Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Protein Structure Prediction by Combining Deep Learning with Physics-Based Simulations using the Force Field X Software
Principal Investigator:	Michael J. Schnieders, DSc, Assoc. Prof. of Biochemistry and Biomedical Eng.
Prepared by (IIAI):	Avinash Mudireddy
Other investigators:	Mallory R Tollefson (Graduate), Guowei Qi (undergraduate)
Date:	

Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y
If yes … Planned submission date	2021
Funding agency	NIH
Grant mechanism	R01
If no … Why not? What went wrong?	

Brief summary of accomplished results:

Research report:

Aims (provided by PI):

Approximately only 40% of the human proteome has protein structural coordinates available from experiment (*i.e.,* from X-ray, NMR, or CryoEM) or homology modeling with quality templates (*e.g.,* 30% sequence identity or greater), leaving the majority of the human proteome structurally unsolved. Deep learning (DL) methods for predicting protein coordinates can help close knowledge gaps where experimental and homology models are difficult or infeasible to obtain, ultimately providing the scientific community with more protein structural coverage.

> *To assess uncertainty in the quality of predicted protein backbone structures from DL and integrate successful predictions into a complete protein folding workflow, we propose two Aims.*
>
> *Aim 1) We will assess the stability of DL protein backbone predictions to both a) missense variations and b) GPU-accelerated MD simulations. This will test the hypothesis that accurate backbone predictions demonstrate greater stability under both perturbations.*
>
> *Aim 2) We will compare the efficiency of physics-based protein folding algorithms that begin from fully extended conformations compared to those that begin from DL conformational predictions. This will test the hypothesis that folding simulations that begin from accurate DL backbone predictions converge faster than beginning from an extended conformation.*

Addressed Problem:

Mohammed AlQuraishi, in his paper "ProteinNet: a standardized data set for machine learning of protein structure" helped build ProteinNet to facilitate ML research on protein structure by providing a standardized data set, and standardized training / validation / test splits, that any group can use with minimal effort to get started.

Additionally in his paper "End-to-End Differentiable Learning of Protein Structure", he developed the Recurrent Geometric Network (RGN) that predicts protein structure from sequence.

This work augments the ProteinNet DL dataset by incorporation of additional protein experimental data and then describes a maximum likelihood style loss for the Recurrent Geometric Network (RGN) that improves its ability to predict protein structures. Compared to the original least squares loss, our novel maximum likelihood loss improves training convergence and ultimately protein structure prediction by addressing the root mean square deviation (RMSD) of backbone atoms, the Global Distance Test (GDT), the Global Distance Test High Accuracy (GDT-HA), and the TM-Score. It is observed that the maximum likelihood loss also predicts structures with more physically realistic backbone angles and thus, are better suited for physics-based simulation.

Data:

ProteinNet is a standardized data set for machine learning of protein structure. It provides protein sequences, structures (secondary and tertiary), multiple sequence alignments (MSAs), position-specific scoring matrices (PSSMs), and standardized training / validation / test splits. ProteinNet builds on the biennial CASP assessments, which carry out blind predictions of recently solved but publicly unavailable protein structures, to provide test sets that push the frontiers of computational methodology. It is organized as a series of data sets, spanning CASP 7 through 12 (covering a ten-year period), to provide a range of data set sizes that enable assessment of new methods in relatively data poor and data rich regimes.

ProteinNet datasets are available at:

https://github.com/aqlaboratory/proteinnet

Our maximum likelihood RGN implementation, dataset and trained models are publicly available (https://github.com/mallory- tollefson/rgn).

AI/ML Approach:

Traditional computational methods for predicting a protein fold combine a physics-based model of intermolecular forces, an explicit or continuum solvation model, and a sampling algorithm that builds upon molecular dynamics simulations. However, a limiting factor of a physics-based simulation approach is that protein folding often occurs on millisecond (10^{-3} seconds) or longer timescales, and similarly, GPU-accelerated molecular dynamics simulations are largely limited to microsecond (10^{-6} seconds) time scales due to the computational expense of integrating Newton's equations of motion over all atoms in a protein system.

Recent developments in machine learning—specifically, deep learning (DL)—have prompted the development of new data-driven approaches to predicting protein structure. These DL algorithms are trained using experimental protein structure data, such as 3D coordinates, evolutionary data, and multiple sequence-alignments to generate a prediction of a protein structure from its amino acid sequence. Two benefits that DL methods of protein structure prediction provide are

1) predicting protein coordinates using DL methods is faster than using physics-based protein folding methods, therefore, DL coordinates can be used as starting conformations for physics-based folding algorithms such as MELD and

2) results from DL methods can be used for computational analyses (i.e., free energy perturbation, docking, etc.) where protein coordinates from experiment or homology modeling did not previously exist.

However, currently, most DL models for protein structure prediction, including the RGN, are trained and evaluated using a least- squares style target function (*e.g.,* RMSD between the predicted structure and the experimental structure). This trains the neural network as if all atomic coordinates within a structure are equally certain, which is not the case for most experimentally determined structures due to factors such as the intrinsic flexibility of the protein and the overall quality of the experiment.

In fields such as experimental biology, structural refinement is performed using a maximum likelihood approach, where the target function is modified to account for the uncertainty of each atomic coordinate prior to optimization.

In this work, we modify the loss function used by the RGN to apply the principle of maximum likelihood refinement to the training of DL models for protein structure prediction. Structures available in the PDB vary in quality; the majority were determined based on a resolution worse than 2 Å37. Using a maximum likelihood target to train a neural network allows higher quality structures to have a greater impact on the resulting DL model, while poorer quality structures still contribute to the model to a lesser extent.

To accomplish this,

1. we first develop an improved, homogenous training dataset that incorporates B-factors from X-ray crystallography as experimental uncertainty data.
2. Following the generation of the dataset, we derive a maximum likelihood loss function based on the electron density function from X-ray crystallography that incorporates B-factors into model training. Models were trained using this likelihood loss function and predictions were generated for a series of target structures from CASP12.
3. Then compare the performance over structures predicted by the original RGN based on several geometry metrics, such as the RMSD of backbone atoms, the Global Distance Test (GDT), the Global Distance Test High Accuracy (GDT- HA), and the TM-Score.

A. <u>Curating a Dataset with Temperature Factors:</u>

Temperature factor of B-factor of an atom is its vibrational motion about a mean position and thereby influences the X-ray diffraction pattern of the structural model. The B-factor is computed using the relationship:

$$B = \frac{8\pi^2}{3}\langle u^2 \rangle$$

where $\langle u^2 \rangle$ is the mean squared displacement of the atom. A large B-factor indicates less certainty in the coordinates of an atom and is correlated with structural regions that have higher flexibility or are disordered, whereas a small B-factor is consistent with folded regions of a protein structure that have reduced conformational uncertainty. For these reasons, B-factors can serve to indicate uncertainty in the protein interatomic distances that a DL algorithm is trained on.

Protein structures determined via X-ray crystallography make up over 85% of the CASP 12 ProteinNet dataset, while the remaining 15% of structures were determined using NMR spectroscopy or cryogenic electron microscopy. Using BioJava, a software tool that obtains a protein's structural information from the RCSB based on its PDB ID40, we obtained B-factors for the backbone atoms of each X-ray crystallography structure. These B-factors were combined with the ProteinNet dataset into a new dataset called ProteinNetX.

For NMR structures, B-factors are modified differently. We derive experimental uncertainties for NMR structures by computing the root mean square fluctuation (RMSF) of each atom over the set of NMR models. This per-atom RMSF can be computed in Angstroms using the following relationship:

$$RMSF_i = \sqrt{\frac{\sum_{k=1}^{m} \left| \boldsymbol{r}_{i,k} - \bar{\boldsymbol{r}}_i \right|^2}{m}}$$

where $r_i = \frac{\sum_{k=1}^{m} r_{i,k}}{m}$ is the average position of an atom over m models and $r_{i,k}$ is the position of atom i in the kth model. To mirror the scale of crystallographic B-factors, we multiply each RMSF by a constant factor to obtain the NMR uncertainty value:

$$B_{i,NMR} = \frac{8\pi^2}{3}(RMSF_i)^2$$

We then add a constant factor of 20 Å to each B-factor to shift the peak of the distribution of computed NMR B-factors to match the distribution of B-factors from X-ray crystallography. These NMR B-factors were computed for the N, C_α, and C atom of each protein to be included in the ProteinNetX database. Computing B-factors for single-model NMR structures and cryoEM structures is beyond the scope of this work, but could be explored in the future to increase the size of ProteinNetX.

For NMR structures, B-factors are modified differently. We derive experimental uncertainties for NMR structures

B. <u>Reformulating Training of the Recurrent Geometric Network Using a Maximum Likelihood Loss Function:</u>

The trained Likelihood-RGN takes as input in an amino acid sequence and its corresponding position specific scoring matrix, and ultimately returns a 3D protein backbone. Likelihood-RGN is comprised of three stages: computation, geometry, and assessment. In the first stage, structural and evolutionary

information from the amino acids is integrated into adjacent units. Three values are output for each unit, corresponding to the backbone torsional angles (i.e., ϕ, ψ, and ω) of each residue. In the second stage, the protein backbone 3D Cartesian coordinates are defined by iteratively extending the amino acid chain by one amino acid based on the predicted torsional angles and known bond lengths and bond angles via internal coordinates. The third stage outputs the final 3D structure of the protein, however, during training, the third stage evaluates the loss between predicted and experimental structures, and that loss is minimized with backpropagation (Figure 1).

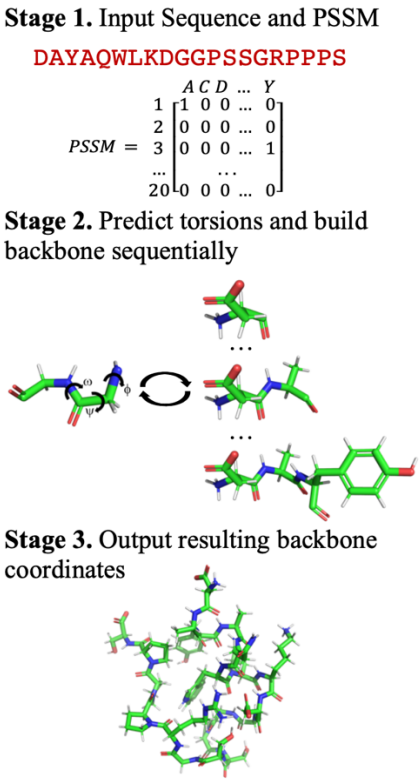


Figure 1. The three stages of Likelihood-RGN. In stage 1, a sequence and Position Specific Scoring Matrix (PSSM) are submitted to the Likelihood-RGN; in stage 2, three backbone torsions (*i.e.*, ψ, ω, and ϕ) are predicted for each amino acid and the backbone is sequentially built; stage 3 outputs the final 3D structure.

The overall likelihood of the experimental coordinates X_0, given the RGN computed coordinates X_C, is then given by the product of all interatomic distance likelihoods

$$P(\mathbf{X}_o;\mathbf{X}_c) = \prod_{i=1}^{n-1}\prod_{j=i+1}^{n} f(\boldsymbol{d}_{ij})$$

$$f(\boldsymbol{d}_{ij}) = \left(\frac{4\pi}{B_i + B_j}\right)^{3/2} \exp\left[-\frac{4\pi^2\left|\boldsymbol{d}_{ij} - \mathbf{r}_{ij}\right|^2}{(B_i + B_j)}\right]$$

The negative of the natural log of the total likelihood is the loss that is minimized during Likelihood-RGN training (ignoring constants)

$$F(\mathbf{X}_o; \mathbf{X}_c) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{4\pi^2 \left| \boldsymbol{d}_{ij} - \mathbf{r}_{ij} \right|^2}{(B_i + B_j)}$$

Where d corresponds to position of electron density, r corresponds to atomic centers, B represents B-factors and n is the number of atoms.

We determine the success of our maximum likelihood loss reformulation by training Likelihood-RGN for protein structure prediction using our ProteinNetX dataset and comparing results to the originally published least-squares loss RGN and corresponding ProteinNet dataset. Hyperparameters for training our model mirror the hyperparameters selected in previous work, and each model is evaluated using the distance-based root-mean- square-deviation (dRMSD) metric, which was used as the loss for RGN in previous work. The dRMSD is computed by first evaluating the pairwise distances between all atoms in the predicted structure and all atoms in the experimental structure individually followed by evaluating the RMSD between the two sets of distances. We evaluated and compared the quality of proteins generated from the two different loss functions by computing the Global Distance Test (GDT), the High Accuracy Global Distance Test (GDT-HA), and the Template Modeling Score (TM-Score).

C. Reformulating Training of the Recurrent Geometric Network Using a Maximum Likelihood Loss Function:

We optimized the testing set proteins predicted by the two trained models (i.e., the original, least-squares RGN versus our modified Likelihood-RGN) using the 2018 AMOEBA, and the fixed-charge Amber force fields under the generalized Kirkwood implicit solvent. The proteins were locally optimized using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm to a root mean square (RMS) gradient convergence criterion of 0.1 kcal/mol/Å. Minimizing to a tight convergence criterion prior to any future optimization or analyses allows relaxation of the tightly folded predicted backbones ultimately reducing steric clashes. We compared the proteins from the original least-squares RGN and our Likelihood-RGN by analyzing the differences in how they perform under the influence of a fixed-charge versus a polarizable force field. We evaluated the RMSD, GDT, GDT-HA, TM-Score, and Ramachandran plots prior to physics-based optimization, and after minimization with both force fields. Minimizing the proteins as done here helps prepare the coordinates for future physics-based simulation and analyses such as molecular dynamics, global side-chain optimization or free energy perturbation of missense variations.

IIAI involvement:

IIAI consultant Avinash Mudireddy, was only involved in helping the team understand the code repository for Recurrent Geometric Network (RGN) developed by Mohammed AlQuraishi. The involvement is intermittent whenever the investigators needed the help.

Experimental methods, validation approach:

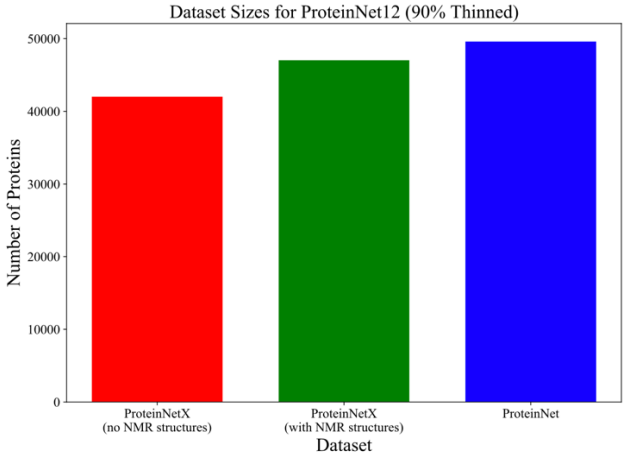
Results:

A. The ProteinNetX Structure Prediction Dataset with Temperature Factors:

When limited by only protein structures solved by X-ray crystallography, ProteinNetX covers 86.47% of structures in the 90% thinned ProteinNet dataset for CASP12. After including computed B-factors for multi-model NMR structures, ProteinNetX contains 97.9% of the structures in the original ProteinNet dataset (Figure 2a). When computing NMR B-factor equivalents through the use of RMSF, a constant of 20 Å is added to the RMSF so that the peak of the NMR distribution of B-factor equivalents parallels the X-ray structure B-factor distribution (Figure 2b).

Our two ProteinNetX datasets (one with and one without NMR data) are publicly available (https://github.com/mallory-tollefson/rgn).

A)



B)

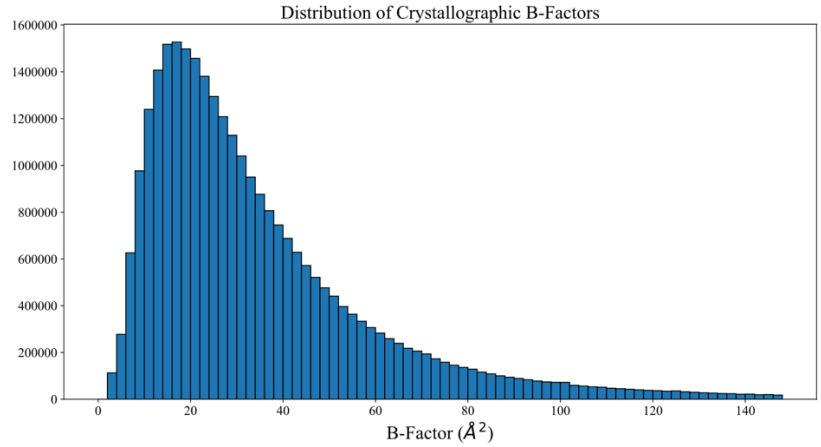
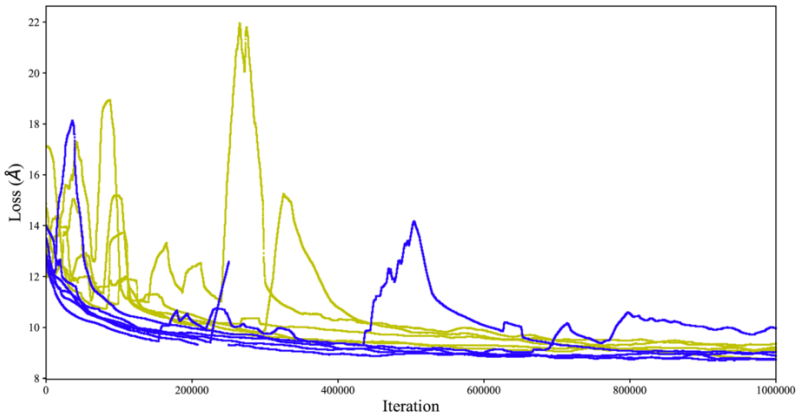


Figure 2. A) Dataset sizes for three variations of ProteinNet. Red (left) shows the ProteinNetX dataset with only structures from X-ray crystallography totaling at 42,019 structures, green (middle) shows ProteinNetX with NMR structures included and a total of 47,035 structures, and blue (right) shows the dataset as originally published with no B-factors and 49,600 structures. B) B-factor distribution for all X-ray crystallography structures in ProteinNetX.

B. Improved Structure Prediction with a Maximum Likelihood Loss

Five pairs of models were trained using only the X-Ray structures from the ProteinNetX dataset for CASP12. Each pair was initialized from the originally published hyperparameters and random seed, controlling for all factors aside from the loss function. Within each pair of models, one model was trained using the original, least-squares loss and the other was trained using our maximum likelihood loss. Each model was trained for 1.5 million iterations (i.e., one batch of training with 32 proteins per batch), followed by reducing the learning rate by a factor of 10 with 10,000 additional iterations of training. Plotting a running average of the least-squares loss versus training iteration for each pair of the five trials reveals that using a maximum likelihood loss to downweigh the contributions of highly disordered regions of proteins results in the initial training iterations being much more stable (Figure 3a). Additionally, training with a maximum likelihood on average converges to a smaller loss than training with least-squares.

A)



B)

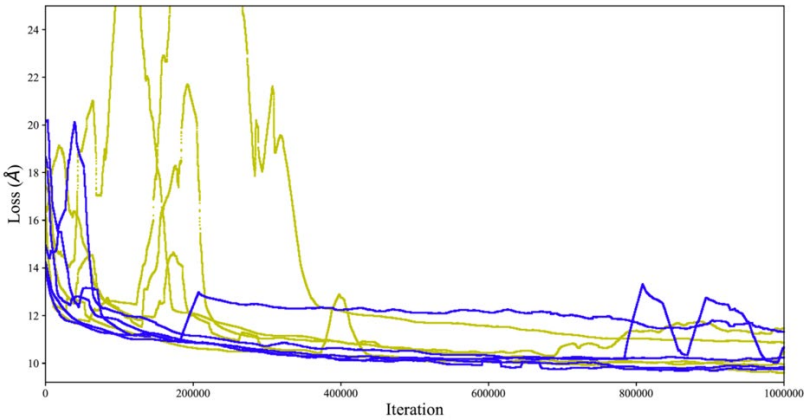


Figure 3. Running average least-squares loss of the testing dataset over the first 1,000,000 training iterations of models trained by A) the dataset including only X-ray structures and B)the full ProteinNetX dataset with NMR models included. Five pairs of trials were run for both datasets, where each pair was initialized from the same random seed. Training with a maximum likelihood loss (blue curves) shows smoother gradient descent over the initial training period compared to training with the least-squares loss (yellow curves), as the network places less weight on highly disordered regions of proteins when B-factors are included in the loss function. Additionally, the maximum likelihood RGN trials converge to smaller losses on average than the least-squares RGN.

Using each final trained model, blind predictions were generated for a testing dataset of 63 CASP12 target structures that were not present in the training dataset. On average, the maximum likelihood loss models outperform the original models when the predicted structures were evaluated using dRMSD (i.e., the originally published least-squares loss function), RMSD, the Global Distance Test (GDT), the High Accuracy Global Distance Test (GDT-HA), and the Template Modeling Score (TM-Score) (Table 1).

Table 1. Average scores for 63 testing set proteins generated by the RGN (least-squares) and Likelihood-RGN loss functions. The two neural networks here were trained on a dataset consisting only of X-ray structures.

	dRMSD	RMSD	GDT	GDT-HA	TM-Score
Likelihood-RGN	8.80	14.80	0.181	0.085	0.300
RGN	8.95	15.34	0.169	0.079	0.284

The same procedure was used to train models from the full ProteinNetX dataset (i.e., including multi-model NMR structures) generated from the CASP12 ProteinNet. Using the full dataset, the maximum likelihood model again showed increased stability during initial training (Figure 3b). The final trained models also outperformed the original model, though to a lesser extent than when using only the X-ray crystallography structures (Table 2). The original least- squares RGN consistently trains and converges to a minimum that predicts global protein topology quite well, however, the minimum predicts unphysical torsional angles that result in an accordion- like backbone fold. Importantly, when using a least-squares loss and the full ProteinNetX, our Likelihood-RGN trains to an alternate minimum that provides more realistic torsional angles in the protein predictions, which is evident in Ramachandran plots (Figure 5) and in the Ramachandran outlier and favored torsion evaluations (Table 2). The addition of B-factors to the dataset and implementation of a maximum likelihood loss function makes this alternate minimum accessible, ultimately resulting in more physically realistic structures that are amenable to downstream physics-based optimization. Structures for six selected CASP12 targets demonstrate that Likelihood-RGN achieves a smaller RMSD to the known protein fold when compared to the RMSD achieved by the least-squares RGN (Figure 4).

Interestingly, though adding NMR structures to the training dataset improved the backbone torsions predicted by the final model, the global distance metrics were slightly worse than those predicted by the models trained using only X-ray structures. This suggests that the representation of NMR models in the dataset could be improved (e.g., currently only the coordinates of the first NMR model from experiment are included in the dataset) to improve on global distance metrics in addition to

predicted backbone torsions. Future work could include investigating alternative methods to represent NMR coordinates and their uncertainty.

All code for this work and our two trained Likelihood-RGNs (with and without NMR data) are publicly available (https://github.com/mallory-tollefson/rgn).

Table 2. Average scores for 63 testing set proteins generated by the RGN (least-squares) and Likelihood-RGN loss functions across five separate training trials. The two neural networks here were trained on the full ProteinNetX dataset consisting of X-ray and NMR protein structures.

	dRMSD	RMSD	GDT	GDT-HA	TM-Score	Outlier Torsions	Favored Torsions
RGN	8.95	15.42	0.162	0.075	0.277	58.9	21.4
Likelihood-RGN	9.01	15.04	0.169	0.079	0.286	41.4	40.3

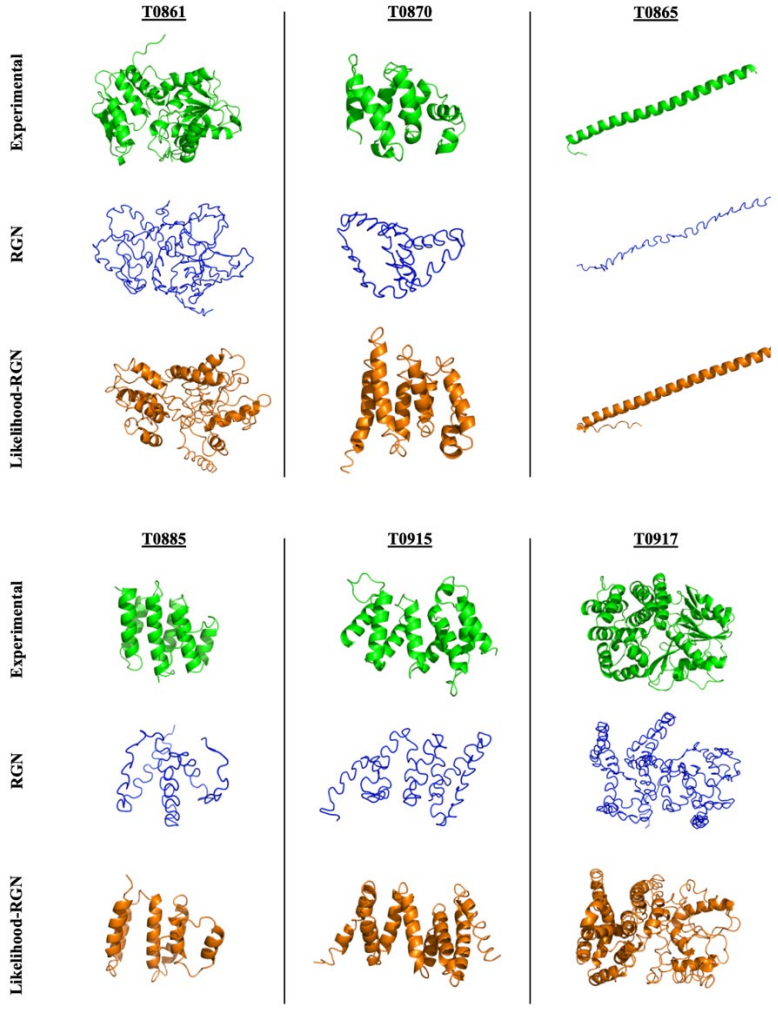


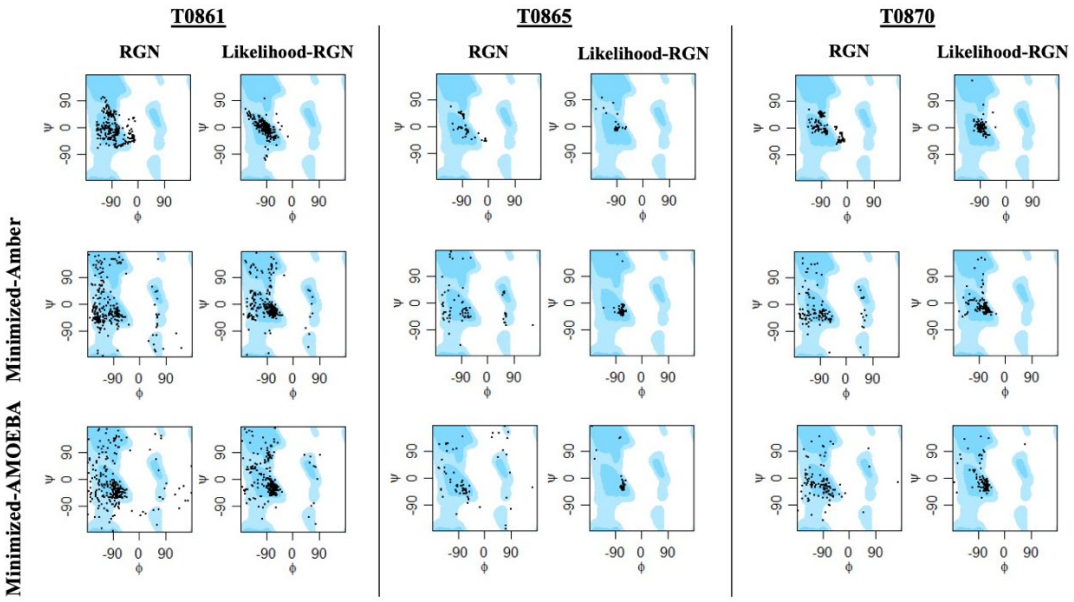
Figure 4. Six targets from the CASP12 competition shown in their experimentally solved coordinates (green), the original RGN predicted coordinates (blue), and the Likelihood-RGN predicted coordinates (orange) from the training trial that provided the best backbone torsions. The Likelihood-RGN structures have a smaller RMSD to the known experimental fold compared to the original RGN.

C. Physics-Based Optimization of Predicted Backbone Structures

Each of the 63 proteins in the testing dataset for both the maximum likelihood loss and original least-squares loss were minimized using the AMOEBA polarizable and Amber fixed charge force fields with a Generalized Kirkwood implicit solvent. Our Likelihood-RGN and RGN networks predicted structures with better torsional angles after training on our full ProteinNetX dataset (i.e., the dataset containing both X-ray and NMR structures), therefore, the protein structures optimized here were initially created from the Likelihood-RGN that was trained on our full ProteinNetX dataset. Minimization on the testing set demonstrates that proteins generated from Likelihood-RGN are more amenable to physics-based simulation because the RMSD, GDT, GDT-HA and TM-Score each decrease a smaller amount than in the least-squares RGN proteins (Table 3). Favored and outlier torsions increase and decrease, respectively, by a larger percentage with the least-squares RGN testing set than with the Likelihood-RGN testing set, but this result is likely an artifact of the significantly worse starting proportions of favored and outlier torsions present in RGN compared to Likelihood-RGN. Ramachandran plots (Figure 5) for six proteins from the testing set show that Likelihood-RGN consistently has fewer torsion outliers and more favored torsions than RGN, both before minimization and after minimization with the AMOEBA and Amber force fields.

Table 3. Physics-based optimization of the 63 testing set proteins under least-squares and Likelihood-RGN averaged across five trials.

	RMSD	GDT	GDT-HA	TM-Score	Torsion Outliers	Torsion Favored
Likelihood-RGN	15.04	0.169	0.079	0.286	41.46	40.25
AMOEBA Minimized	16.19	0.152	0.070	0.261	21.80	55.54
Amber Minimized	16.01	0.152	0.070	0.261	16.03	57.76
RGN	15.42	0.162	0.075	0.277	58.95	21.39
AMOEBA Minimized	16.65	0.140	0.062	0.248	24.91	50.50
Amber Minimized	16.46	0.139	0.061	0.248	17.31	50.46



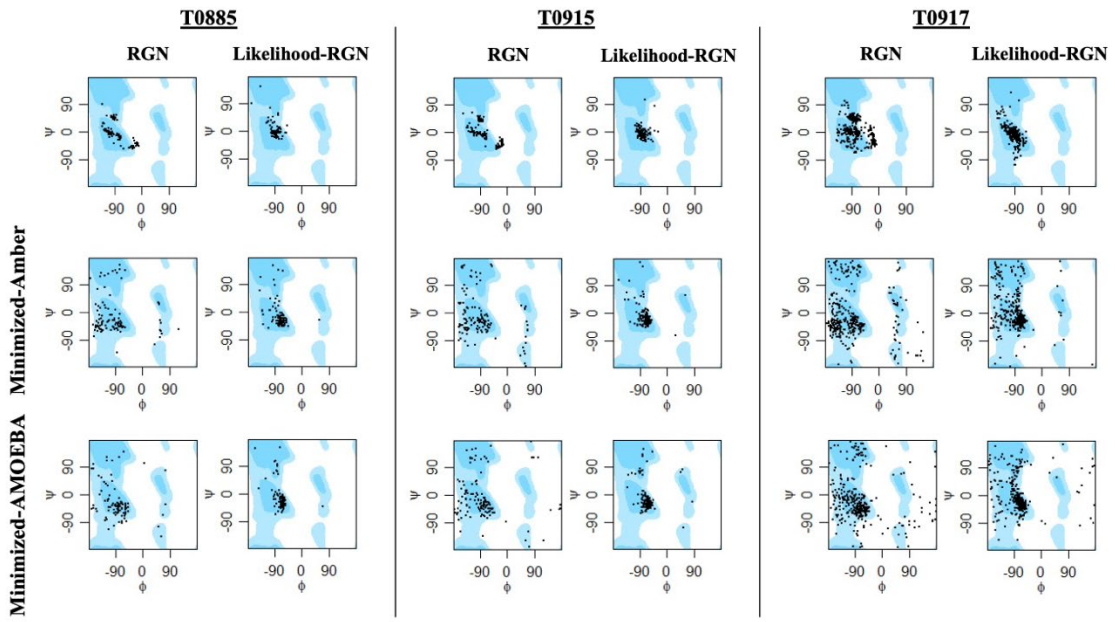


Figure 5. Ramachandran plots for six CASP12 targets for RGN and Likelihood-RGN. Ramachandran plots are also show for the six targets after minimization with the Amber and AMOEBA force fields.

Discussion

In this work, we described an improved dataset and loss function for deep learning approaches to protein structure prediction. We generated the ProteinNetX, which incorporates crystallographic B-factors into the existing DL protein structure prediction dataset, ProteinNet. For NMR structures, we computed B-factor equivalents from a per-atom RMSF over each NMR model, though future work includes improving upon NMR coordinate and uncertainty representation in ProteinNetX, along with developing a method to compute B-factors for single- model NMR and cryoEM structures.

By reformulating the loss function in the RGN model from a least squares target to a maximum likelihood target, we were able to incorporate these B-factors as experimental uncertainty values in model training. Our maximum likelihood model consistently improved network training over a series of trials, both in the initial stability of the training and in the structural metrics collected on backbones predicted by the final models. Our maximum likelihood network predicted protein models with more physically realistic backbone torsions, as the defined regions of secondary structure have smaller B-factors and contribute more to the calculated training loss.

These improvements in secondary structure predictions proved evident in physics-based optimizations. The backbones predicted by Likelihood-RGN were more amenable to physics- based simulations and retained their overall fold better than the structures predicted by the least- squares RGN after a series of energy minimizations and MD simulations. This suggests that a maximum likelihood loss model is better suited for downstream biophysical analyses, such as global backbone folding refinement like MELD global side-chain optimization or free energy perturbation of missense variations. Specifically, beginning physics-based protein folding from a backbone predicted by deep learning, rather than attempting to fold a protein *ab initio*, decreases simulation time.

Our reformulation of RGN's least-squares loss to a maximum likelihood is a novel approach in the effort to apply DL methods to protein structure prediction. Though our results do not attain AlphaFold's predictive accuracy, improvements to loss functions, training datasets, network architectures and optimization processes are necessary to improve the public effort at solving the protein folding problem.

Ideas/aims for future extramural project:

Future directions include predicting B-factors alongside protein coordinates in order to quantify the uncertainty in a DL coordinate prediction. The ability to quantify uncertainty in coordinate predictions would benefit downstream physics refinement by guiding refinement simulations toward improving lower confidence protein regions.

Publications resulting from project:

The paper was published in bioRxiv with title "**Protein Structure Prediction Using a Maximum Likelihood Formulation of a Recurrent Geometric Network**"

doi: https://doi.org/10.1101/2021.09.03.458873