1 Introduction

Approximately $\sim 40\%$ of the human proteome currently has resolved structural information, either through experimental methods or homology modeling. Understanding protein structures is essential for identifying therapeutic targets, enabling the development of effective drug delivery strategies. Traditional methods of predicting protein folds rely on physics-based models and molecular dynamics simulations. However, protein folding occurs in milliseconds (10^{-3} s), while Molecular dynamics simulations are generally constrained to timescales of microseconds (10^{-6} s) because of their significant computational demands. In contrast, deep learning techniques offer a time-efficient alternative for predicting protein structures from amino acid sequences. These methods are particularly valuable for downstream analyses where structural or homology information is unavailable. The potential of deep learning in protein structure prediction was exemplified by AlphaFold2, which achieved an impressive median Global Distance Test (GDT) score of 92.4 during the 14th Critical Assessment of Structure Prediction (CASP14), covering $\sim 98.5\%$ of the human proteome with high confidence. Another prominent model, Recurrent Geometric Network (RGN), constructs the protein backbone sequentially using predicted torsion angles. RGN was a top performer in CASP12, and its successor, RGN2, recently outperformed AlphaFold2 in certain benchmarks.

2 Hypothesis

Models like AlphaFold2, RGN, and RGN2 typically use a least-squares loss function to minimize the difference between predicted and actual protein structures, treating all structural parts as equally accurate. However, this approach overlooks variability in structure quality, particularly in regions with higher uncertainty or disorder. To address this, structural refinement methods that incorporate maximum likelihood principles have proven effective in accounting for experimental uncertainty and improving prediction accuracy. The incorporation of the maximum likelihood loss function into models like Recurrent Geometric Network (RGN) refines prediction accuracy by accounting for experimental uncertainty and physical realism. In fact, models like RGN2 have shown improved predictions, achieving good accuracy even in the case of orphan proteins. Furthermore, the utilization of maximum likelihood approaches allows the refinement of model training, where high-quality data has significantly improved the final structure, ultimately enhancing the overall reliability of protein structure prediction.

3 Methods and Approach

The ProteinNet database is a widely used machine learning resource for protein structural, functional, and evolutionary data. It was enhanced by incorporating atomic displacement parameters (ADPs), specifically B-factors from X-ray crystallography, and per-atom root mean square fluctuation (RMSF) data, with a scaling factor applied to RMSF to estimate B-factors in multimodal NMR, enabling training of the Likelihood-RGN. B-factors capture atomic vibrational motion and reflect positional uncertainty. This enhancement led to ProteinNetX, integrating structural and displacement data. Likelihood-RGN improves protein prediction by replacing distance-based root-mean-square deviation (dRMSD) with a maximum likelihood approach, modeling atomic positions as Gaussian distributions and utilizing ADPs (B-factors or RMSF) for uncertainty. Hyperparameters are consistent with the previous model. Model predictions are evaluated using geometric metrics like Global Distance Test (GDT), Global Distance Test High Accuracy (GDT-HA), Template Modeling Score (TM-Score), and favorable backbone torsions. Predicted structures obtained from RGN and maximum likelihood-RGN undergo physics-based optimization, first with Amber using the BFGS algorithm, then with AMOEBA forcefields. This reduces steric clashes and improves torsions. Structural quality is evaluated using geometric metrics before and after refinement. Final refinement employs GPU-accelerated many-body optimization with AMOEBA and Kirkwood sol-

vent, refining side chains and performing energy minimization. Quality is assessed with MolProbity for geometric accuracy.

4 Results

ProteinNetX includes $\sim 84.7\%$ of CASP12 ProteinNet structures, incorporating B-factors for crystallographic structures, which increases to $\sim 98.5\%$ with the addition of multimodal NMR structures and atomic displacement parameters (ADPs). A 20 Å shift to NMR aligns NMR parameters with crystallographic B-factors, creating a distribution like B-factors. Five pairs of models were trained on CASP12 ProteinNetX data, each pair sharing identical hyperparameters and random seed generation but differing in loss functions: one with least squares (dRMSD) and the other with maximum likelihood. The training process included 1.5 million iterations and 10,000 additional iterations at a reduced learning rate. Maximum likelihood improved convergence and accuracy by reducing the effects of disordered structures. When trained on the full dataset, including NMR structures, maximum likelihood showed increased stability and accuracy. Tested on 63 CASP12 structures, the models produced realistic protein structures, excelling in backbone torsion predictions. Maximum likelihood outperformed both in global accuracy and local secondary structure prediction. Global evaluation metrics included RMSD, DGT, DGT-HA, TM-scores, and backbone torsions. Torsion angles, particularly for alpha helices, also improved. Including NMR structures enhanced local predictions but slightly reduced global accuracy, as NMR focuses on local interactions, while X-ray crystallography captures both local and global interactions. When predicted structures from both models, RGN and Likelihood-RGN, were optimized using physics-based methods like Amber and AMOEBA, RMSD increased slightly, but Likelihood-RGN displayed more realistic torsion angles with fewer outliers compared to RGN. After minimization, Likelihood-RGN outperformed in geometry metrics, indicating better fold stability. Side-chain optimizations showed that both models improved MolProbity scores, with Likelihood-RGN achieving lower MolProbity and clash scores, indicating higher quality for biophysical simulations.

5 Discussion

ProteinNetX enhances protein structure prediction by integrating deep learning with crystallographic B-factors and NMR-derived atomic displacement parameters (ADPs). By shifting RGN loss function to maximum likelihood, the model generates more realistic torsion angles and improves secondary structure prediction, as it accounts for experimental uncertainties, leading to enhanced stability, convergence, and accuracy. Likelihood-RGN structures better maintain their fold than RGN structures during physics-based optimizations, demonstrating their potential for efficient downstream refinement. Additionally, ProteinNetX facilitates the integration of experimental constraints into generative models, enabling the prediction of structures with higher fidelity to physical and experimental observations. Comparative analyses reveal that incorporating B-factors and ADPs not only boosts model robustness but also aids in identifying regions with dynamic flexibility or disorder, which are critical for understanding protein function. Future work will refine NMR representations, predict B-factors from coordinates, and enhance protein structure prediction, advancing models like RGN2 and AlphaFold2 towards solving complex protein folding problems and enabling scalable protein design.

Reference

Qi, G., Tollefson, M. R., Gogal, R. A., Smith, R. J. H., AlQuraishi, M., & Schnieders, M. J. (2021). Protein structure prediction using a maximum likelihood formulation of a recurrent geometric network. *bioRxiv*, *2021*, 458873. https://doi.org/10.1101/2021.09.03.458873