Assessing the Quality of Loops in Predicted Protein Structures

Florian Wieser*, Gustav Oberdorfer

Institute of Biochemistry, Graz University of Technology, Austria

1. Introduction

In contemporary protein design, advanced machine learning algorithms are used to generate novel amino acid sequences with unprecedented functionalities. Examples of which include novel enzymes with enhanced catalytic rates or new reactivities. Usually, the designed sequences are assessed *in-silico* through advanced protein structure prediction methods, such as AlphaFold2 (AF2) or RosettaFold2. Despite the success of these deep learning models in predicting rigid protein structures, the prediction of loop regions, critical for function, flexibility and specificity in proteins, remains challenging due to their unstructured nature.

To this end, we introduce the Equivariant Loop Evaluation Network (ELEN), a deep learning based local model quality assessment (MQA) method, specifically designed to assess the quality of protein loops. The network utilizes 3D equivariant group convolutions to map the local geometric environment of each atom. By incorporating sequence embeddings from large language models (LLM), such as Meta's Evolutionary Scale Model 2 (ESM-2) or SaProt, the model gains insights into sequence variation and evolutionary patterns. Moreover, by incorporating physicochemical features such as solvent accessible surface area (SASA), Rosetta energies, and partial charges, the network achieved competitive accuracy relative to the consensus method ModFOLD9 [1] and outperformed other state-of-the-art model quality assessment (MQA) methods in accuracy on a CAMEO dataset over a three-month period.

Although ELEN was primarily developed for assessing loop quality, its architecture also demonstrates promising potential for general MQA tasks, ultimately allowing per residue accuracy assessments. We believe that this method will be of critical use in the study and engineering of flexible regions in protein structures.

ELEN's Network Architecture

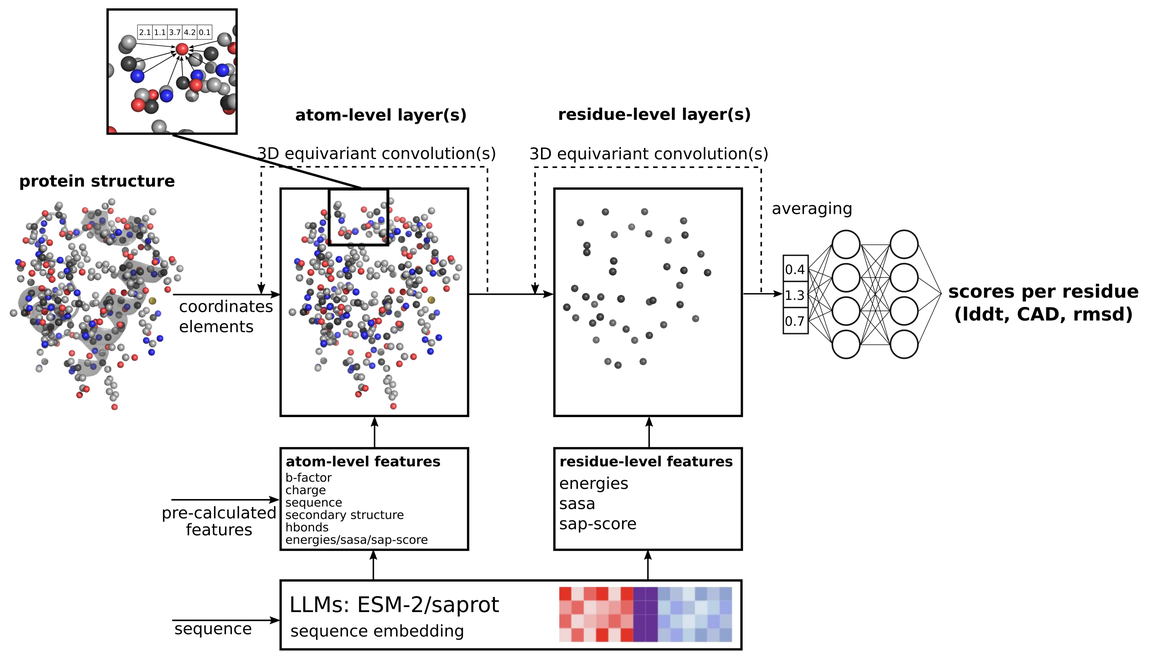
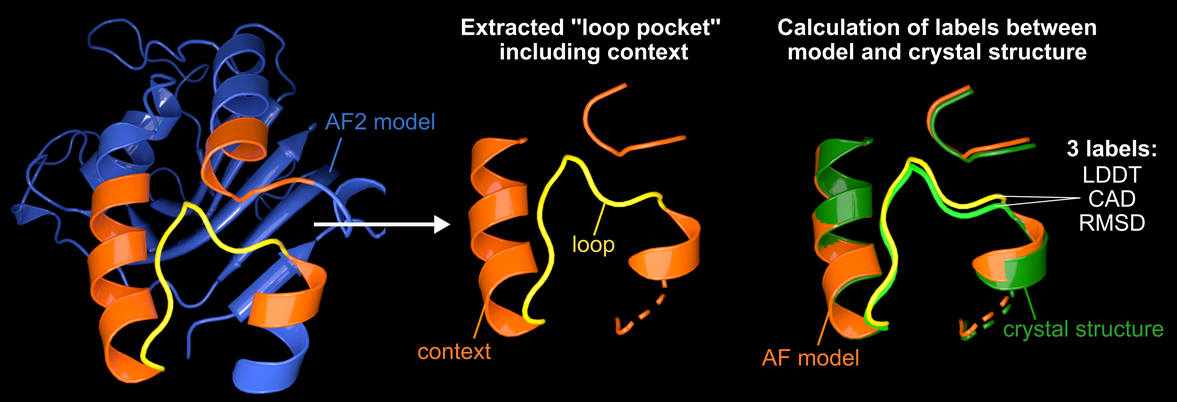


Fig. 1: ELEN's network architecture is designed as follows: Initially, the model inputs atomic coordinates and element types, applying multiple rounds of equivariant 3D convolutions to capture the local geometry around each atom. Features from individual atoms are then aggregated at the C-alpha carbons, followed by another round of equivariant convolution. In parallel, both atom and residue features are enhanced with sequence embeddings from LLMs and physicochemical properties such as SASA and Rosetta energies. These enriched features are averaged and processed through a dense neural network. The model outputs three scores for each residue —LDDT, CAD-score, and RMSD and compares them with the ground truth labels.

Dataset Preparation



2. Methods

An image of the Protein Data Bank (PDB) from September 2023 was filtered for high-quality crystal structures using the PISCES server. For each of the resulting sequences, five AF2 models were predicted using localcolabfold. Loop positions were then detected by computing the secondary structure via DSSP algorithm. Loops were subsequently extracted as so-called "loop pockets" [2] - these include not only the loop residues itself, but also their broader geometric context (compare with left figure). Finally, residue-wise LDDT (local Distance Difference Test), CAD-score (Contact Area Difference score) and RMSD (Root Mean Square Deviation) were computed between the extracted loops of the AF2 models and the respective loops in the native crystal structures. These metrics represent the distance of the computational model to the ground truth crystal structure and are hence used as training labels.

ELEN's network architecture is based on Eismann et al. [3] and is depicted in Figure 1. The currently best-performing ELEN model has 341,000 parameters and was trained on a dataset of approximately 1 million computational loop models for three epochs on an NVIDIA A100 graphics card with 40GB of RAM. The learning rate was set to 0.0001 with a batch size of 32.

3. Results

ELEN's predictive performance was tested on an MQA dataset from the CAMEO online server over a three-month period (December 2023). The correlation between the predicted scores from ELEN and the corresponding ground truth LDDT values is comparable to that of the consensus method, ModFOLD9, and is at least 0.3 higher for Pearson and 0.2 higher for Spearman correlations compared to all other methods (see Figure 3). In global MQA (picking the best model of an ensemble), the top performers were ModFOLD9 and ProQ3, with average top1 losses of 0.620 Å and 0.626 Å, respectively. ELEN ranked third, with a loss of 0.626 Å.

To better understand the model's ability to learn specific loop features, we performed principal component analyses (PCA) on the activation values from the penultimate layer of the network. The resulting data were color-coded by relevant loop features, as illustrated in Figure 2. For several features, a clear separation of colors could be observed.

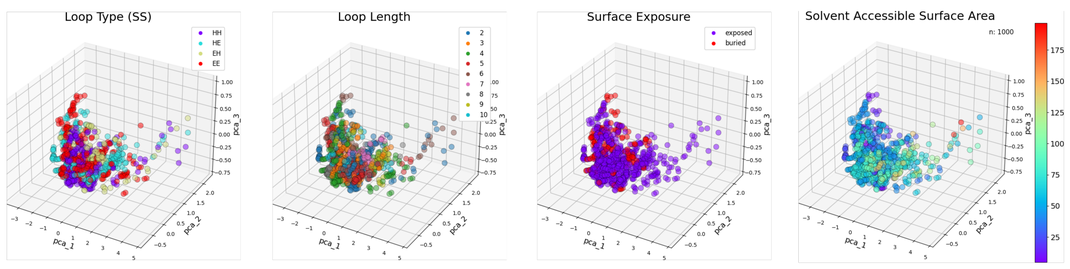


Fig. 2: PCA of activation values from the penultimate layer, color-coded by loop type (based on adjacent secondary structures), loop length, surface exposure (buried or exposed), and SASA.

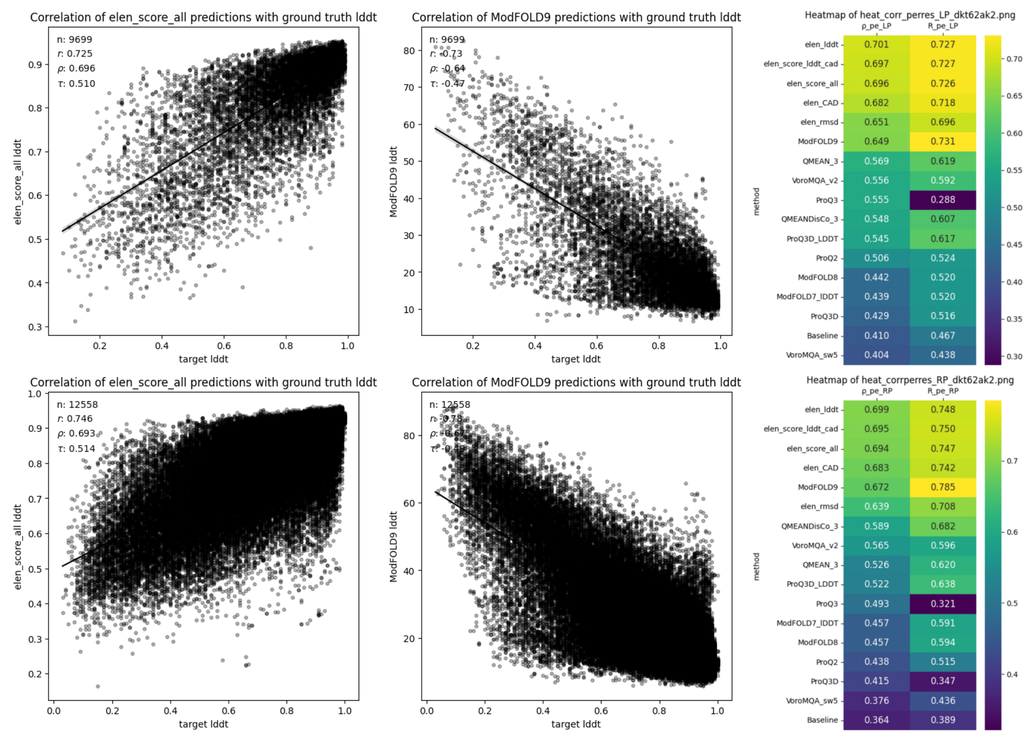
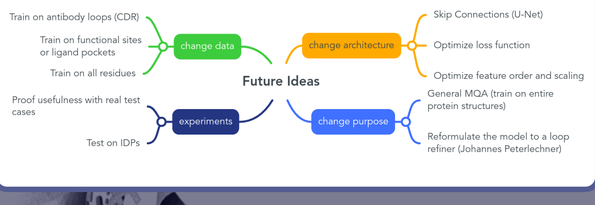


Fig. 3: Correlation analysis on a CAMEO dataset over a three-month period. Top row - Only loop residues were scored. Bottom row - All residue was scored. Left and mid column - Correlation of the combined ELEN score and the ModFOLD9 score with the ground truth labels. Right column - Color-coded Spearman and Pearson correlation for all methods.

Outlook



4. Discussion

To the best of our knowledge, ELEN is the first MQA method specifically developed for unstructured protein regions. Although not yet fully optimized, the model has already achieved results comparable to the consensus method, ModFOLD9, and has outperformed other MQA methods in a correlation analysis. Motivated by these promising results, a general MQA model (inclusive of all secondary structures) is currently being developed. Additionally, a Master's student is currently adapting the model to a loop refinement protocol.

***e-mail:** florian.wieser@tugraz.at
Rosetta slack: FlorianWieser

References:

[1] McGuffin et al., JMB (2024)
[2] Wang et al., SPS (2023)
[3] Eismann et al., Proteins (2023)