EmbedSimScore: Advancing Protein Similarity Analysis with Structural and Contextual Embeddings

Gourab Saha[*]
Bangladesh University of Engineering and Technology
gourabsaha567@gmail.com

Md Toki Tahmid[*†]
Bangladesh University of Engineering and Technology
sharifulislamtoki@gmail.com

Md. Shamsuzzoha Bayzid
Bangladesh University of Engineering and Technology
shams_bayzid@cse.buet.ac.bd

Abstract

Accurately computing protein similarity is challenging due to the intricate interplay between local substructures and the global structure within protein molecules. Traditional metrics like TM-score often focus on aligning the global structures of the proteins in a rather geometry-based algorithmic way, potentially overlooking critical local-global relations and contextual comparisons. We introduce EmbedSimScore, a novel self-supervised method that generates structural and contextual embeddings by jointly considering both local substructures and global proteins' structures. Utilizing contrastive language-structure pre-training (CLSP) and structural contrastive learning, EmbedSimScore captures comprehensive features across different scales of protein structure. These embeddings provide a more precise and holistic means of computing protein similarities, resulting in the identification of intrinsic relations among proteins that traditional approaches overlook.

1 Introduction

Accurately computing protein similarities is a fundamental challenge in computational biology, with significant implications for understanding protein function, evolution, and interactions. Traditional metrics such as TM-score [1] focus primarily on global structural alignments, which may overlook subtle yet critical local structural features and contextual information encoded in protein molecules. These limitations hinder the ability to fully capture the multifaceted nature of protein similarities, particularly when proteins share functional similarities despite low sequence or structural identity.

Recent advances in self-supervised learning have demonstrated the potential of deep learning models to extract meaningful representations from large datasets without explicit labels [2, 3, 4], including domains of biology[5, 6, 7]. In the context of proteins, self-supervised models have been employed to learn from amino acid sequences [8] and structures [9], but often focus on either the global structure or the sequence alone, without effectively integrating local structural nuances and contextual information.

[*] These authors contributed equally. Authors listed alphabetically.
[†] Corresponding author.

NeurIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice.

In this work, we introduce **EmbedSimScore**, a novel self-supervised method designed to generate structural and contextual embeddings for proteins by jointly considering local substructures, global architecture, and sequence context. Our approach leverages a combination of techniques to enhance the capture of protein similarities:

- **Structural Alignment of Multiscale Subgraphs**: We adapt the self-supervised knowledge distillation framework introduced by DINO [10] for protein graphs to align representations of local and global subgraphs, ensuring that local structural nuances contribute effectively to the overall protein embedding.
- **Incorporating Contextual Similarity via Language Models**: By integrating embeddings from pre-trained protein language model ESM [5], we enrich the structural embeddings with contextual information derived from amino acid sequences, capturing functional and evolutionary relationships that may not be apparent from structure alone.
- **Contrastive Learning between Subgraphs**: We employ contrastive learning to refine local structural representations, encouraging the model to distinguish between similar and dissimilar substructures across different proteins.

2 Methodology

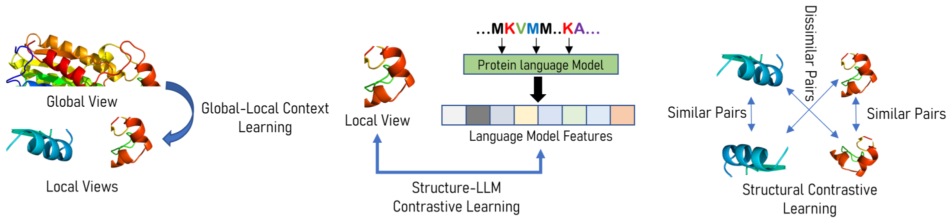


Figure 1: Overall training process of EmbedSimScore. Three different aspects of structural-contextual learning is employed: global-local context learning, structure-language model constrastive learning, and structural contrastive learning.

In this section, we discuss the key components of **EmbedSimScore** (Figure 1). First, we present the local and global structural alignment approach. Then, we focus on integrating contextual information through structure-language model joint contrastive learning. Finally, we explain how these components are combined into **EmbedSimScore** for comprehensive structure representation learning.

2.1 Structural Representation Alignment

To capture both local and global structural features, we generate multiple views of each protein graph $\mathcal{G}$, inspired by the advances of DINO [10] in computer vision. We generate a **Global View** $\mathcal{G}_{\text{global}}$, representing the larger protein substructure, and a **Local View** $\mathcal{G}_{\text{local}}$, focusing on local substructures. In this setting of knowledge distillation, the **teacher network** f_t operates on the global view, and the **student network** f_s operates on both global and local views. Both networks share the same architecture but have separate parameters. The student network is trained to align its representations with those of the teacher network by minimizing the alignment loss $\mathcal{L}_{\text{align}} = \sum_{v_g \in G_{global}} \sum_{\substack{v \in G_{global} \cup G_{local} \\ v_g \neq v}} -\mathcal{P}^t_{v_g} \log(\mathcal{P}^t_x)$, where $\mathcal{P}^t_g$ and $\mathcal{P}^s_g$ are the probability distribution of the pseudo-labels for graph g produced by f_t and f_s respectively.

2.2 Integrating Contextual Similarity via Language Models

To incorporate contextual information from protein sequences, we utilize a pre-trained language model $\mathcal{M}$ that generates sequence embeddings $\mathbf{h}_S$ from the amino acid sequence S corresponding to the protein graph $\mathcal{G}$. The structural embeddings $\mathbf{h}_\mathcal{G}$ of the student network are aligned with the sequence embeddings using a contrastive loss inspired by CLIP [11]. The loss is defined as $\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp\left(\text{sim}\left(\mathbf{h}_{\mathcal{G}}^{(i)},\mathbf{h}_{S}^{(i)}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(\mathbf{h}_{\mathcal{G}}^{(i)},\mathbf{h}_{S}^{(j)}\right)/\tau\right)}+\log\frac{\exp\left(\text{sim}\left(\mathbf{h}_{S}^{(i)},\mathbf{h}_{\mathcal{G}}^{(i)}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(\mathbf{h}_{S}^{(i)},\mathbf{h}_{\mathcal{G}}^{(j)}\right)/\tau\right)}\right]$, where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity, and τ is a temperature parameter controlling the sharpness of the distribution.

2.3 Contrastive Learning Between Subgraphs

To improve local structural representation learning, we apply contrastive learning between the augmented views of local subgraphs. For each protein, we create two augmented versions (check Appendix A.3 for details on augmentation) of a local subgraph, resulting in embeddings $\mathbf{h}^{(i,1)}$ and $\mathbf{h}^{(i,2)}$. The contrastive loss is defined as $\mathcal{L}_{\text{contrast}} = -\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}\left(\mathbf{h}^{(i,1)},\mathbf{h}^{(i,2)}\right)/\tau\right)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}\exp\left(\text{sim}\left(\mathbf{h}^{(i,1)},\mathbf{h}^{(k)}\right)/\tau\right)}$.

2.4 Overall Objective

The total loss function combines the above components: $\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{context}}\mathcal{L}_{\text{context}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}}$, where λ_{context}, and $\lambda_{\text{contrast}}$ are hyperparameters (check Appendix A.4 for details on hyperparameter selection) that balance the contributions of each loss term. With this combined approach, we train a graph neural network (GNN) backbone with five layers of GVP [12] on 48k proteins curated from protein data bank(PDB) [13] and their respective language model features generated using ESM-2[5] 650M model for 300 epochs. This backbone can embed any protein sequence, enabling the computation of structural and contextual relationships between proteins by comparing their **EmbedSimScore** embeddings.

3 Results

In this section, we discuss our findings on proteins' structural and contextual similarity calculation on a representative set with 50 protein molecules for whom we extract their 3d structure from the corresponding PDB file. This selection of protein structures is diverse across different protein families.

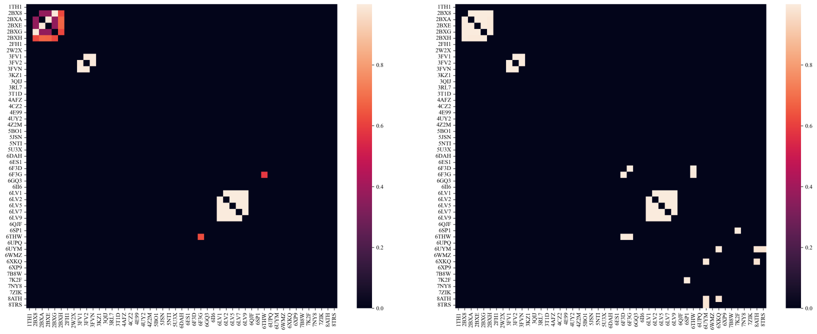


Figure 2: The heatmap in figure (a) represents similarity as shown by TM-score, and the one in figure (b) represents similarity scores as predicted by EmbedSimScore

In Figure 2 (a), we plot the TM-score between each pair of protein molecules in our representative set. A score less than 0.2 has been discarded as means that the similarities are rather random [14]. In Figure 2 (b), we have plotted the embedding similarity between each pair of protein embedding generated by EmbedSimScore. From the figure, we see that 24 protein pairs that are identified as similar by TM-score metrics are all identified by EmbedSimScore embedding as similar. However, EmbedSimScore captures another seven proteins that are structurally and contextually similar as calculated from their embedding similarity (similarity score > 0.98).

In Table 3, we show the 3D structural configurations of each pair of proteins that are identified as similar by EmbedSimScore, but not captured with TM-score. We see that, most of them have local structural similarity (similar folding structure in subgraphs), even though the overall global conformations may vary. Note that, we did not compare the structural configuration of 6F3D-6F3G pair, as they are the different conformation of the same protein molecule.

Score Type	6F3D - 6F3G	6UYM - 8TRS	6UYM - 6XKQ	6UYM - 8ATH	6XKQ - 8ATH	6F3D - 6THW	6SP1 - 7K2F
TM Score	0.011	0.024	0.003	0.010	0.003	0.005	0.059
Normalized Embedding Similarity	0.997	0.982	0.991	0.981	0.980	0.998	0.998

Table 1: Comparison of TM and Embedding Similarity Scores for 3D Structure of PDB Pairs

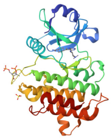
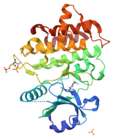
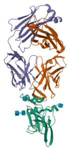
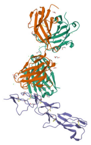
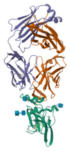
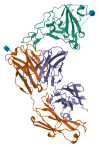
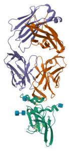
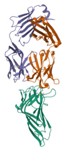
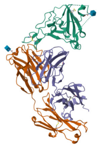
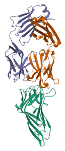
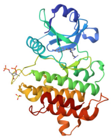
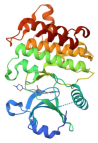
PDB ID Pair	3D Configuration (Protein 1)	3D Configuration (Protein 2)
6SP1 - 7K2F		
6UYM - 8TRS		
6UYM - 6XKQ		
6UYM - 8ATH		
6XKQ - 8ATH		
6F3D - 6THW		

Table 2: PDB pairs and their respective conformations

4 Conclusion

We presented **EmbedSimScore**, a self-supervised approach that enhances protein similarity computation by considering both local substructures and global protein features. Using structural alignment, contrastive learning, and language model integration, **EmbedSimScore** generates embeddings that capture relationships between proteins, often overlooked by traditional metrics like TM-score. **EmbedSimScore** sets a promising direction for advancing protein similarity analysis. Notably, aligning two proteins through their local substructure opens up new possibilities for various tasks, including protein engineering with specific functions (e.g., designing enzymes that catalyze novel reactions), drug discovery (e.g., identifying proteins with similar binding pockets for new therapeutic targets), and more, offering a powerful tool for targeted applications in these fields.

References

[1] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[2] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[5] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

[6] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

[7] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

[8] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[9] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael M Bronstein, and Bruno E Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

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

[14] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[15] Janani Durairaj, Andrew M Waterhouse, Toomas Mets, Tetiana Brodiazhenko, Minhal Abdullah, Gabriel Studer, Gerardo Tauriello, Mehmet Akdel, Antonina Andreeva, Alex Bateman, et al. Uncovering new families and folds in the natural protein universe. *Nature*, 622(7983):646–653, 2023.

[16] Mesih Kilinc, Kejue Jia, and Robert L Jernigan. Improved global protein homolog detection with major gains in function identification. *Proceedings of the National Academy of Sciences*, 120(9):e2211823120, 2023.

[17] Lorenzo Pantolini, Gabriel Studer, Joana Pereira, Janani Durairaj, Gerardo Tauriello, and Torsten Schwede. Embedding-based alignment: combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone. *Bioinformatics*, 40(1):btad786, 2024.

[18] Irina Ponamareva, Antonina Andreeva, Maxwell L Bileschi, Lucy Colwell, and Alex Bateman. Investigation of protein family relationships with deep learning. *Bioinformatics Advances*, 4(1):vbae132, 2024.

[19] Janani Durairaj, Andrew M Waterhouse, Toomas Mets, Tetiana Brodiazhenko, Minhal Abdullah, Gabriel Studer, Mehmet Akdel, Antonina Andreeva, Alex Bateman, Tanel Tenson, et al. What is hidden in the darkness? deep-learning assisted large-scale protein family curation uncovers novel protein families and folds. *bioRxiv*, pages 2023–03, 2023.

[20] Mindaugas Margelevičius. Gtalign: Spatial index-driven protein structure alignment, superposition, and search. *Nature Communications*, 15(1):7305, 2024.

[21] Ron S Boger, Seyone Chithrananda, Anastasios N Angelopoulos, Peter H Yoon, Michael I Jordan, and Jennifer A Doudna. Functional protein mining with conformal guarantees. *bioRxiv*, pages 2024–06, 2024.

[22] Eli J Draizen, Stella Veretnik, Cameron Mura, and Philip E Bourne. Deep generative models of protein structure uncover distant relationships across a continuous fold space. *Nature Communications*, 15(1):8094, 2024.

[23] R Prabakaran and Y Bromberg. Functional profiling of the sequence stockpile: a review and assessment of in silico prediction tools. *bioRxiv*, pages 2023–07, 2023.

[24] Qiqige Wuyun, Yihan Chen, Yifeng Shen, Yang Cao, Gang Hu, Wei Cui, Jianzhao Gao, and Wei Zheng. Recent progress of protein tertiary structure prediction. *Molecules*, 29(4):832, 2024.

A Appendix

A.1 Related Works

Structural and contextual similarity searching in macromolecules such as proteins is essential in bioinformatics and computational biology, particularly in uncovering new protein families and structural folds. Various techniques have been developed to navigate the natural protein universe, with recent advancements leveraging genomic context information, deep learning, and homology searches to predict protein functions and detect structural similarities. Here, we outline recent advancements and provide some insights into protein similarity matching.

One innovative approach involves the integration of genomic context information, which leverages remote homology searches on genomic neighbors to guide function prediction. Deep learning-based tools [15] are used for structure-guided function prediction, making it possible to reveal new families and folds in protein data. Sequence similarity searches are widely used to detect homologs, with traditional scoring methods like pairwise sequence alignment serving as a foundation [16].

A newer technique, embedding-based alignment (EBA), has been introduced to capture structural similarities in protein sequences. This method generates embedding-based alignments and has proven effective in identifying similarities in structure even when high sequence similarity is lacking [17]. Furthermore, deep learning models [18] have shown promising results in evaluating protein family relationships, employing benchmarks based on structure similarity, such as TM-align scores, to assess these relationships.

A combination of sequence, structure, and genomic context similarities is being used to reveal hidden insights in large-scale protein data. Researchers can identify less obvious structural features in protein sequences by employing deep learning-based function prediction methods [19]. Additionally, spatial index-driven protein structure alignment tools like GTalign [20] have emerged as robust methods for evaluating structural similarity using TM scores, a measure independent of protein length ratios.

Homology search methods also play a pivotal role in generating scores indicative of similarity between query proteins and known proteins. Such scores, including TM scores, provide insights into the structural similarity of protein structures, helping further research in protein structure prediction [21]. Deep generative models have additionally been developed to uncover distant structural similarities, emphasizing the significance of granularity in viewing protein structural similarities [22].

The use of metrics like the TM score to assess protein structure similarity remains a focal point in this field. Research continues to evolve as data types such as structure similarity, homology, and sequence length are integrated to improve predictive models [23]. Techniques like Position-specific Scoring Matrix (PSSM) are commonly employed in protein tertiary structure prediction, capturing amino acid preferences at each sequence position [24].

A.2 TM-score for Protein Similarity Matching

The Template Modeling (TM) score is a widely used metric for evaluating structural similarity between protein structures by aligning their C_α (alpha carbon) atoms. Unlike RMSD (Root Mean Square Deviation), the TM-score is independent of protein size and emphasizes overall topology rather than local structural variations. The TM-score ranges between 0 and 1, where higher scores indicate greater structural similarity.

The TM-score between two protein structures X and Y, each with L aligned residues, is defined as:

$$\text{TM-score} = \max\left(\frac{1}{L_{\text{target}}}\sum_{i=1}^{L}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right)$$

where,

- d_i is the distance between the i-th aligned C_α atoms of structures X and Y,
- L_{target} is the length of the target protein (often the shorter of the two structures),
- $d_0(L_{\text{target}})$ is a scaling factor given by:

$$d_0(L_{\text{target}}) = 1.24\sqrt[3]{L_{\text{target}} - 15} - 1.8.$$

A.2.1 Understanding TM-score values

The TM-score has an established interpretive range:

- **TM-score > 0.5:** High structural similarity. Structures are likely to share the same fold.
- **0.2 $<$ TM-score ≤ 0.5:** Moderate similarity. Structures may have some common motifs but differ in global topology.
- **TM-score ≤ 0.2:** Low similarity, indicating likely unrelated structures with different folds.

A.3 Augmentation Techniques for Local Subgraph Embeddings

To enhance local structural representation learning, we create augmented views of each protein's local and global subgraphs. Let us consider a subgraph(G) of a given protein graph $\mathcal{G}$. These augmentations introduce subtle variations that encourage the model to focus on invariant structural features. The following augmentation techniques are used:

- **Random Rotation:** Apply a random rotation $\mathbf{R} \in \mathrm{SO}(3)$ (where $\mathrm{SO}(3)$ denotes the 3D rotation group) uniformly to all nodes in G. The coordinates of each node v in the rotated graph G' become $\mathbf{x}'_v = \mathbf{R} \cdot \mathbf{x}_v$. This transformation ensures that the learned representation is invariant to the orientation of G in 3D space.
- **Random Gaussian Noise Addition:** Add random Gaussian noise $\mathbf{n}_v \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with small variance σ^2 to the coordinates of each node v. The perturbed position $\mathbf{x}'_v$ for each node v is given by $\mathbf{x}'_v = \mathbf{x}_v + \mathbf{n}_v$. This augmentation models minor structural fluctuations in the protein subgraph and enhances robustness to spatial variability.
- **Random Flipping:** Independently flip each coordinate axis with probability p by applying a transformation matrix $\mathbf{F} \in \{\mathbf{I}, \mathrm{diag}(-1, 1, 1), \mathrm{diag}(1, -1, 1), \mathrm{diag}(1, 1, -1)\}$ to all nodes in G. After flipping, each node v has new coordinates $\mathbf{x}'_v = \mathbf{F} \cdot \mathbf{x}_v$. This transformation introduces symmetry by reflecting G along different axes, promoting invariance to mirroring operations.

These transformations ensure that the learned representations capture essential local structural properties while maintaining invariance to minor spatial and topological variations.

A.4 Choosing Values for Coefficients of Loss Terms

In our overall objective function $\mathcal{L}$ we had three components $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{context}}$, and $\mathcal{L}_{\text{contrast}}$. In the loss function, the latter two terms are multiplied by constant factors λ_{context}, and $\lambda_{\text{contrast}}$ respectively. The purpose of these two constant terms was to control the contribution of those two loss components to the overall learning process. We wanted the contrastive components of the loss function to improve the learning without completely taking over as the major loss components from the alignment portion. In our experiments we varied the values of λ_{context}, and $\lambda_{\text{contrast}}$ between 0.1 and 0.5 at a 0.1 interval. We have found that 0.1 works best for both the loss components for best representation learning.