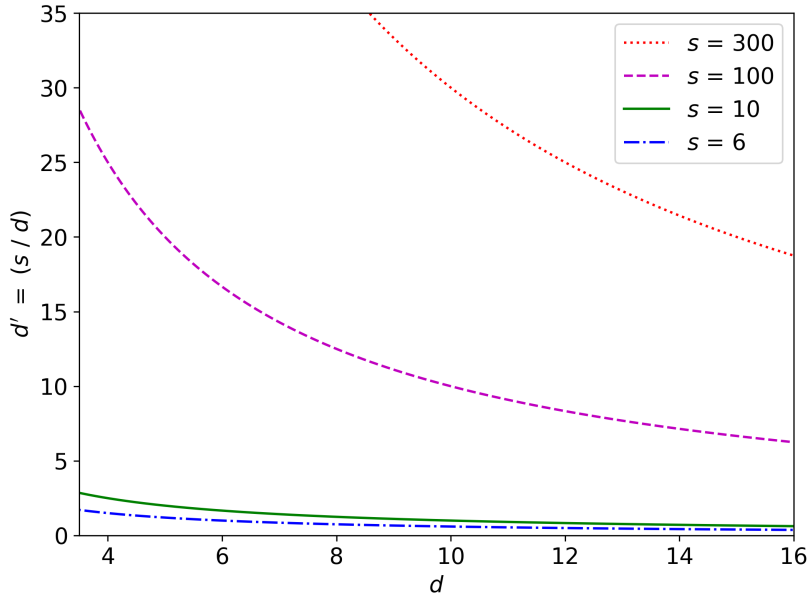

Figure 3.1: Distance transformations of the form $d' = s/d$ for $s = \{6, 10, 100, 300\}$. For a residue pair i and j where $i \neq j$, since d is always greater than around 3.5 Å, the range for x-axis is chosen to be > 3.5.

to the region below. These results suggest that we focus our search for optimal transformation function between $6/d$ and $100/d$.

The concept behind the development of a transformation function is to convert small physical distances, i.e., around 8 Å, into larger transformed values in order to observe a very high loss for small distances. Thus, the next step in our search was to refine the transformation function such that the distribution of transformed values stretches the distribution for smaller input distances. A plausible idea was to increase the steepness with which the translated distances d' become large as the true distance d approaches 3.5 Å (the minimum input distance). To this end, we examined the effects of exponentiating the function. We initially squared the transformation function, resulting in $(10/d)^2$, which generates a curve in which y approaches 8.2 as x approaches 3.5 Å. This yielded a significant breakthrough in terms of precision, outperforming the original transformation functions of the

form s/d. This new transformation function also performed similar to the contact precision method (binary). Specifically, on the PSICOV set, $(10/d)^2$ has a precision of 68.5% and the contact prediction method has a precision of 68.4%, when top L long-range contacts are evaluated. We further generalized this transformation function into the form $(s/d)^k$ where k is the exponent that requires further optimization. Next we tested $k = 3$, which performed worse. In summary, transformation functions with $k = 1, 2, 3$ resulted in precision values of 67.5%, 68.5%, and 67.7% respectively. This result suggests that high precision is observed for $1 < k < 3$. Therefore, next we tested additional values for k including 1.8 (9/5), 2.2 (11/5), 2.33 (7/3), and 2.5 (5/2), and observed the highest precision around $k = 7/3$. The precision of top L long-range contacts with this transformation function $(10/d)^{7/3}$ is 68.5% which is similar to the results of a binary predictor and with $k = 2$. This function, however, performs better than with $k = 2$ and the binary prediction method when top NC contacts are evaluated. Since we observed similar performance between s=6 and s=10, we also tested $s = 6$ with k set to 7/3, and obtained precision similar to $s = 10$. **Table 3.1** summarizes our evaluations and **Figure 3.1** visualizes all of the transformation functions plotted on the range [3.5, 16].

The average precision of the transformation functions tested above, on the contact evaluation metrics $P_{L/2}$ and $P_{L/5}$, seems to be slightly lower than the contact-based predictions evaluated on the same metrics. This gap is likely due to contact predictions having a slight advantage on smaller input numbers of contacts, such as $P_{L/5}$, as the top most confident contacts may be slightly more accurate in this range over the top most confident distances. Since the transformation-based methods are more precise in terms of P_L and P_{NC} than the contact-based method, we can observe the trend that the more rigorous the evaluation of contacts, i.e., considering more contacts, the better real-valued distance-based predictions perform.

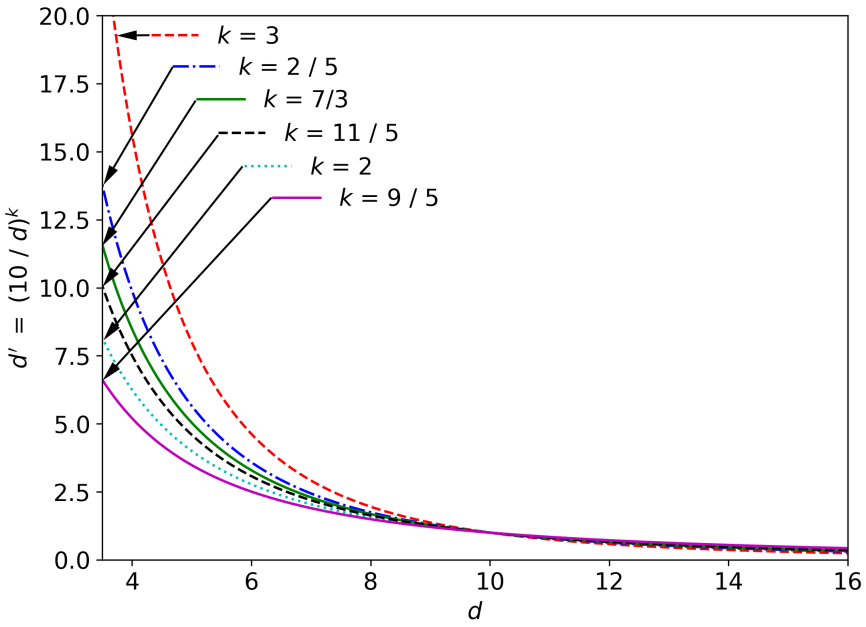

Figure 3.2: Distance transformations of the form $d' = (10/d)^k$ for $k = \{3, 2/5, 7/3, 11/5, 2, 9/5\}$. k around $7/3$ delivers optimal precision. For a residue pair i and j where $i \neq j$, since d is always greater than around 3.5 Å, the range for x-axis is chosen to be > 3.5.

Table 3.1: Comparison of the contact precision of top $L/5$, $L/2$, L and NC long-range contacts when various transformation functions are used for label engineering. For all experiments, similar ResNet models were trained (residual blocks = 64, filters per layer = 64, epochs = 128, and training window = 128). L is the length of the protein sequence and NC is the total number of true contacts in the corresponding native structure. Precision values of a contact prediction method are listed in the last row for reference.

		PSICOV 150				**CAMEO 131**			
Transformation	**Recovery**	$\mathbf{P_{L/5}}$	$\mathbf{P_{L/2}}$	$\mathbf{P_{L}}$	$\mathbf{P_{NC}}$	$\mathbf{P_{L/5}}$	$\mathbf{P_{L/2}}$	$\mathbf{P_{L}}$	$\mathbf{P_{NC}}$
$d' = (10/d)^{7/3}$	$d = 10/d'^{3/7}$	91.3	82.6	68.5	61.4	71.6	61.0	**48.0***	45.1
$d' = (6/d)^{7/3}$	$d = 6/d'^{3/7}$	91.0	**83.1**	**69.2***	**61.9***	71.1	60.4	47.6	**45.2***
$d' = (10/d)^{5/2}$	$d = 10/d'^{2/5}$	91.2	82.7	68.7	61.1	70.7	60.4	47.5	44.3
$d' = (10/d)^{11/5}$	$d = 10/d'^{5/11}$	**91.9**	83.0	68.3	60.5	71.1	60.4	47.4	44.1
$d' = (10/d)^{9/5}$	$d = 10/d'^{5/9}$	91.6	82.5	68.1	60.1	**71.8**	**61.6**	47.8	44.9
$d' = (10/d)^{2}$	$d = 10/\sqrt{d'}$	91.8	82.8	68.5	60.6	71.0	59.8	47.0	44.4
$d' = (10/d)^{3}$	$d = 10/\sqrt[3]{d'}$	91.9	83.0	67.8	59.5	70.4	59.2	46.4	42.5
$d' = 10/d$	$d = 10/d'$	90.8	81.6	67.5	60.3	69.2	58.6	45.7	42.7
$d' = 6/d$	$d = 6/d'$	91.0	80.8	66.9	60.2	70.1	58.7	46.7	44.0
$d' = 300/d$	$d = 300/d'$	90.4	81.2	66.7	58.8	67.4	57.4	45.2	41.8
$d' = 100/d^{**}$	$d = 100/d'$	91.7	82.1	67.1	59.3	70.3	59.4	46.2	42.9
Binary (contacts)	N/A	93.4	84.2	68.4	61.0	74.3	61.3	47.2	44.5

*Cases with precision higher than the contact predictor

**Method used in PDNET

3.2 How Distance Transformation Changes the Distribution of Distances

To study and visualize how various transformation functions reciprocate the distribution of inter-residue distances, we plotted the distributions of distances before and after the transformation, using the 150 representative proteins in the PSICOV set. As shown in **Figure 3.3**, the distribution of protein distances is roughly normal with a mean of around 20 Å[42]. **Figure 3.3** shows how reciprocating the distances flips the highlighted range from the left of distribution to the right. For example, the $100/d$ transformation translates the range [3.5, 8] in the original distribution to [28.6, 12.5] in the transformed distribution. In the distribution plot we highlight the distribution range for $3.5 < d < 8$ Å, the range where contacts are defined, and $3.5 < d < 8$ Å, the range useful for model reconstruction. These transformations also stretch the range in the distribution for smaller distances less than our scalar s and compress the range for large distance values greater than s. Our hypothesis is that this reciprocating and stretching/squeezing effect on the distribution allows the model to optimize its loss on the originally smaller distances. The pairing of distribution stretching and loss optimization allows the model to more easily discriminate amongst small distances.

3.3 Comparison with PDNET-Distance and PDNET-Contact

Our experiments to optimize the transformation function were performed using a shallower version of the ResNet architecture used in PDNET with depth set to 64 instead of 128 and training window set to 128 instead of 256. This allowed our deep learning training jobs to complete faster. For a complete comparison

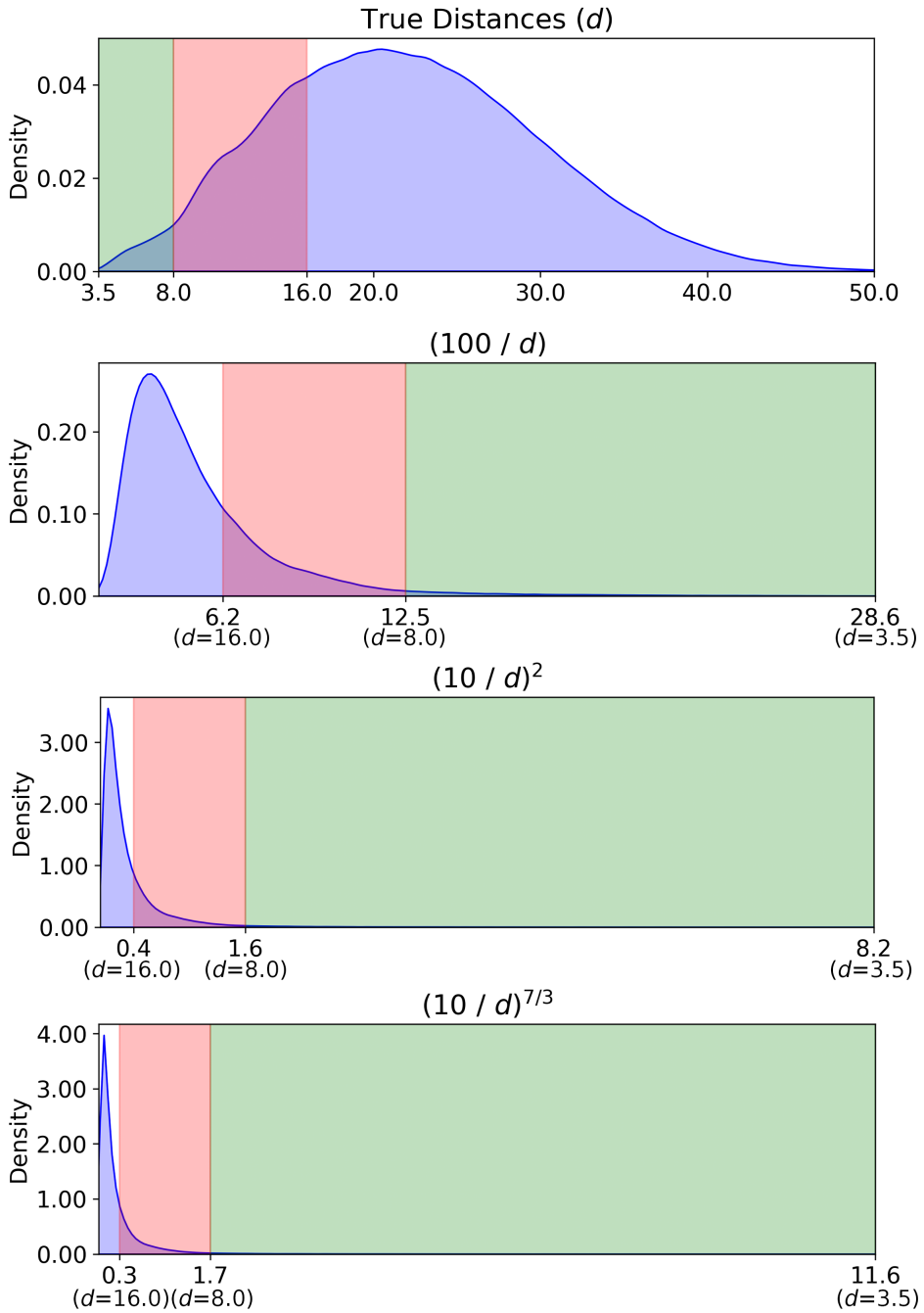

Figure 3.3: Distribution of inter-residue distances (d) in protein structures (1^{st} plot), $100/d$ (2^{nd} plot), $(100/d)^2$ (3^{rd} plot), and $(100/d)^{7/3}$ (4^{th} plot). A representative set of 150 proteins in the PSICOV set were used to obtain the distance distribution. In all plots, two distance ranges of interest, $3.5 < d < 8$ and $8 \leq d < 16$, are highlighted using green and red color respectively. The first range defines an inter-residue contact, and the second range is important for building 3D models.

with PDNET-Distance, which uses the $100/d$ transformation function, we also trained new models at depth 128 and window size 256 as done in the PDNET method using the optimal transformation function $(10/d)^{7/3}$. Our new model with the optimal transformation function performed considerably better than PDNET-Distance with around 3 to 4 percentage points higher precision when top L or top NC long-range contacts are evaluated, on both PSICOV and CAMEO sets (see **Table 3.2**). The new transformation function used in our real-valued distance prediction model also demonstrated an approximate 1 percentage point improvement in P_{NC} over PDNET-Contact, the contact prediction method, and 0.8 percentage points improvement in P_L on both the PSICOV and CAMEO sets. Notably, all these models—PDNET-Contact, PDNET-Distance (with $100/d$ transformation), and our distance prediction method (with $(10/d)^{7/3}$ transformation)—have the same number of training parameters.

3.4 Flooring Threshold Optimization

As an alternative approach to predicting real-valued distances, instead of reciprocating the distances using transformation functions, we trained various ResNet models by flooring the maximum distances to thresholds $t = 9, 10, ..., 30$, i.e., all distances higher than t are set to t during training. We also trained models by combining these two approaches, i.e., distance transformation and flooring. Specifically, we trained three sets of ResNet models at various thresholds: a) trained using the $(10/d)^{7/3}$ using 'LOGCOSH' loss function, b) trained without distance reciprocation using the 'LOGCOSH' function, and c) with mean squared error loss function. We evaluated the sets of models using three metrics—precision of top L long-range contacts, $C\beta$-LDDT score, and mean absolute error (MAE) of all medium and long-range distances predicted to be below 15 Å—with the help of DISTEVAL available at http://deep.cs.umsl.edu/disteval/. Our results,

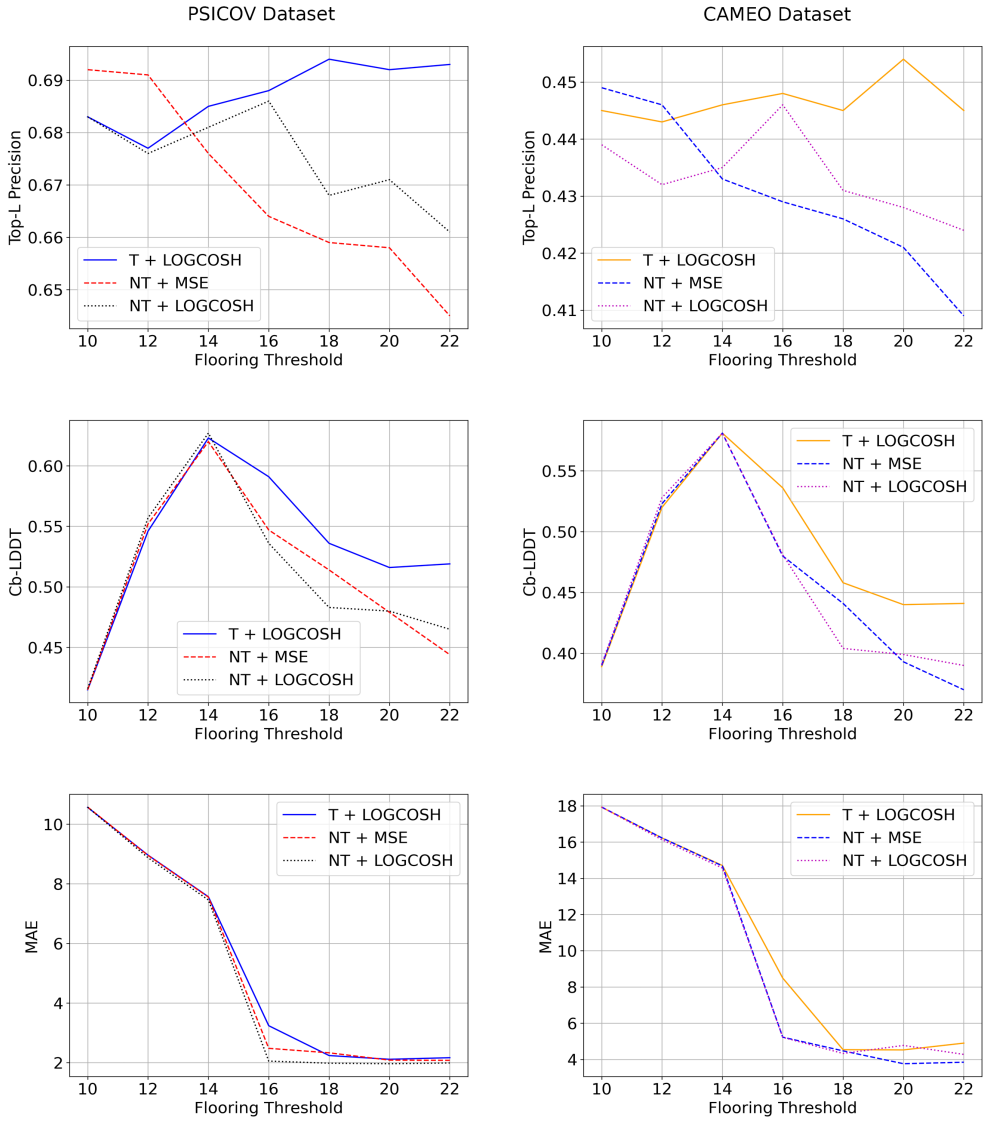

Figure 3.4: Evaluation of distances predicted for the PSICOV dataset (left column) and CAMEO (right column) dataset using the metrics, precision of top L long-range contacts, $C\beta$-LDDT, and mean absolute error (MAE), for three methods—transformation using $(10/d)^{7/3}$ along with LOGCOSH loss (T+LOGCOSH), no transformation along with mean squared loss (NT+MSE), and no transformation along with LOGCOSH loss (NT+LOGCOSH).

summarized in **Figure 3.4**, show that when no transformation function is used, increasing t decreases the precision and $C\beta$-LDDT scores. However, when optimal transformation is applied, flooring has minimal effect on precision and other metrics. When comparing the transformation and flooring approaches, we can see that transformation compresses large distances into a small range, whereas flooring removes them entirely. This gives the transformation-based model an advantage as it is able to optimize for large distance prediction when possible. Significantly high $C\beta$-LDDT scores are observed for all three sets around $t = 16$. This is likely because of the 'radius' parameter set to 15 Å, by default, in calculating the score [39]. These results also reveal the weakness of this metric—by training a model at $t = 16$ high $C\beta$-LDDT scores can be obtained—highlighting why multiple metrics should be used when evaluating predicted distances. However, when paired with the use of a transformation function, flooring can significantly improve $C\beta$-LDDT. Overall, the question of what threshold to use, we find, depends on the purpose of predicting real-valued distances. If only the predicted distances below a lower threshold such as 10 Å will be used for building 3D models, training a model at a threshold of 14 Å or 16 Å may work slightly better than no flooring. In general, however, simply using the optimal transformation function without any flooring, should work well for most applications of predicted real-valued distances.

3.5 Model Reconstruction Using Real-valued Distances

For a more rigorous evaluation of predicted real-valued distances, we reconstructed 3D models using CONFOLD [40] due to its reliance solely on distance or contact based information. All of the protein chains in the PSICOV and CAMEO datasets were used as input for the CONFOLD model reconstruction experiments. We evaluated the top-one model (not the best model) using TM-score [43] and GDT-TS

[43]. To establish a baseline for 3D model quality, we generated models using contact predictions generated by PDNET-Contact, which yielded a TM-Score of 0.51 and a GDT-TS of 49. Next, to assess the reconstruction value of our optimal translation function's ($(10/d)^{7/3}$) predictions to that of PDNET-Contact, we converted the real-valued distances into binary contacts by translating all distances below 8 Å as contacts and the rest as non-contacts. This generated 3D models with an accuracy similar to the PDNET-Contact method. Next, we built models using the real-valued distances predicted up to 8 Å, without relaxation, capped to 8 Å. Ideally, this should improve the reconstruction accuracy because it provides more granulation information for the reconstruction tool to build models. We observed this expected improvement when building models using CONFOLD. These results demonstrate that when we build models using distance constraints capped at the threshold of contact definition, the models' accuracy is on par or better than using contact constraints. The true significance of real-valued distances should be uncovered if we utilize all predicted distances up to a certain threshold, higher than the 8 Å threshold for defining contacts. Although, when we step away from these constraints and allow for the usage of the distance constraints up to 12 Å, we see a significant jump in model accuracy. Top-one models generated by the real-valued distance predictor with non-relaxed constraints up to 12 Å had an average TM-Score of 0.70 and GDT-TS of 62 when they were built using CONFOLD. Due to the incorporation of larger distances, we hypothesized it would be beneficial to the model building process to utilize constraints which relaxed more the larger the predicted distance is. This method yielded models with TM-Score and GDT-TS marginally more precise than the static constraint generation method. **Table 3.2** summarizes our reconstruction results. Similar trends were observed for reconstructions using Rosetta[41].

As an example, to demonstrate the value of predicting real-valued distances over

Table 3.2: Evaluation of transformation and flooring methods using contact precision metric, distances evaluation metrics, and 3D model evaluation. All metrics were calculated using the DISTEVAL tool. All ResNet models have same total number of parameters and were trained with same hyper-parameters (256 x 256 window size, 128 residual blocks, 64 filters per layer). Models were reconstructed using CONFOLD and Rosetta, and top-one models were evaluated.LDDT is calculated only using $C\beta$-atoms with minimum separation 6 and R value of 15 Å. PCC is Pearson corr. coeff. between $d_{pred} < 15$ with d_{true} with minimum separation 12. P_L is precision of top L long-range contacts.

Method	P_L	LDDT	MAE	PCC	TM-Score	GDT-TS
PSICOV 150 Dataset:						
PDNET-Contact	69.5	N/A	N/A	N/A	0.51	0.49
PDNET-Distance	67.5	0.47	**1.9**	0.67	0.64	0.57
$d' = (10/d)^{7/3}$ (using $d < 8$Å)	-	-	-	-	0.58	0.50
$d' = (10/d)^{7/3}$ & LOGCOSH loss	**70.3**	0.53	2.0	0.67	**0.70**	0.62
$d[d > 16] = 16$ & MSE loss	67.1	0.54	2.4	0.65	0.60	0.52
$d[d > 16] = 16$ & LOGCOSH loss	67.9	0.54	2.0	**0.70**	0.57	**0.65**
$d[d > 16] = 16$ & $d' = (10/d)^{7/3}$	**70.3**	**0.59**	2.6	0.65	0.68	0.61
CAMEO 131 Dataset:						
PDNET-Distance	46.7	0.40	**3.7**	0.48	0.40	0.30
$d' = (10/d)^{7/3}$ (using $d < 8$Å)	-	-	-	-	0.38	0.23
$d' = (10/d)^{7/3}$ & LOGCOSH loss	49.1	0.45	4.4	0.47	**0.43**	**0.33**
$d[d > 16] = 16$ & MSE loss	46.8	0.50	5.4	0.47	0.38	0.28
$d[d > 16] = 16$ & LOGCOSH loss	47.8	0.48	4.4	**0.49**	0.40	0.30
$d[d > 16] = 16$ & $d' = (10/d)^{7/3}$	**49.5**	**0.53**	6.8	0.46	0.42	**0.33**

contacts, we discuss the case of reconstructing chain A of the protein '1vhu'. The set of reconstructions implemented for this case were carried out using predicted contacts and real-valued distances as constraints for guiding Rosetta's *ab intio* reconstruction method. Additional information on the Rosetta configuration used can be found in **S1**. For this protein, we select the top-one model reconstructed using Rosetta with contacts predicted using PDNET-Contact as input, had a TM-score 0.49. Next, we used our new deep learning model trained using the new translation function $((10/d)^{7/3})$ and predicted real-valued distances for this protein chain. From this distance map, we first kept only the distances predicted below 8 Å and reconstructed models using Rosetta. The top-one in this case has a TM-score 0.55, where the slight improvement highlights the value of real-valued distances over the use of binary information. When we use predicted distances up to 16 Å, however, the TM-score of the top-one model increases considerably to 0.8. This demonstrates that the granularity of real-valued distance maps provide an advantage to the reconstruction process. This model also captures the beta-sheets observed in the true structure and the orientation of the helices are more aligned with the true model (see **Figure 3.5**). Also, the disparity of information provided by contacts and distances is illustrated in the visualization of predicted and true contact/distance heatmaps in **Figure 3.5**.

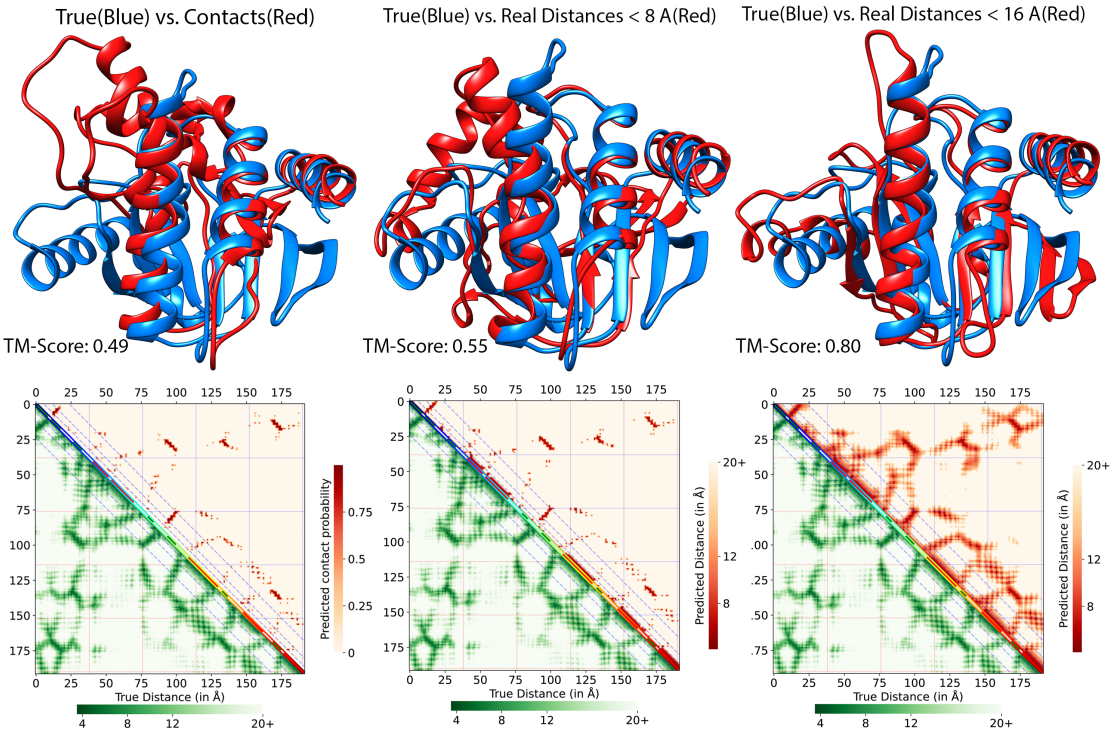

Figure 3.5: Chain A of *1vhu* is shown for three different model building strategies, the first being contact, then real distance at 8 and 16 Å thresholds. It effectively shows the difference in structure accuracy as well as the granularity of information provided by distance maps as opposed to contacts.

Chapter 4

Conclusion

Initial exploration into the performance of label engineering strategies for real-valued distances shows promise for accurate *de novo* structure prediction. When comparing these strategies, such as flooring and transformation, we found that both strategies may prove advantageous depending on the situation. Flooring is simple to implement. Transformation, however, shows promise on difficult targets, and when the target is likely to have many large ($d > 12$ Å) distances. Both of these methods show promise in terms of contact prediction precision, besting the PDNET-contact method on P_{NC} and P_L, hinting toward the idea that real-valued distances perform better than contacts when evaluated upon a larger number of known contacts. The final layer of validation, 3D model reconstruction, displayed similar trends to those observed in the other precision metrics. The models generated by both real-valued distance based strategies, transformation and flooring, outperformed the PDNET-Contact generated models on both the PSICOV and CAMEO sets. When compared with each other, the two real-valued distance based methods generated models of similar accuracy for the PSICOV set, although the models generated by the transformation based strategies, including PDNET-distance, outperformed any flooring strategies on the more difficult CAMEO set.

We look forward to seeing the rise of real-valued distance based prediction methods, and anticipate that others will propose methods to compensate for the hurdles accompanied with regression based prediction. Transformation and flooring may provide a stepping stone to further progress the accuracy of regression techniques, and this paper may lay a foundation for those looking to predict inter-residue distances as close as possible to how they appear in nature.

Chapter 5

Supplementary Material

5.1 S1: Model Reconstruction Using Rosetta

All model building carried out using Rosetta was done via the Static AbinitioRelax tool. In order to convert real-valued distance predictions into the Rosetta constraints format, the constraints function SCALARWEIGHTEDFUNC was used with a constant weight of 0.1, which was found to yield the most accurate models among the constant values 0.01, 0.1, and 1.0. The constraints function BOUNDED was used to build the models constrained by the various distance bound generation methods discussed below due to the ease with which the distances can be converted to a range. The margin of error was kept at a constant 0.5 throughout these experiments. Each constraint line took the following format:

AtomPair CB a **CB** b **SCALARWEIGHTEDFUNC 0.1 BOUNDED** u l **0.5 NOE**

With a and b denoting the carbon beta atoms the distance is predicted to be between, and u and l denoting the upper and lower bounds, methods for the generation of which are discussed below.

The first constraint generation method used to build Rosetta models applied a non-relaxed bound strategy with a ceiling set to 16 Å; as retaining predicted distances greater than this threshold tended to decrease model accuracy. The non-relaxed bounds were calculated via taking the predicted distance d, and then calculating the upper bound u via $u = d + 0.1$ and the lower bound l via $l = d - 0.1$. Then, to constrain the distances to a range on par with that of contacts, we applied the same non-relaxed constraint generation method except with a ceiling set to 8 Å. This allowed us to compare real-valued distance prediction performance to contacts on the same [0,8] Å range. Lastly, we took the contact predictions generated by PDNET-Contact, and generated constraints with static bounds when a contact is predicted to occur, i.e., $p > 0.5$. These static bounds were set according to the contact range and the minimum distance our model generator processes, or $u = 8.0$ and $l = 3.5$ Å. Each of these constraint generation methods were applied to build each chain in the PSICOV 150 set, with 200 Rosetta models being generated for each.

Bibliography

[1] Ken A Dill et al. "The protein folding problem". In: *Annu. Rev. Biophys.* 37 (2008), pp. 289–316.

[2] Robert W Doms et al. "Folding and assembly of viral membrane proteins." In: *Virology* 193.2 (1993), p. 545.

[3] JM Yon. "Protein folding: a perspective for biology, medicine and biotechnology". In: *Brazilian Journal of Medical and Biological Research* 34.4 (2001), pp. 419–435.

[4] Lubert Stryer. "Implications of X-ray crystallographic studies of protein structure". In: *Annual review of biochemistry* 37.1 (1968), pp. 25–50.

[5] MS Smyth and JHJ Martin. "x Ray crystallography". In: *Molecular Pathology* 53.1 (2000), p. 8.

[6] Herbert A Hauptman. "The Phase Problem of X-ray Crystallography: Overview". In: *Electron Crystallography*. Springer, 1997, pp. 131–138.

[7] Michael Levitt and Arieh Warshel. "Computer simulation of protein folding". In: *Nature* 253.5494 (1975), pp. 694–698.

[8] Seiji Tanaka and Harold A Scheraga. "Model of protein folding: inclusion of short-, medium-, and long-range interactions". In: *Proceedings of the National Academy of Sciences* 72.10 (1975), pp. 3802–3806.

[9] ID Kuntz, GM Crippen, and PA Kollman. "Application of distance geometry to protein tertiary structure calculations". In: *Biopolymers: Original Research on Biomolecules* 18.4 (1979), pp. 939–957.

[10] Anne-Frances Miller. "Superoxide dismutases: active sites that save, but a protein that kills". In: *Current Opinion in Chemical Biology* 8.2 (2004), pp. 162–168. ISSN: 1367-5931. DOI: https://doi.org/10.1016/j.cbpa.2004.02.011. URL: http://www.sciencedirect.com/science/article/pii/S1367593104000262.

[11] Andrej Sali. "Comparative protein modeling by satisfaction of spatial restraints". In: *Molecular medicine today* 1.6 (1995), pp. 270–277.

[12] Marcin J Skwark et al. "Improved contact predictions using the recognition of protein like contact patterns". In: *PLoS Comput Biol* 10.11 (2014), e1003889.

[13] Jesse Eickholt and Jianlin Cheng. "Predicting protein residue–residue contacts using deep networks and boosting". In: *Bioinformatics* 28.23 (2012), pp. 3066–3072.

[14] Pietro Di Lena, Ken Nagata, and Pierre Baldi. "Deep architectures for protein contact map prediction". In: *Bioinformatics* 28.19 (2012), pp. 2449–2457.

[15] Jaume Bacardit et al. "Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features". In: *Bioinformatics* 28.19 (2012), pp. 2441–2448.

[16] Yunqi Li, Yaping Fang, and Jianwen Fang. "Predicting residue–residue contacts using random forest models". In: *Bioinformatics* 27.24 (2011), pp. 3379–3384.

[17] Allison N Tegge et al. "NNcon: improved protein contact map prediction using 2D-recursive neural networks". In: *Nucleic acids research* 37.suppl_2 (2009), W515–W518.

[18] Patrik Björkholm et al. “Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue–residue contacts”. In: *Bioinformatics* 25.10 (2009), pp. 1264–1270.

[19] Sitao Wu and Yang Zhang. “A comprehensive assessment of sequence-based and template-based methods for protein contact prediction”. In: *Bioinformatics* 24.7 (2008), pp. 924–931.

[20] Jianlin Cheng and Pierre Baldi. “Improved residue contact prediction using support vector machines and a large feature set”. In: *BMC bioinformatics* 8.1 (2007), p. 113.

[21] George Shackelford and Kevin Karplus. “Contact prediction using mutual information and neural nets”. In: *Proteins: Structure, Function, and Bioinformatics* 69.S8 (2007), pp. 159–164.

[22] Alessandro Vullo, Ian Walsh, and Gianluca Pollastri. “A two-stage approach for improved prediction of residue contact maps”. In: *BMC bioinformatics* 7.1 (2006), p. 180.

[23] Piero Fariselli et al. “Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations”. In: *Proteins: Structure, Function, and Bioinformatics* 45.S5 (2001), pp. 157–162.

[24] Protein Data Bank. “Protein data bank”. In: *Nature New Biol* 233 (1971), p. 223.

[25] Mike Carson and Charles E Bugg. “Algorithm for ribbon models of proteins”. In: *Journal of Molecular Graphics* 4.2 (1986), pp. 121–122.

[26] Andrew W Senior et al. “Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)”. In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019), pp. 1141–1148.

[27] Wenze Ding and Haipeng Gong. “Predicting the Real-Valued Inter-Residue Distances for Proteins”. In: *Advanced Science* 7.19 (2020), p. 2001314.

[28] Jinbo Xu. "Distance-based protein folding powered by deep learning". In: *Proceedings of the National Academy of Sciences* 116.34 (2019), pp. 16856–16865.

[29] Tianqi Wu et al. "DeepDist: real-value inter-residue distance prediction with deep residual network". In: *bioRxiv* (2020).

[30] Andrew W Senior et al. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792 (2020), pp. 706–710.

[31] Jianyi Yang et al. "Improved protein structure prediction using predicted interresidue orientations". In: *Proceedings of the National Academy of Sciences* 117.3 (2020), pp. 1496–1503.

[32] Adhikari Badri. "A fully open-source framework for deep learning protein real-valued distances". In: *Scientific Reports (Nature Publisher Group)* 10.1 (2020).

[33] Jin Li and Jinbo Xu. "Study of Real-Valued Distance Prediction For Protein Structure Prediction with Deep Learning". In: *bioRxiv* (2020). DOI: 10.1101/2020.11.26.400523. eprint: https://www.biorxiv.org/content/early/2020/11/27/2020.11.26.400523.full.pdf. URL: https://www.biorxiv.org/content/early/2020/11/27/2020.11.26.400523.

[34] Anne-Frances Miller. "Superoxide dismutases: active sites that save, but a protein that kills". In: *Current opinion in chemical biology* 8.2 (2004), pp. 162–168.

[35] David T Jones and Shaun M Kandathil. "High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features". In: *Bioinformatics* 34.19 (2018), pp. 3308–3315.

[36] Badri Adhikari. "DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout". In: *Bioinformatics* 36.2 (2020), pp. 470–477.

[37] Rojan Shrestha et al. “Assessing the accuracy of contact predictions in CASP13”. In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019), pp. 1058–1068.

[38] Aleix Lafita et al. “Assessment of protein assembly prediction in CASP12”. In: *Proteins: Structure, Function, and Bioinformatics* 86 (2018), pp. 247–256.

[39] Valerio Mariani et al. “lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests”. In: *Bioinformatics* 29.21 (2013), pp. 2722–2728.

[40] Badri Adhikari et al. “CONFOLD: residue-residue contact-guided ab initio protein folding”. In: *Proteins: Structure, Function, and Bioinformatics* 83.8 (2015), pp. 1436–1449.

[41] Carol A Rohl et al. “Protein structure prediction using Rosetta”. In: *Methods in enzymology*. Vol. 383. Elsevier, 2004, pp. 66–93.

[42] MG Reese et al. “Distance distributions in proteins: a six-parameter representation”. In: *Protein Engineering, Design and Selection* 9.9 (1996), pp. 733–740.

[43] Yang Zhang and Jeffrey Skolnick. “Scoring function for automated assessment of protein structure template quality”. In: *Proteins: Structure, Function, and Bioinformatics* 57.4 (2004), pp. 702–710.