A Unified Protein Embedding Model with Local and Global Structural Sensitivity

Introduction

Many research tasks depend on identifying structural homologs of proteins, including evolutionary analysis, peptidomimetics, and functional annotation. However, standard structural alignment algorithms often utilize atomic distance matrix alignment, a quadratic-time operation. My research utilizes PLMs (protein language models) to speed up structural comparison. In addition, by producing both overall protein embeddings and per-residue embeddings, as well as utilizing a dual global-local loss function, my model is able to capture both global and local structural similarity simultaneously.

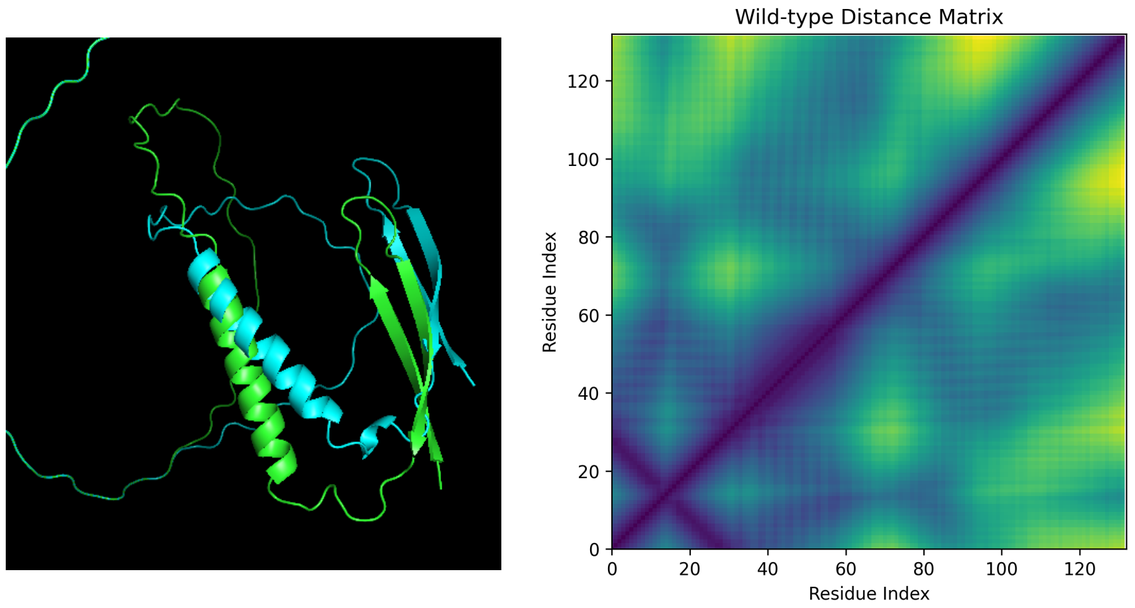

Figure 1: Structure of p14ARF and a common mutant (left), and the wild-type C_α distance matrix (right). Graphics generated by the student researcher using PyMOL and Matplotlib, 2025 [1].

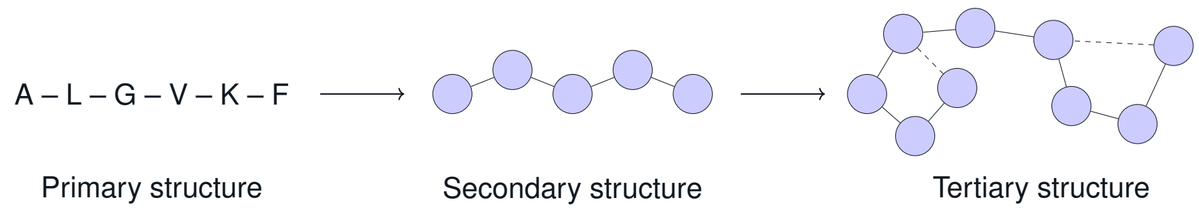

Figure 2: Depiction of the biophysical prior that sequence determine structure, utilized in my PLM. Graphic generated by the student researcher using TikZ, 2026.

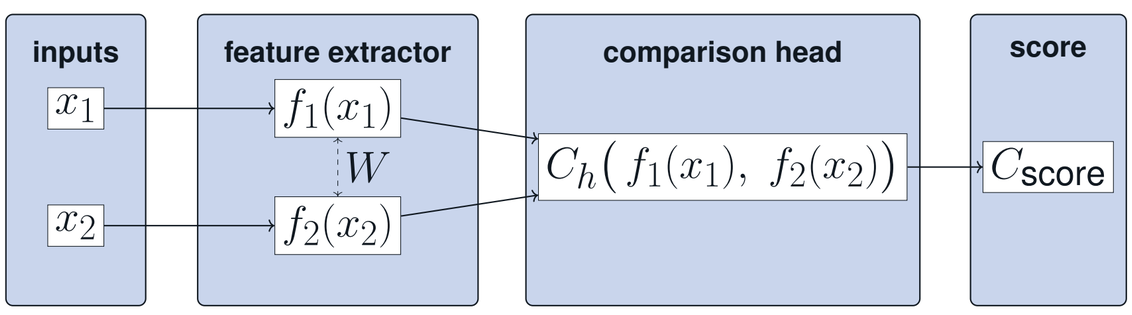

Figure 3: The typical architecture of a siamese neural network consists of a feature extractor (the two neural networks of equal weights) and a comparison head. Graphic generated by the student researcher using TikZ, 2025 [1].

Methodology

My PLM was designed as a transformer-based siamese neural network. It is able to generate both per-residue structural embeddings, and per-protein structural embeddings. Transformers are utilized to process data in parallel via self-attention, enabling the model to capture relationships across all positions in a sequence simultaneously. Siamese neural networks are an extension of the neural network architecture that allows for pairwise comparisons of inputs. My contrastive loss function combined local metrics (IDDT-scores) and global metrics (TM-scores) of structural similarity.

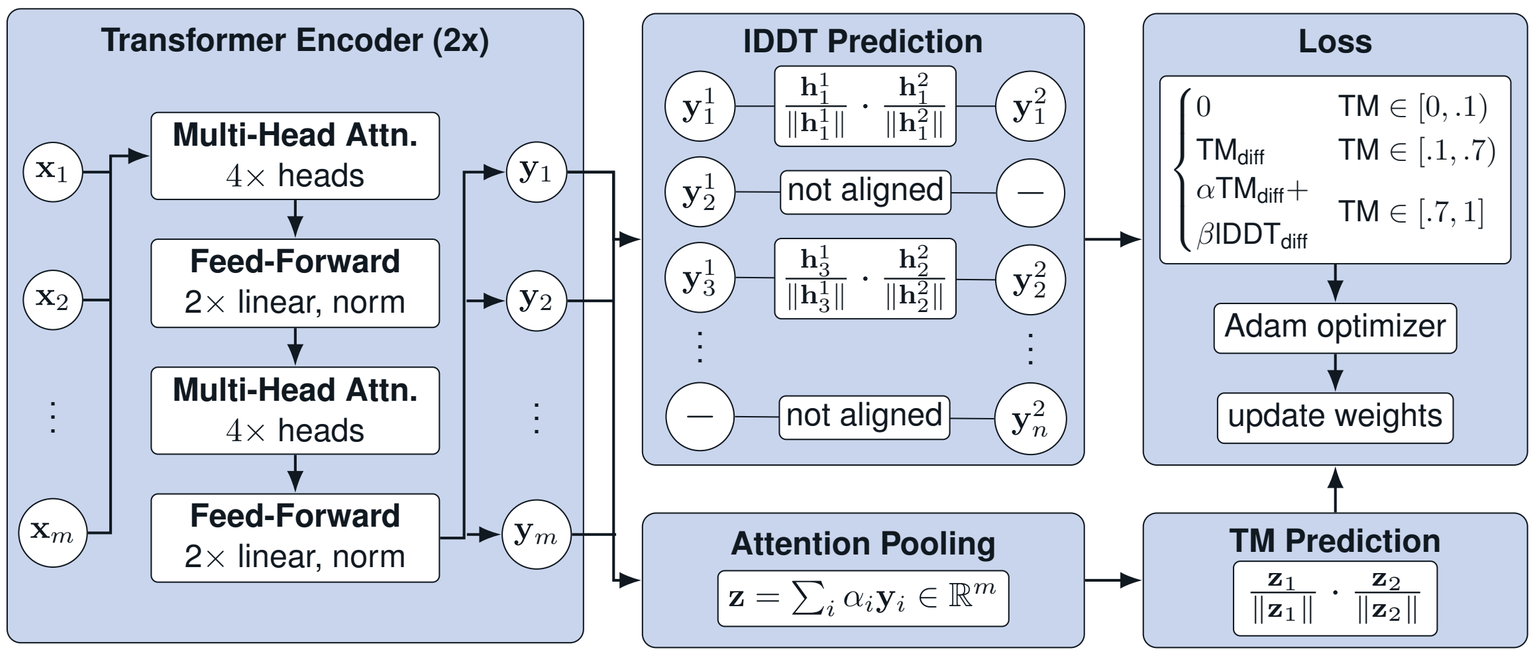

Figure 4: The architecture of a single NN in my PLM. Graphic generated by the student researcher using TikZ, 2025 [1].

Results

I tested my PLM on two datasets, designed to evaluate my PLM's capabilities at predicting both global and local structural similarity.

Metric	Model	MAE	MSE	Error Stdev
IDDT (per-residue)	My model	0.0788	0.0344	0.1224
TM (per-protein)	My model	0.0741	0.0103	0.1010
TM (per-protein)	TM-Vec	0.0792	0.0126	0.1113

Table 1: Performance of my model and TM-Vec on the TM-Vec dataset. Table generated by the student researcher using LaTeX, 2025 [1].

Metric	Model	MAE	MSE	Error Stdev
IDDT (per-residue)	My model	0.0038	0.0001	0.0095
TM (per-mutant)	My model	0.0583	0.0096	0.0980
TM (per-mutant)	TM-Vec	0.0617	0.0102	0.1011

Table 2: Performance of my model and TM-Vec on the VIPUR dataset. Table generated by the student researcher using LaTeX, 2025 [1].

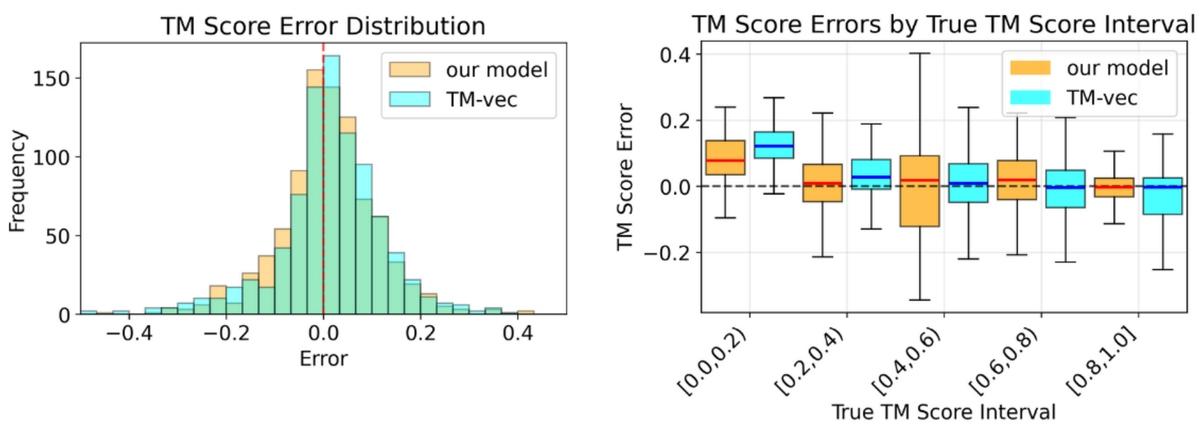

Figure 5: The TM-score error histogram and box plot on the TM-Vec dataset for both my model and TM-Vec. Graphic generated by the student researcher using Matplotlib, 2025 [1].

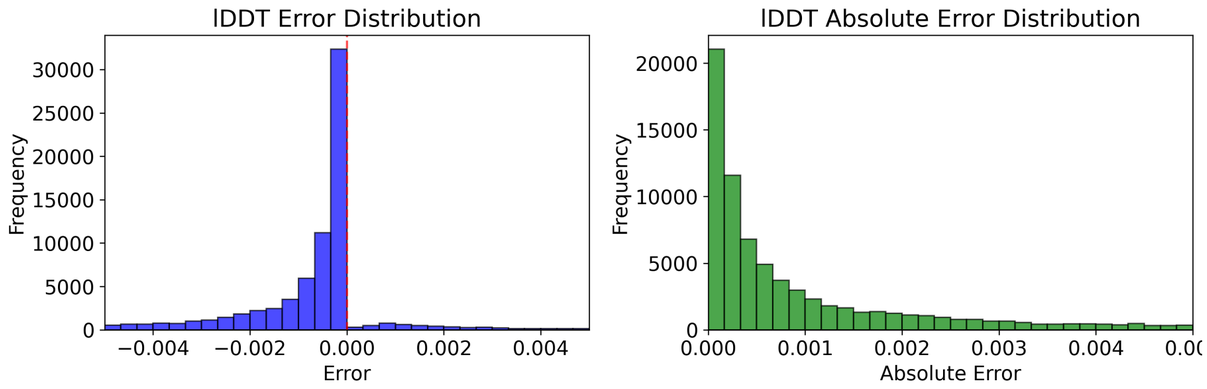

Figure 6: The IDDT-score error histogram on the VIPUR dataset for my model. Graphic generated by the student researcher using Matplotlib, 2025 [1].

[1] Jerry Xu, Shaojun Pei, and Gil Alterovitz. "A Unified Protein Embedding Model with Local and Global Structural Sensitivity". In: (Oct. 2025). DOI: 10.1101/2025.10.27.684815. URL: http://dx.doi.org/10.1101/2025.10.27.684815.

Xu.J