A Physics-informed Inverse Folding Model for Protein Free Energy Prediction

Zhaoyang Li
Department of Bioengineering
Stanford University
Stanford, CA 94305
zhaoyangli@stanford.edu

Pengwei Sun
Department of Radiology
Stanford University
Stanford, CA 94305
pengwei@stanford.edu

1 Motivation

Inverse protein folding (protein sequence design) seeks a sequence that folds to a given backbone. This is a central challenge in protein engineering, with applications ranging from enzyme design to vaccine development. Modern deep learning models like ProteinMPNN [1] have achieved breakthrough performance on this task, substantially outperforming traditional physics-based methods (e.g. Rosetta) in native sequence recovery (52.4% vs 32.9% for ProteinMPNN vs. Rosetta).

Beside its success on protein sequence design, recent studies also revealed that the log-likelihood predicted by the model is highly correlated with the free energy of the folded protein, and there is an opportunity to inject more physical context into the model. In particular, side-chain packing and solvent exposure that strongly influence protein stability and binding are not explicitly modeled by using only backbone atoms. We propose to extend graph-based protein design models with more physically meaningful structural features to improve generalization to stability and binding tasks. Our project will improve a ProteinMPNN-style model that leverages side-chain geometry, improved neighbor definitions, and occupancy (burial) features. We will evaluate it on inverse folding as well as (zero-shot) free energy prediction using large-scale experimental datasets. Ultimately, our goal is to create a GNN-based model that not only excels at the core sequence design task, but also generalizes to predict experimental measures of stability and protein-protein interaction strength.

2 Datasets and Tasks

We will pre-train the model on the Protein Data Bank (PDB) dataset [2] as what is done in ProteinMPNN [1] (and other inverse folding models like ESM-IF1 [3]), and then fine-tune it on the two public datasets to benchmark the model's performance on stability and binding generalization. We will use the official architecture and training recipe as a strong starting point.

1. MegaScale stability dataset: a "mega-scale" experimental dataset of protein folding stabilities [4]. This dataset contains 1.84 million measurements of folding free energy (ΔG) or stability changes ($\Delta\Delta G$) across 10^3 diverse proteins. We will primarily use the high-quality single-mutation subset (776k data points) for quantitative evaluation. We will use the public release on Hugging Face [5].

2. SKEMPI v2.0 binding affinity dataset: a comprehensive database of mutations in protein–protein interfaces with measured binding free energy changes [6, 7]. SKEMPI v2.0 includes 7,085 mutations (single and multiple) in various complexes. We will focus on single-point mutations (over 4,000 entries after quality filtering), using the reported $\Delta\Delta G$ as ground truth for binding affinity change.

Using these datasets, we plan to evaluate our model on three fronts: (i) *Sequence recovery* (%), (ii) *Stability prediction*: Pearson r and RMSE between model-derived scores and experimental $\Delta\Delta G$ on MegaScale, and (iii) *Binding affinity prediction*: RMSE (kcal/mol) and rank correlations for SKEMPI $\Delta\Delta G$.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

3 Models and Methods

Protein energetics is driven by geometry-dependent, many-body interactions (packing, burial, and side-chain orientations). Distance-cutoff graphs, virtual-center features [8], and exposure encodings directly reflect these determinants, while equivariant GNNs [9, 10] align with the $\mathrm{SE}(3)$ symmetries of structure.

Given a protein backbone graph $G = (V, E)$ with residue nodes $i \in V$ and geometry-derived edge features, learn $p_\theta(\mathbf{a} \mid G)$ over sequences $\mathbf{a} = (a_1, \ldots, a_{|V|})$. Evaluation also uses zero-shot scoring: for a structure and mutant at site k, define a sequence score $s_\theta(\mathbf{a} \mid G)$ (e.g., $\log p_\theta$ or an energy head); correlate Δs_θ for WT→mutant with experimental $\Delta\Delta G$.

Baseline. ProteinMPNN [1], a message-passing neural network implemented in PyTorch Geometric (PyG) [11]. ThermoMPNN (ProteinMPNN features $\to \Delta\Delta G$ via transfer learning) [12] for stability prediction.

Physics-informed extensions (novelty). ProteinMPNN originally used backbone distances and orientations (pairwise C_α distances, frame angles, etc.) as input features. We will incorporate additional features to augment this representation:

1. **Connectivity by distance cutoff.** Replace fixed k-NN with edges for residue pairs whose C_α distance is below a threshold (e.g., $d_c \in [8, 10]$ Å). This matches the biophysical notion that interaction strength decays with Euclidean distance and yields degree heterogeneity consistent with core vs. surface residues.

2. **Side-chain virtual centers.** Augment nodes with a side-chain virtual center V constructed from backbone geometry as in Foldseek [8]. Virtual centers are defined by angle $\theta(V - C_\alpha - C_\beta)$, dihedral $\tau(V - C_\alpha - C_\beta - N)$, and length $l = |V - C_\alpha|$; the optimal parameters are $\theta = 270^\circ$ $(3\pi/2)$, $\tau = 0^\circ$, $l = 2 \times 1.53$ Å. We add edges/features using V–V and (C_α, V) distances to inform side-chain packing.

3. **Occupancy (solvent exposure).** Define a differentiable occupancy for node i:

$$\mathrm{Occ}(i) = \sum_{j \neq i,\ d_{ij} < r} w_j \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \quad S(i) = \frac{\mathrm{Occ}(i) - \mathrm{Occ}_{\min}(t_i)}{\mathrm{Occ}_{\max}(t_i) - \mathrm{Occ}_{\min}(t_i)}, \tag{1}$$

   where d_{ij} are V–V distances, t_i is residue type, and w_j are per-residue weights (e.g., approximate side-chain volumes). $S(i)$ encodes burial vs. exposure to guide hydrophobic/polar preferences.

Architectural exploration. In addition to the ProteinMPNN message-passing encoder, we may evaluate (i) GVP-GNN layers that pair scalar/vector features with rotation-aware updates [9], and (ii) EGNN layers providing $\mathrm{E}(n)$-equivariance with lightweight coordinate updates [10]. We will keep the order-agnostic autoregressive decoder of ProteinMPNN [1], optionally conditioning it on the proposed features.

References

[1] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Han Bai, Ryan J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Antoine Courbet, Richard J. de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

[2] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[3] Ching-Hui Hsu, Robert Verkuil, Jiayi Liu, Zeming Alan Lin, Brian Hie, Tom Sercu, Alexander Rives, et al. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *PMLR*, pages 8946–8970, 2022.

[4] Kotaro Tsuboyama, Weijun Chen, and et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620:434–444, 2023.

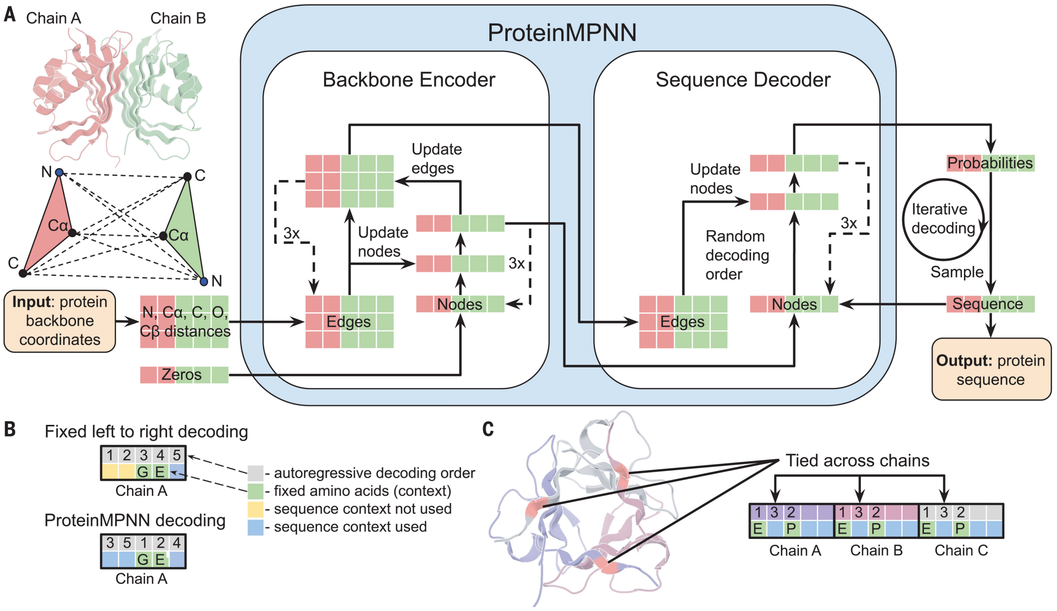

Figure 1: ProteinMPNN architecture (encoder–decoder). Adapted from [1]. We extend the encoder with distance-cutoff connectivity, virtual-center side-chain features, and occupancy signals.

[5] RosettaCommons. RosettaCommons/MegaScale (hugging face dataset). `https://huggingface.co/datasets/RosettaCommons/MegaScale`, 2024. Accessed: 2025-10-21.

[6] Justina Jankauskaitė, Benoît Jiménez-García, Jurgis Dapkūnas, Juan Fernández-Recio, and Iain H. Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

[7] Protein Interactions and Docking Group. SKEMPI v2.0 website. `https://life.bsc.es/pid/skempi2/`, 2019. Accessed: 2025-10-21.

[8] Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42:243–246, 2024.

[9] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron O. Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pages 9323–9332, 2021.

[11] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019.

[12] Henry Dieckhaus, Michael Brocidiacono, Nicholas Z. Randolph, and Brian Kuhlman. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121(6):e2314853121, 2024.