Inverse Folding with Protein Language Model

Mahan Naseri†, Xiao Wang†

Abstract: We propose an integrated framework for protein inverse folding that combines two modalities: a protein sequence and a protein structure backbone image. The sequence modality is processed by a pretrained protein language model (PLM) trained autoregressively via multi-head latent attention on millions of sequences, while the structure modality is transformed into rich, informative images. These images are generated from 3D backbone coordinates (N, Cα, C, Cβ) into distograms that capture inter-residue distances and angles, embedding essential differential geometry. By using image representations, we avoid the need for equivariant methods, as the relative geometry is already encoded in the image channels, and leverage the scalability of vision transformers. Our architecture interleaves pretrained GPT blocks with cross-attention blocks that condition the language model on multiscale image features, facilitating efficient and high-fidelity sequence generation. We also incorporate repetition and n-gram penalties during sampling for improved diversity and quality, while latent attention provides parameter efficiency and fast inference.

Background: Inverse folding is the computational challenge of determining an amino acid sequence that will fold into a given three-dimensional protein structure backbone. Essentially, it reverses the conventional protein folding problem by starting with a desired backbone and asking, "What sequence will reliably form this structure?" Traditionally, this task relied on physics-based methods that evaluated sequence–structure compatibility through energy functions. However, these methods were often computationally intensive and struggled with the complexity of natural protein interactions.

Recent advances in deep learning have revolutionized inverse folding by introducing deep generative models. These models, which include graph neural networks, transformers, and diffusion models, learn the complex relationship between a protein's structure and its sequence from large datasets. By capturing geometric details and evolutionary patterns, deep generative models can:

1. Generate diverse and high-quality protein sequences conditioned on a target structure.
2. Achieve significantly higher sequence recovery rates compared to traditional methods.
3. Accelerate the design process, paving the way for engineering novel proteins with desired functions.

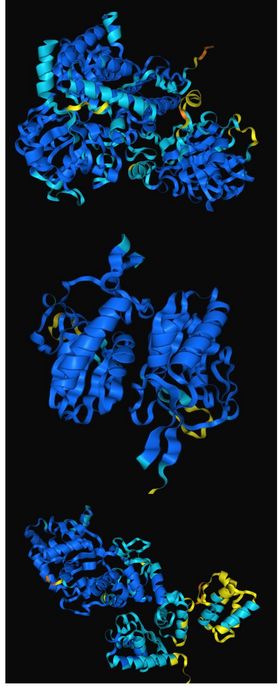
This synergy of deep learning and protein design is pushing the boundaries of de novo protein engineering, making it possible to design proteins that not only fold correctly but also exhibit novel or enhanced functionalities.

Supervised Fine-Tuning via Prompt Engineering for PPI Design

We adapt our pretrained protein language model to generate binding partners through supervised fine-tuning using prompt engineering. In this approach, interacting protein sequences are fused into a single input, where Protein A serves as the prompt and Protein B is generated as the complementary binding partner. For example, an input is formatted as:

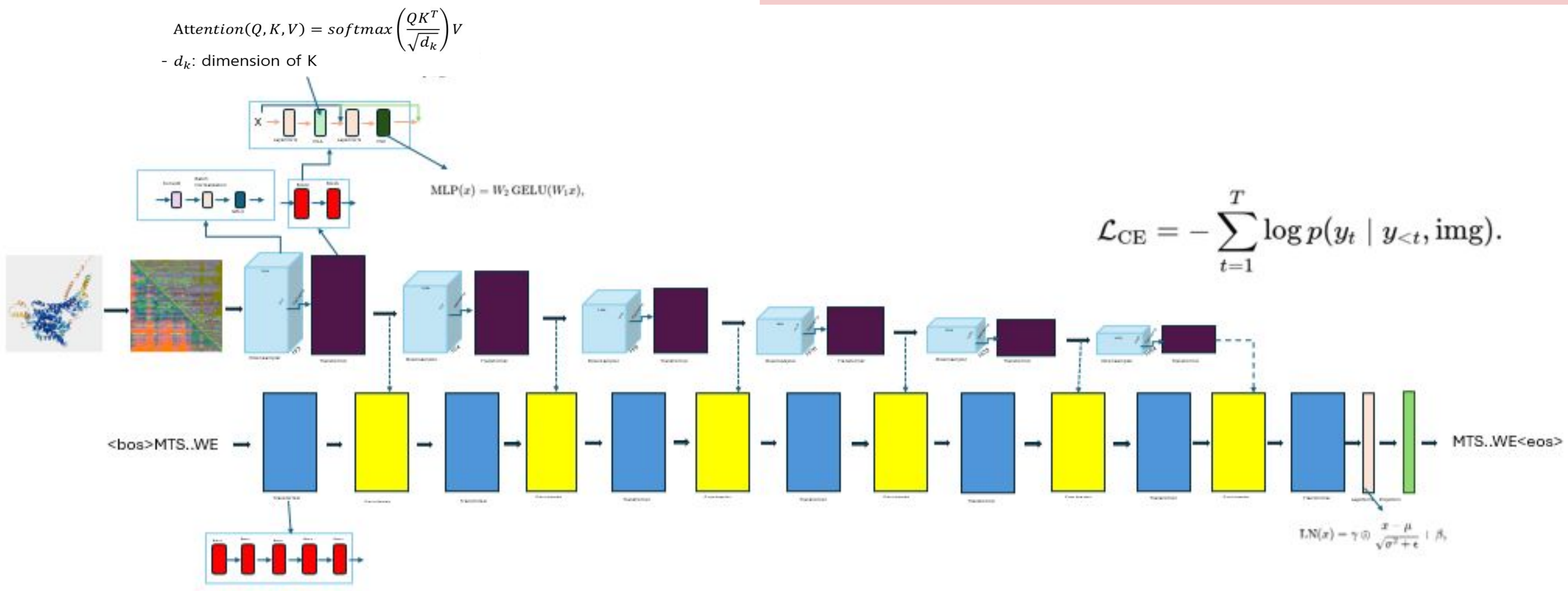
[BOS] Protein A [EOS] [SEP] [BOS] Protein B [EOS]

During training, the model minimizes the cross-entropy loss to predict the continuation (Protein B) based on Protein A, learning the contextual dependencies necessary for protein–protein interaction (PPI) design. We leverage both FASTA and curated PPI datasets to create these fused sequence examples. Additional sampling penalties—such as repetition and n-gram blocking—ensure diversity and prevent degenerate outputs. This supervised fine-tuning strategy effectively tailors the PLM to generate high-quality binding partners for a given protein target, significantly enhancing its performance on PPI design tasks.



$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- d_k: dimension of K



Model Architecture: This figure illustrates our multimodal architecture for protein inverse folding. On the left, raw protein structures from PDB files are processed: non-standard residues are remapped, and backbone atoms (N, Cα, C) are extracted. These coordinates are transformed into a "distogram" image—where each channel encodes a different geometric feature (normalized distances, dihedral angles, bond angles)—thus capturing relative spatial relationships without requiring explicit equivariance.

The image then flows into a multi-level Vision Transformer encoder. At each of six downsampling stages, overlapping convolutional layers reduce the spatial dimensions while transformer blocks capture increasingly abstract representations of the protein backbone. These multiscale features are then used as conditioning tokens.

On the bottom, a pretrained protein language model (a GPT-based network) processes tokenized protein sequences. Its transformer layers are interleaved with cross-attention blocks that integrate the structure's latent features via scaled dot-product attention. In this way, the model conditions sequence generation on the underlying backbone structure.

Finally, the output logits predict the next amino acid in an autoregressive fashion. During sampling, additional penalties (for repeated tokens and n-gram redundancy) further refine the generated sequence. Overall, the figure captures the complete data flow—from raw 3D coordinates to image generation, feature extraction via a Vision Transformer, and sequence generation with cross-modal conditioning—demonstrating an efficient, scalable approach to protein design.

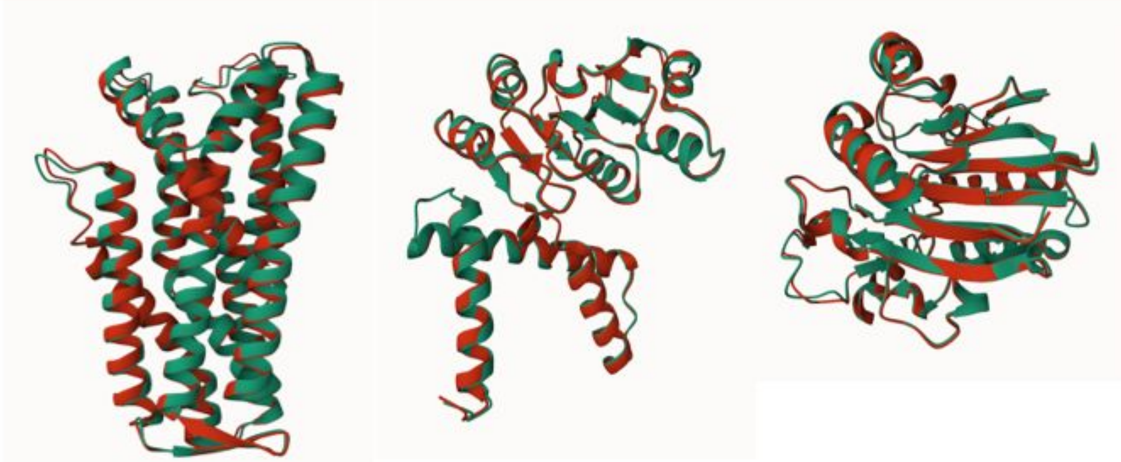
Performance & Results

This figure demonstrates our model's evaluation on three challenging protein targets (from the MGnify dataset, comparing the **actual** structure (green) with our **generated** design (red). The table above shows three metrics:

1. **pTM** (Predicted TM-score) – Reflects how well the designed sequence folds into a structure that aligns with the native fold (higher is better). Our pTM values (~0.88–0.93) indicate strong structural agreement.
2. **RMSD** (Root-Mean-Square Deviation) – Measures the average atomic displacement between the native and generated structures (lower is better). Our RMSDs (0.65–0.97 Å) confirm close backbone alignment for large proteins.
3. **Sequence Recovery** – Percentage of amino acids in the designed sequence that match the native sequence (higher is better). We achieve up to ~73% recovery, surpassing typical benchmarks on proteins of this size.

Collectively, these results demonstrate **high structural fidelity and strong sequence accuracy** for challenging, large proteins. Our method shows a notable improvement over existing approaches, especially in the 300–500 residue range, highlighting the effectiveness of our inverse folding architecture on more complex protein targets.

IF-GPT Evaluation	MGYP003146536281	MGYP004445122525	MGYP001500075107
pTM ↑	0.93	0.92	0.88
RMSD ↓	0.65	0.87	0.97
Sequence Recovery ↑	73.4	69.3	67.2



†: School of Biological and Health Systems Engineering, Arizona State University