Enhancing Protein Mutation Effect Prediction through a Retrieval-Augmented Framework

Ruihan Guo^{*1}**, Rui Wang**^{*1}**, Ruidong Wu**^{*1}**, Zhizhou Ren**^{1}**, Jiahan Li**^{1}
Shitong Luo^{1}**, Zuofan Wu**^{1}**, Qiang Liu**^{1,2}**, Jian Peng**^{1}**, Jianzhu Ma**^{1,3}
^{1}Helixon Research, ^{2}The University of Texas at Austin University
^{3}Institute for AI Industry Research, Tsinghua
guoruihan.sansi@gmail.com
majianzhu@tsinghua.edu.cn

Abstract

Predicting the effects of protein mutations is crucial for analyzing protein functions and understanding genetic diseases. However, existing models struggle to effectively extract mutation-related local structure motifs from protein databases, which hinders their predictive accuracy and robustness. To tackle this problem, we design a novel retrieval-augmented framework for incorporating similar structure information in known protein structures. We create a vector database consisting of local structure motif embeddings from a pre-trained protein structure encoder, which allows for efficient retrieval of similar local structure motifs during mutation effect prediction. Our findings demonstrate that leveraging this method results in the SOTA performance across multiple protein mutation prediction datasets, and offers a scalable solution for studying mutation effects.

1 Introduction

Protein fitness plays significant roles in diverse applications in pharmaceutical industry [Amara, 2013], drug design [De Carvalho, 2011], biofuel production [Huang et al., 2020], and environmental bioremediation [Lu et al., 2022]. Deciphering mutation effects on protein fitness is crucial for understanding their functional dynamics and yet remains a central challenge in molecular biology. Most recent breakthroughs on computational predictions of mutation effects are driven by coevolutionary information [Riesselman et al., 2018, Luo et al., 2021, Notin et al., 2022]. The mutations on contacting residue pairs would become correlated under the evolutionary pressure to maintain protein stability and optimize functional efficiency within cellular environments. Consequently, conserved patterns within protein sequences and structures typically signify their stability and functionality. The predominant methodology to exploit such coevolutionary patterns is to perform multiple sequence alignments (MSA) [Thompson et al., 1994, 1997] and fit either statistical [Seemayer et al., 2014] or machine learning models [Rao et al., 2021]. In addition to sequence-level alignments, performing domain-level structure clustering [Orengo et al., 1997, Dong et al., 2018] is also a promising approach to extract evolutionary information from a protein family.

In this paper, we explore an alternative perspective to retrieve information for mutation effect prediction. In contrast to the global protein representation considered by MSA and domain-level structure alignments, we extract coevolutionary information in the scope of local microenvironments. We focus on the alignment of local structure motifs, *i.e.*, a central amino-acid with a few contacting neighbors. Such a local representation of coevolutionary information is specialized to the scenarios of protein engineering, where a common practice involves introducing a few point mutations to enhance a desired function [Shroff et al., 2020]. It is widely observed that the effects of point mutations are mainly given by the alteration of local biochemical microenvironments [Kim et al., 2011, Lu et al.,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2022]. Given these empirical insights, we propose to retrieve local structure fragments with similar backbone positions as auxiliary information for mutation effect prediction, while withdrawing the constraint on global sequence/structure similarity. This microscopic retrieval mechanism enables us to extract atom-level information from the whole protein universe rather than restricting to molecule-level instances within a certain protein family.

To elucidate the intricate details of local coevolutionary patterns in proteins, we employed a structure-based embedding approach, ESM-IF [Hsu et al., 2022], to encode local structure motifs into latent embeddings, assuming the metric space of such embeddings measures the similarity of motif structures. We preprocess the entire Protein Data Bank (PDB) [Berman et al., 2003] and build a database, we call Structure Motif Embedding Database (SMEDB), to support fast information retrieval by GPU-accelerated k-nearest neighbors (kNN) search. This retrieval procedure, we call Multiple Structure Motif Alignment (MSMA), is designed to extract coevolutionary information from protein fragments with similar local structure. We rigorously evaluate these extracted embeddings, focusing on their structural similarity and predictive accuracy for mutation effects. Remarkably, the embeddings derived from local coevolutionary patterns demonstrate superior performance compared to those obtained from traditional Multiple Sequence Alignments (MSA). Furthermore, the distribution of these embeddings was found to be complementary to those derived from MSA profiles, indicating that they capture distinct aspects of coevolutionary information, thereby enhancing our understanding of protein structure and function dynamics.

In addition, we introduce a novel model architecture, called Multi-Structure Motif Invariant Point Attention (MSM-IPA), to aggregate the retrieved coevolutionary motif information for predicting the structural fitness of proteins. This model, which effectively incorporates mutational information, has an outstanding capability to generalize across various mutations. It is trained to predict the change in binding free energy ($\Delta\Delta$G) on protein surfaces, an essential element for assessing protein-protein interactions. We extensively evaluate our model on a suite of widely-used protein stability and binding affinity benchmarks, including S669 [Pancotti et al., 2022], cDNA [Tsuboyama et al., 2023], and SKEMPI [Jankauskaitė et al., 2019] and demonstrate substantial improvement over baseline methods.

Our contributions are as follows:

- We develop Structure Motif Embedding Database (SMEDB), a comprehensive local structure alignment database encompassing all structures from the Protein Data Bank (PDB), which is organized to accelerate GPU-based kNN search.
- We propose Multiple Structure Motif Alignment (MSMA) to retrieve local coevolutionary motifs based on embeddings of protein structure encoders, which is shown to be complementary to classical sequence-level retrievals.
- Our model, Multi-Structure Motif Invariant Point Attention (MSM-IPA), pretrained on our novel database and retrieval mechanism, demonstrates superior performance in predicting $\Delta\Delta$G, surpassing other models on benchmark datasets, i.e., S669 and SKEMPI.

2 Preliminaries

Definitions A protein consists of multiple residues, possibly from different chains. For each residue i, we represent the residue as its residue type $a_i \in \{1, \ldots, 20\}$ and its positions of backbone heavy atoms with $\boldsymbol{p}_{i,C}, \boldsymbol{p}_{i,C\alpha}, \boldsymbol{p}_{i,N} \in \mathbb{R}^3$. A structure motif $\mathcal{M}$ is a fragment of the whole protein structure with N residues. For the raw structure motif $\mathcal{M}_{\text{raw}}$, we choose $N_{\text{raw}} = 256$ to match the pretraining settings and keep enough information about raw structure. For the retrieved structure motif (denoted as $\{\mathcal{M}_i\}_{i=1}^{L}$, where L is the size of the retrieved motif set), we choose $N_{\text{retr}} = 16$ to ensure sufficient interactions from similar environments are captured.

Our approach uses ESM-IF as a pretrain model and CUHNSW for vector database implementation. Here we make a brief introduction to these two methods.

ESM-IF ESM-IF [Hsu et al., 2022] is a model designed for protein sequence prediction using backbone structures. This approach treats inverse folding as a sequence-to-sequence problem with an autoregressive encoder-decoder architecture, allowing the model to recover native sequences from backbone atom coordinates. By incorporating a large number of sequences with predicted structures as additional training data, ESM-IF effectively learns even when experimental structures are unavailable. The augmented data and backbone-only input make ESM-IF a suitable model to encode the nearby backbone structure of a residue.

CUHNSW CUHNSW [Malkov and Yashunin, 2020] is a CUDA implementation of Hierarchical NSW (HNSW), a novel approach of vector database that can do approximate K-nearest neighbor (K-NN) search utilizing navigable small world graphs with a controllable hierarchy. Unlike traditional methods, CUHNSW is fully graph-based and eliminates the need for additional search structures. It incrementally builds a multi-layer structure of proximity graphs, with elements randomly assigned to layers using an exponentially decaying probability distribution. This method enhances performance by starting the search from upper layers and leveraging scale separation, resulting in logarithmic complexity scaling. Performance evaluations show that CUHNSW outperforms previous state-of-the-art vector-only approaches and its similarity to the skip list structure allows for straightforward distributed implementation.

3 Methodology

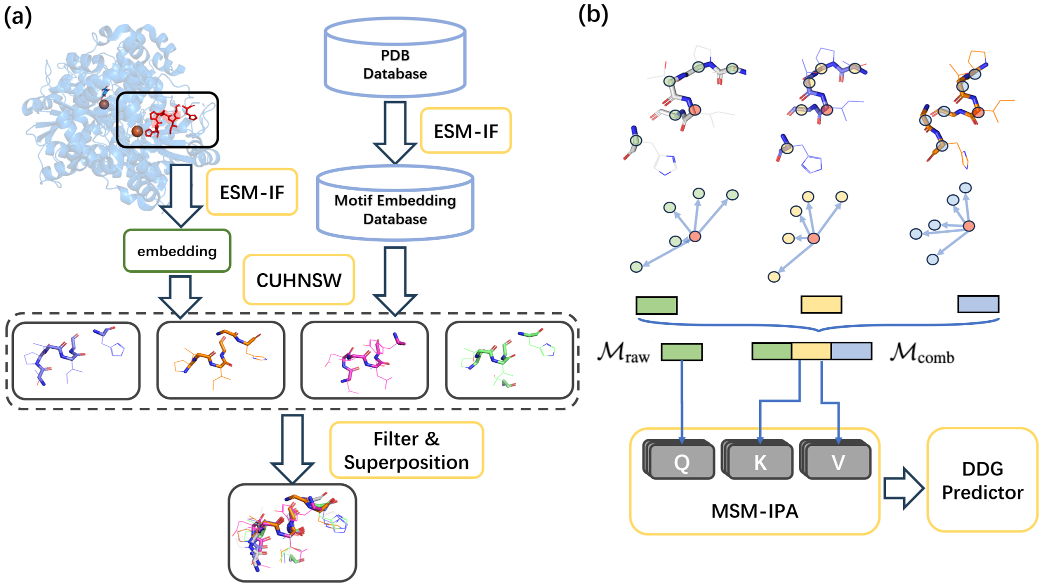

Figure 1: Overview of the retrieval augmented framework. (a) Multiple Structure Motif Alignment process. (b) A diagram illustrating MSM-Mut, a model that can predict mutation effect with information get from multiple structure motif alignment.

3.1 Multiple Structure Motif Alignment

We use per-residue embeddings extracted by ESM-IF model [Hsu et al., 2022] for each of the protein chains in PDB database and perform local structure search by HNSW. To obtain embeddings for ESM-IF, we utilize the encoder module named GVPTransformerEncoder in ESM-IF to generate a 512-dimensional embedding for each residue. As we exclusively use the encoder module of ESM-IF, the computational cost remains manageable, requiring approximately 3 days on 32 A100 GPUs.

The PDB dataset comprises more than 130 million residues, making it impractical to load the entire dataset into memory for querying. Traditional databases lack the capability to leverage GPU acceleration for efficiently querying the top k nearest neighbors. To address this, we integrate CUHNSW as a module to obtain the k-nearest neighbors in the ESM-IF embedding space. The structure motifs are retrieved for their similar local interactions between residues that have similar backbone positions. The interaction between residues occurs only when they are in close proximity. Therefore, when processing the retrieved structure motifs, a large motif size is not necessary. For simplicity, in this paper, we always use $N_{\text{retr}} = 16$. With a highly efficient CUDA implementation [Yoon, 2021] the time consumption querying the top 10^5 neighbors of a result is about 8 seconds in 8 A100 GPUs and only 0.5 seconds for the top 10^3 neighbors.

The advantage of using an inverse folding model as a structure encoder, compared to traditional methods that rely on retrieving from a database based on continuous structural fragments, lies in the model's inherent ability to generate embeddings tailored for predicting the surrounding environment based on a given backbone structure. The inverse folding model encodes positions within the spatial context that significantly influence the central amino acid type, rather than merely contiguous positions along the sequence.

Since we use embeddings from a structure encoder (ESM-IF in this case) instead of directly search on structures, we would like to validate the performance of the embedding retrieval method by checking the matched motif size N_{matched} of the search result comparing with that of structure search. We first define the *matched* motif size as

$$N_{\text{matched}} = \sum_{i \in N_{\text{raw}}} \mathbb{1}\left[\min_{j \in N_{\text{retr}}} (\text{dist}(R_{\text{raw},i}, R_{\text{motif},j})) < 2.0\text{Å}\right] \tag{1}$$

The left panel of Figure 2 illustrates the distribution of alignment sizes for the retrieved structures, demonstrating that numerous similar spatial structures can be found centered on each position. The other panel of Figure 2 shows the relationship between local similarity (number of matching amino acids) and overall similarity (TM-score) with different motif size. The figure show that the similar motif may have low TM-score, indicating that our search approach can extract analogous local structures motif from structurally unrelated proteins to aid in prediction.

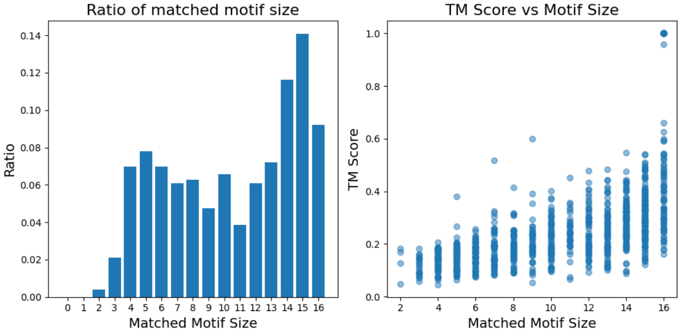

Figure 2: CUHNSW combined with embeddings from ESM-IF successfully retrieved motifs with large *matched* regions, and possibly motifs distant in sequence identity, where matched residues are defined by the indicator function in Eq. (1). Moreover, lower TM-Score is obtained for motifs with high number of locally matched residues, indicating that our search approach can extract analogous local structures motif from structurally unrelated proteins.

3.2 Multi Structure Motif Modelling

In this section, we introduce how we utilize the retrieved structure motifs to help predict the effects of mutations. We begin by discussing the filtering of retrieved structure motifs, followed by the superposition of retrieved data, which enables the model to learn information from diverse structures. Finally, we introduce our module, MSM-IPA, which extracts similarity information from the PDB database.

Retrieved Structure Motif Filter After extraction, we obtain a large volume of neighbor data (on the order of 10^3). For efficiency reasons, we only reserve $L_{\text{filter}} = 16$ data points with most valuable information. This is done in the following manner: Firstly, retrieved structure motifs with central amino acid different from query are discarded. After that, We rank them with a scoring function and retain the top results. The scoring formula can be expressed as follows:

$$\sum_{i \in N_{\text{raw}}} \mathbb{1}\left[\min_{j \in N_{\text{retr}}} (\text{dist}(R_{\text{raw},i}, R_{\text{motif},j})) < 2.0\text{Å}\right] \cdot \exp(-\|p_{i,C\alpha} - p_{0,C\alpha}\|_2) \tag{2}$$

where $\mathbb{1}[\cdot]$ is the indicator function and the distance function $\text{dist}(R_1, R_2) = \|p_{1,C\alpha} - p_{2,C\alpha}\|_2$. Intuitively, we use the weighted term $\exp(-\|p_{i,C\alpha} - p_{0,C\alpha}\|_2)$ to rewrite Eq. (1) to retain as many contacts around the central amino acid as possible.

When dealing with multiple mutations, which involves several central amino acids, we ensure a balanced selection of retrieved structure motif information by using each central amino acid as a query. This approach guarantees an even distribution of the retrieved information across all central amino acids.

In the context of mutation analysis, we conduct separate selection processes for both the pre-mutation (wild type) and post-mutation states. The key distinction between these two processes is the amino acid type in the initial step—one being the wild type and the other the mutated form.

Superposition The choice of superimposition method depends on the nature of the extracted information. We compare two superimposition approaches: alignment based on the central frame and alignment based on the overall structure. Experimental results indicate that the central frame-based alignment method provides better alignment quality for the retrieved structures. Consequently, we select the central frame-based alignment method for subsequent analyses.

For each residue, we can construct the frame from its backbone atom positions $\boldsymbol{p}_C, \boldsymbol{p}_{C_\alpha}, \boldsymbol{p}_N$.

$$\begin{aligned}
\boldsymbol{v}_{N,C_\alpha} &= (\boldsymbol{p}_N - \boldsymbol{p}_{C_\alpha})/\|\boldsymbol{p}_N - \boldsymbol{p}_{C_\alpha}\| \\
\boldsymbol{v}_{C,C_\alpha} &= (\boldsymbol{p}_C - \boldsymbol{p}_{C_\alpha})/\|\boldsymbol{p}_C - \boldsymbol{p}_{C_\alpha}\| \\
\boldsymbol{R} &= [\boldsymbol{v}_{N,C_\alpha}, \boldsymbol{v}_{C,C_\alpha}, \boldsymbol{v}_{N,C_\alpha} \times \boldsymbol{v}_{C,C_\alpha}] \\
\boldsymbol{t} &= \boldsymbol{p}_{C_\alpha},
\end{aligned}$$

where $\times$ is the cross product between two vectors.

Hence with the given raw structure motif $\mathcal{M}_{\text{raw}}$ and any other structure motif $\mathcal{M}_{\text{retr}}$, we firstly take out the central residue $\mathcal{R}^C_{\text{raw}}$ and $\mathcal{R}^C_{\text{retr}}$. Then we can calculate the frame of both the residues, called $(\boldsymbol{R}_C, \boldsymbol{t}_C)$ and $(\boldsymbol{R}_{\text{retr}}, \boldsymbol{t}_{\text{retr}})$ respectively. Then we can define the alignment function as:

$$\mathcal{F}_{\text{align-atom}}(\boldsymbol{p}) = \boldsymbol{R}_C \boldsymbol{R}^T_{\text{retr}}(\boldsymbol{p} - \boldsymbol{t}_{\text{retr}}) + \boldsymbol{t}_C.$$

The alignment procedure involves applying the alignment function to each atom of $\mathcal{M}_{\text{retr}}$ to obtain the aligned motif $\mathcal{M}^{\text{aligned}}_{\text{retr}}$. For simplicity, in the following sections, we will omit the term "aligned" and assume that all motifs are aligned with the raw motif by default.

MSM-IPA To retrieve information from structure motifs, we propose Multi-Structure Motif Invariant Point Attention (MSM-IPA), which is inspired by and similar to the IPA module in AlphaFold2 [Jumper et al., 2021]. The MSM-IPA module takes the raw structure motif $\mathcal{M}_{\text{raw}}$ and the processed retrieved structure motifs $\mathcal{M}_1, \ldots, \mathcal{M}_L$ as input. For simplicity, we merge the raw motif with the retrieved motifs into a single motif, denoted as $\mathcal{M}_{\text{comb}} = \{r \in \mathcal{M}_i, \forall i \in \{1, \ldots, L\}\} \cup \{r \in \mathcal{M}_{\text{raw}}\}$. $\mathcal{M}_{\text{comb}}$ is then used as *key* motifs in the cross attention mechanism. The size of this combined motif is represented as N_{comb}.

To extract information from the motif, we define two encoders, $\text{Enc}_s(\mathcal{M})$ for single node-wise representations and $\text{Enc}_z(\mathcal{M}_1, \mathcal{M}_2)$ for pair edge-wise representations. The single encoder encodes the residue types and positions of atoms into a single representation, $\boldsymbol{s}$. The pair encoder encodes information such as relative positions in the sequence, spatial relative positions, and amino acid type pairs into a pair representation, $\boldsymbol{z}$.

In MSM-IPA, we extract scalar and vector representations from $\boldsymbol{s}$. The attention weights are updated and aggregated from three sources of information: single, pair, and predicted points. This approach ensures that the model efficiently utilizes the information while maintaining invariance to overall rotation and translation. The updated single representation of the raw structure motif is obtained by concatenating the weighted sums of each data representation. The detailed implementation of this algorithm is shown in Algorithm 1.

MSM-Mut The MSM-Mut model mainly consists of two parts: a series of MSM-IPA to fusion the information and a mutation effect predictor consisted of a series of MLP. For both the wild-type and mutated structures, we retrieve structure motifs independently, merge the information using MSM-IPA, and model it with IPA to obtain the single representations for both structures. These representations are then fed into the mutation effect predictor to get the final impact on the structure.

3.3 Model Training Details

Our model training comprises two phases. In the pretraining phase, we utilize a meticulously curated dataset from the Protein Data Bank-REDO (PDB-REDO) [Joosten et al., 2014] for our pretraining data. The dataset is split into training, validation, and test sets in a ratio of 95%:0.5%:4.5%. The pretraining process involves an initial 200,000 steps without the inclusion of retrieved structure motifs, followed by an additional 30,000 steps incorporating retrieved structure motifs. During training, a random amino acid was selected, and its 256 nearest amino acids are extracted with their amino acid types and backbone atom positions. The type of the central amino acid was masked since the model was tasked with predicting it.

In the finetuning phase, to ensure comparability of our model, we follow the settings of StableOracle and RDE. For the stability mutation effect prediction task, we use the CDNA120K [Tsuboyama et al., 2023] subset of the training set, which was deduplicated against s669 in StableOracle. For the PPI surface mutation effect prediction task, we perform 3-fold cross-validation on the SKEMPI dataset, partitioned by PDBID.

4 Results

In this section, we show that our model outperforms other models on a series of tasks with the retrieved MSM information. To begin with, in section 4.1.1, we demonstrate that the retrieved information provides an advantage in predicting mutation effects on protein-protein interfaces. Subsequently, in section 4.1.2, we present a case study on antibody engineering for SARS-CoV-2, illustrating our model's strong performance on real-world targets. In addition, in section 4.2.1, we conducted an experiment to show that our model outperforms others on the protein stability change dataset. To further validate our model's capability in stability prediction, in section 4.2.2, we tested it on a novel enzyme dTM dataset and achieved excellent results.

Baseline models The baselines referenced in this study encompass two main categories: unsupervised models, and semi-supervised/supervised models. The unsupervised models comprise PSSM (position-specific scoring matrix), ESM-1v [Meier et al., 2021], MSA Transformer [Rao et al., 2021], Tranception [Notin et al., 2022], ESM-IF [Hsu et al., 2022], ProteinMPNN [Dauparas et al., 2022], Rosetta [Park et al., 2016, Alford et al., 2017], and FoldX [Delgado et al., 2019]. The semi-supervised category includes MIF-Net [Yang et al., 2023], RDE-Net [Luo et al., 2023], and DiffAffinity [Liu et al., 2024], while supervised models encompass DDGPred [Shan et al., 2022] and the End-to-End method.

Evaluation metrics In this section, we employ six metrics to evaluate our model: Pearson and Spearman correlations to quantify overall trends, per-structure metrics to assess the model's ability to identify beneficial and detrimental mutations within each structure, and RMSE and MAE to measure numerical differences in predicted energy values. Among these, the most critical metric is the per-structure correlation, as practical applications often involve ranking different mutations within the same structure.

4.1 Mutation Effects Prediction on PPI Surface

Following the settings in RDE [Luo et al., 2023], we partitioned the dataset into three folds based on PDB IDs. Two of these folds were further divided into training and validation sets in a 95:5 ratio based on PDB IDs, while the remaining fold was used as the test set. By training in this manner, we obtained three distinct sets of model parameters, each corresponding to the performance measured on the entire dataset used as the test set.

4.1.1 Experimental Results on SKEMPI2.0

This experiment primarily demonstrates two key points. Firstly, we establish a simple baseline that directly utilizes the top 100 retrieved structure motifs obtained from our search to construct a profile, *i.e.*, simple statistics of the amino-acid types according to the retrieval results. This baseline, called MSM-profile, achieves significantly better performance compared to MSA, and even surpass the majority of unsupervised language models. Note that MSM-profile is built upon the embeddings obtained from the ESM-IF model, and we observed that the per-structure Pearson and Spearman correlation coefficients extracted directly are comparable to those predicted by the ESM-IF model. This indicates that our search method effectively retains the information from ESM-IF, allowing us to explicitly extract local structure motifs corresponding to the ESM-IF distribution.

Secondly, the semi-supervised model trained with the enhanced information achieved state-of-the-art performance across almost all metrics. Among the metrics, the per-structure correlation is the most crucial, as it reflects the model's ranking capability for mutations on PPI surfaces. We found that the use of the information we extracted resulted in a significant improvement in the model's prediction correlation, underscoring the value of the retrieved information.

Table 1: Performance comparison to baseline methods on SKEMPI2.0 benchmark.

		Per-Structure		Overall			
Category	Method	Pearson	Spearman	Pearson	Spearman	RMSE	MAE
Energy function	Rosetta	0.3284	0.2988	0.3113	0.3468	1.6173	1.1311
	FoldX	0.3789	0.3693	0.3120	0.4071	1.9080	1.3089
Profile	PSSM	0.0826	0.0822	0.0159	0.0666	1.9978	1.3895
	MSM-profile	0.1905	0.1886	0.1653	0.2063	1.9423	1.3784
Unsupervised	ESM-1v	0.0073	-0.0118	0.1921	0.1572	1.9609	1.3683
	MSA Transf.	0.1031	0.0868	0.1173	0.1313	1.9835	1.3816
	Tranception	0.1348	0.1236	0.1141	0.1402	2.0382	1.3883
	ESM-IF	0.2241	0.2019	0.3194	0.2806	1.8860	1.2857
	ESM2	0.0100	0.0100	0.1700	0.1630	2.6580	2.0210
	EVE	0.1131	0.0898	0.1237	0.1088	2.2622	1.4178
Semi-sup./ Supervised	ESM2(Sup)	0.3330	0.3040	0.6030	0.5290	2.1500	1.6700
	DDGPred	0.3750	0.3407	0.6580	0.4687	1.4998	1.0821
	End-to-End	0.3873	0.3587	0.6373	0.4882	1.6198	1.1761
	MIF-Net.	0.3965	0.3509	0.6523	0.5134	1.5932	1.1469
	RDE-Net.	0.4448	0.4010	0.6447	0.5584	1.5799	1.1123
	DiffAffinity.	0.4220	0.3970	0.6690	0.5560	1.5350	1.0930
	MSM-Mut (w/o retrieval)	0.4325	0.4031	0.6233	0.4954	1.6076	1.2155
	MSM-Mut	**0.4736**	**0.4354**	**0.6814**	**0.5786**	**1.4703**	**1.0212**

4.1.2 SARS-COV-2 Antibody Optimization

In Shan et al. [2022], five single mutations are identified that enhance the neutralization effectiveness of antibodies against SARS-CoV-2. Within the three CDR regions in the heavy chain, spanning a total of 26 positions, there are 494 possible mutations. Our task is to select these five beneficial mutations from the pool. We compare our results with a subset of the baseline methods that performed well in Section 4.3. Experimental results indicate that our model excels in predicting the ranking of RH103M, while also maintaining high ranks for the other four mutations. This improvement is attributed to our model's ability to extract similar structure motifs from the database.

Table 2: Rankings of the five beneficial mutations on an anti-SARS-CoV-2 antibody.

Category	Method	TH31W	AH53F	NH57L	RH103M	LH104F
Energy function	Rosetta	**10.73%**	76.72%	93.93%	**11.34%**	27.94%
	FoldX	**13.56%**	**6.88%**	**5.67%**	**16.60%**	66.19%
DL-based	RDE-Net	**5.06%**	**12.15%**	35.47%	50.61%	**9.51%**
	DiffAffinity	**7.28%**	**3.64%**	**18.82%**	81.78%	**10.93%**
	MSM-Mut (w/o retrieval)	**9.51%**	**11.94%**	36.44%	50.61%	23.19%
	MSM-Mut	**6.48%**	**10.12%**	**16.19%**	**19.23%**	20.04%

Case Study: RH103M Given that this task focuses on antibody optimization, we incorporated a post-processing step in the retrieval process. We prioritize those retrieved data corresponding to an antibody or nanobody and if the relevant position was located in a loop region. At position H103, out of the top 1000 candidates, only 5 structure motifs remained after filtering, with one having the central residue type as methionine (M). Experimental results indicated that including this motif significantly influenced the outcome. Analysis revealed that this structure corresponds to residue 576 of chain T in the 5KOV structure. The aligned structures are depicted in Figure 3.

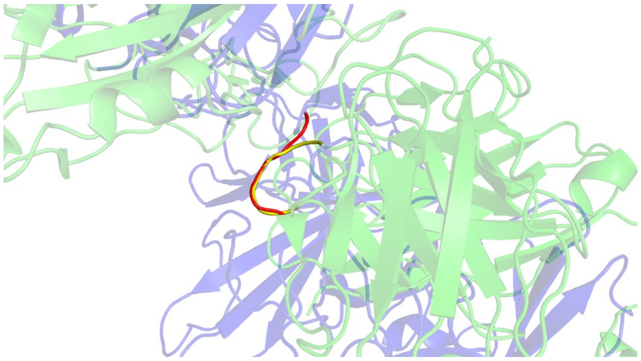

Figure 3: A diagram illustrating a set of highly similar local antibody structures obtained through Multiple Structure Motif Alignment. The figure compares the local structures of the T chain 576 from 5KOV and the H chain 103 from 7FAE.

This demonstrates that our data retrieval method can effectively assist in extracting relevant local structures for antibody design, potentially offering an interpretable approach to the antibody engineering process.

4.2 Stability Change Prediction

Training MSM-Mut(or other model name) for protein engineering To train the model's affinity prediction module, we fine-tuned the stability prediction component using the cDNA dataset. To fully utilize the mutation data and reduce bias, ensuring a more balanced representation of mutation types, we followed the settings described in Stability Oracle . Following Diaz et al. [2023], we use Thermodynamic Permutations (TP) to perform data augmentation to balance the mutation type distribution of dataset cDNA120K [Tsuboyama et al., 2023] and train on 2M thermodynamically valid ddG measurements. TP achieves this by exploiting properties of Gibbs free energy and dataset characteristics. Specifically, the original 19 mutations at each position are augmented to cover all 380 (19×20) pairwise mutations.

4.2.1 Experimental Results on S669

In this study, we utilized a newly curated dataset (S669) [Pancotti et al., 2022] derived from the latest version of ThermoMutDB. This dataset comprises 669 protein variants not found in commonly used training sets, offering a robust basis for evaluating prediction models. Although our model does not match Stability Oracle in predicting stability changes, the integration of retrieved structure motifs enabled our model to surpass the such methods, thereby establishing a new state-of-the-art in this task.

4.2.2 Thermostability Optimization on Novozymes Dataset

To demonstrate our model's robust generalization capability on new data, we tested it on a novel enzyme thermostability dataset provided by Novozymes [Pultz et al., 2022]. This dataset includes experimental measurements of melting temperature of point mutations on a novel enzyme sequence that has no high-similarity match in PDB database with sequence identity higher than 30%. An AlphaFold prediction of wild-type protein structure is released with the dataset as the reference to perform structure-based methods. We adopt this predicted structure to retrieve local motifs and feed to MSM-Mut. As the results shown in Table 3b, our method significantly outperforms both classical force fields and machine-learning-based baselines. It suggests that our method can be applied to novel proteins and predicted structures.

Table 3: Performance comparison to baseline methods on S669 and Novozymes datasets.

(a) Results on S669 dataset.

Method	Pearson	RMSE
ESM-1v	0.16	3.05
ESM-IF	0.27	2.43
FoldX	0.22	2.30
Rosetta	0.39	2.70
ProteinMPNN	0.26	3.32
Stability Oracle	0.52	1.43
MSM-Mut (w/o retrieval)	0.45	1.73
MSM-Mut	**0.54**	**1.51**

(b) Results on Novozymes dataset.

Method	Spearman
ESM-1v	0.174
ESM-IF	0.255
FoldX	0.415
Rosetta	0.438
ProteinMPNN	0.231
MSM-Mut (w/o retrieval)	0.323
MSM-Mut	**0.484**

5 Related Work

Mutation effect prediction Mutation effect prediction plays a pivotal role in in-silico protein engineering. There are three major categories of methods for mutation effect prediction–biophysical [Schymkowitz et al., 2005, Park et al., 2016, Alford et al., 2017, Steinbrecher et al., 2017], statistical [Geng et al., 2019, Zhang et al., 2020], and deep learning-based methods [Rao et al., 2021, Liu et al., 2021, Shan et al., 2022, Yang et al., 2023, Luo et al., 2023].

Different information is provided as inputs to different deep learning models. Models based on protein language models (PLMs) usually require only the primary sequence of both wild type and mutation [Meier et al., 2021, Notin et al., 2022]. MSAs also serve as inputs to mutation prediction models [Hopf et al., 2017, Riesselman et al., 2018, Rao et al., 2021, Luo et al., 2021, Frazer et al., 2021] based on the observation that co-evolutionary information correlates with the information needed. In addition to sequences, structures of protein also provide additional information. Along this line, a number of models are pretrained on protein structure data to encode the structure information [Hsu et al., 2022, Yang et al., 2023, Zhang et al., 2023]. However, MLStrA-IPA is, to our knowledge, the first model that employs multiple structures for single mutation effect prediction.

Retrieval-based deep learning Retrieval-based deep learning has made its splash in language modeling recently [Guu et al., 2020, Wu et al., 2022b] that achieves higher performance-per-parameter.

Retrieval-based predictions have been long employed in protein structure prediction where people searches for evolutionarily related sequences of the sequence of interest and gather them into one MSA input to a model, which has been an important building block for models including AlphaFold [Jumper et al., 2021] and RoseTTAFold [Baek et al., 2021], while replacing MSA modules with language model demonstrates a small but noticeable drop in performance [Wu et al., 2022a, Lin et al., 2023].

Retrieval mechanism other than MSA has also been investigated previously. Retrieved Sequence Augmentation [Ma et al., 2023] augments PLMs with a collection of related sequences, either in sequence or in structure, and claims substantial improvements over MSA Transformer [Rao et al., 2021] both in performance and in prediction speed. Wang et al. [2023] use retrieval techniques to improve the performance of controllable molecule generation.

6 Conclusion

In this study, we addressed the challenge of predicting the effects of protein mutations by introducing a novel retrieval-augmented framework that leverages local structure motifs. By creating the Structure Motif Embedding Database (SMEDB) and using Multiple Structure Motif Alignment (MSMA), we efficiently retrieved and utilized local coevolutionary information. Our Multi-Structure Motif Invariant Point Attention (MSM-IPA) model demonstrated superior performance on benchmark datasets, proving the value of focusing on local structural environments. Our model offers a scalable and robust method for studying protein mutations, with important implications for protein engineering and genetic disease research.

7 Acknowledgements

This work is supported by the National Key RD Program of China No.2021YFF1201600

References

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Jr. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017. doi: 10.1021/acs.jctc.7b00125. URL `https://doi.org/10.1021/acs.jctc.7b00125`. PMID: 28430426.

Amro Abd-Al-Fattah Amara. Pharmaceutical and industrial protein engineering: where we are? *Pakistan Journal of Pharmaceutical Sciences*, 26(1), 2013.

Minkyung Baek, Frank Dimaio, Ivan V. Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy M. DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina Aike van Dijk, Ana C. Ebrecht, Diederik Johannes Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a 3-track neural network. *Science (New York, N.Y.)*, 373:871 – 876, 2021.

Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature structural & molecular biology*, 10(12):980–980, 2003.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Carla CCR De Carvalho. Enzymatic and whole cell catalysis: finding new strategies for old processes. *Biotechnology advances*, 29(1):75–83, 2011.

Javier Delgado, Leandro G Radusky, Damiano Cianferoni, and Luis Serrano. Foldx 5.0: working with rna, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, 2019.

Daniel J Diaz, Chengyue Gong, Jeffrey Ouyang-Zhang, James M Loy, Jordan Wells, David Yang, Andrew D Ellington, Alex Dimakis, and Adam R Klivans. Stability oracle: a structure-based graph-transformer for identifying stabilizing mutations. *BioRxiv*, pages 2023–05, 2023.

Runze Dong, Zhenling Peng, Yang Zhang, and Jianyi Yang. mtm-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, 34(10):1719–1725, 2018.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, Nov 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04043-8. URL `https://doi.org/10.1038/s41586-021-04043-8`.

Cunliang Geng, Anna Vangone, Gert E. Folkers, Li C. Xue, and Alexandre M. J. J. Bonvin. isee: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins: Structure, Function, and Bioinformatics*, 87(2):110–119, 2019. doi: https://doi.org/10.1002/prot.25630. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25630`.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, Feb 2017. ISSN 1546-1696. doi: 10.1038/nbt.3769. URL https://doi.org/10.1038/nbt.3769.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Jie Huang, Peng Zhao, Xin Jin, Yiwen Wang, Haotian Yuan, and Xinyuan Zhu. Enzymatic biofuel cells based on protein engineering: Recent advances and future prospects. *Biomaterials science*, 8 (19):5230–5240, 2020.

Justina Jankauskaitė, Brian Jiménez-García, Justas 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.

Robbie P Joosten, Fei Long, Garib N Murshudov, and Anastassis Perrakis. The pdb_redo server for macromolecular structure model optimization. *IUCrJ*, 1(4):213–220, 2014.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Sung Bae Kim, Hideyuki Suzuki, Moritoshi Sato, and Hiroaki Tao. Superluminescent variants of marine luciferases for bioassays. *Analytical chemistry*, 83(22):8732–8740, 2011.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Shiwei Liu, Tian Zhu, Milong Ren, Chungong Yu, Dongbo Bu, and Haicang Zhang. Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36, 2024.

Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, and Jian Peng. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLOS Computational Biology*, 17(8):1–28, 08 2021. doi: 10.1371/journal.pcbi.1009284. URL https://doi.org/10.1371/journal.pcbi.1009284.

Hongyuan Lu, Daniel J Diaz, Natalie J Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J Acosta, Bradley R Alexander, Hannah O Cole, Yan Zhang, et al. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, 2022.

Shitong Luo, Yufeng Su, Zuofan Wu, Chenpeng Su, Jian Peng, and Jianzhu Ma. Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. *bioRxiv*, pages 2023–02, 2023.

Yunan Luo, Guangde Jiang, Tianhao Yu, Yang Liu, Lam Vo, Hantian Ding, Yufeng Su, Wesley Wei Qian, Huimin Zhao, and Jian Peng. Ecnet is an evolutionary context-integrated deep learning framework for protein engineering. *Nature Communications*, 12(1):5743, Sep 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25976-8. URL https://doi.org/10.1038/s41467-021-25976-8.

Chang Ma, Haiteng Zhao, Lin Zheng, Jiayi Xin, Qintong Li, Lijun Wu, Zhihong Deng, Yang Lu, Qi Liu, and Lingpeng Kong. Retrieved sequence augmentation for protein representation learning. *bioRxiv*, 2023. doi: 10.1101/2023.02.22.529597. URL https://www.biorxiv.org/content/early/2023/05/23/2023.02.22.529597.

Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42 (4):824–836, apr 2020. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2889473. URL https://doi.org/10.1109/TPAMI.2018.2889473.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N. Gomez, Debora Marks, and Yarin Gal. Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16990–17017. PMLR, June 2022. URL https://proceedings.mlr.press/v162/notin22a.html. ISSN: 2640-3498.

Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Corrado Pancotti, Silvia Benevenuta, Giovanni Birolo, Virginia Alberini, Valeria Repetto, Tiziana Sanavia, Emidio Capriotti, and Piero Fariselli. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Briefings in Bioinformatics*, 23(2):bbab555, 2022.

Hahnbeom Park, Philip Bradley, Per Jr. Greisen, Yuan Liu, Vikram Khipple Mulligan, David E. Kim, David Baker, and Frank DiMaio. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of Chemical Theory and Computation*, 12(12):6201–6212, 2016. doi: 10.1021/acs.jctc.6b00819. URL https://doi.org/10.1021/acs.jctc.6b00819. PMID: 27766851.

Dennis Pultz, Esben Friis, Jesper Salomon, Peter Fischer Hallin, Sarah Baagøe Jørgensen, Walter Reade, and Maggie Demkin. Novozymes enzyme stability prediction, 2022. URL https://kaggle.com/competitions/novozymes-enzyme-stability-prediction.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8844–8856. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, Oct 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0138-4. URL https://doi.org/10.1038/s41592-018-0138-4.

Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(suppl_2):W382–W388, 07 2005. ISSN 0305-1048. doi: 10.1093/nar/gki387. URL https://doi.org/10.1093/nar/gki387.

Stefan Seemayer, Markus Gruber, and Johannes Söding. Ccmpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, 30(21):3128–3130, 2014.

Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, Xuanling Shi, Qi Zhang, Bonnie Berger, Linqi Zhang, and Jian Peng. Deep learning guided optimization of human antibody against sars-cov-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences*, 119(11):e2122954119, 2022. doi: 10.1073/pnas.2122954119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2122954119.

Raghav Shroff, Austin W Cole, Daniel J Diaz, Barrett R Morrow, Isaac Donnell, Ankur Annapareddy, Jimmy Gollihar, Andrew D Ellington, and Ross Thyer. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS synthetic biology*, 9(11):2927–2935, 2020.

Thomas Steinbrecher, Robert Abel, Anthony Clark, and Richard Friesner. Free energy perturbation calculations of the thermodynamics of protein side-chain mutations. *Journal of Molecular Biology*, 429(7):923–929, 2017. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2017.03.002. URL `https://www.sciencedirect.com/science/article/pii/S0022283617301067`.

J Drew Thompson, Desmond G. Higgins, and Toby J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22 22:4673–80, 1994.

J Drew Thompson, Toby J. Gibson, Frédéric Plewniak, F Jeanmougin, and Desmond G. Higgins. The clustal_x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research*, 25 24:4876–82, 1997.

Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973): 434–444, 2023.

Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. Retrieval-based controllable molecule generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=vDFA1tpuLvk`.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022a. doi: 10.1101/2022.07.21.500999. URL `https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999`.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022b. URL `https://openreview.net/forum?id=TrjbxzRcnf-`.

Kevin K Yang, Hugh Yeh, and Niccolò Zanichelli. Masked inverse folding with sequence transfer for protein representation learning, 2023. URL `https://openreview.net/forum?id=2EO8eQ2vySB`.

Jisang Yoon. Efficient cuda implementation of hierarchical navigable small world (hnsw) graph algorithm for approximate nearest neighbor (ann), 2021. URL `https://github.com/js1010/cuhnsw`.

Ning Zhang, Yuting Chen, Haoyu Lu, Feiyang Zhao, Roberto Vera Alvarez, Alexander Goncearenco, Anna R Panchenko, and Minghui Li. MutaBind2: Predicting the impacts of single and multiple mutations on protein-protein interactions. *iScience*, 23(3):100939, March 2020.

Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=to3qCB3tOh9`.

A Resources

The code is available at https://github.com/guoruihan/MSM-Mut

B Comparison of MSA-Profile and MSM-Profile in Predicting Mutation Impact

The conservation of Multiple Sequence Alignment (MSA) in sequences is relatively strong, but when dealing with surface mutation data and predicting the impact of mutations on binding affinity, sequence-based profiles tend to be weaker. Our MSM-Profile addresses this limitation by leveraging local structure motifs, which can share information regardless of whether they are intra-chain or inter-chain.

Table 4 presents the performance results of our MSM-Profile and the traditional MSA-Profile on the SKEMPI2.0 dataset. As shown, our MSM-Profile exhibits a natural advantage in this task.

Table 4: Performance comparison between MSA-Profile and MSM-Profile on the SKEMPI2.0 dataset.

Category	Method	Pearson (P.S.)	Spearman (P.S.)	Pearson	Spearman
Profile	MSA-Profile	0.0826	0.0822	0.0159	0.0666
	MSM-Profile	0.1551	0.1766	0.1433	0.1739
	MSM-Profile (Filtered)	**0.1905**	**0.1886**	**0.1653**	**0.2063**

Additionally, on the s669 dataset, although the distributions of our MSM-Profile and MSA-Profile are similar, combining the two profiles results in an increase in Pearson correlation. This finding suggests that each profile provides complementary information, enhancing overall model performance when both are used together.

Table 5: Performance of MSA-Profile, MSM-Profile, and their combination on the s669 dataset.

Method	Pearson
MSA-Profile	0.17
MSM-Profile	0.19
MSA-Profile + MSM-Profile	**0.23**

C Ablation Study

C.1 Interpretability and Importance of Retrieved Motifs in Mutation Prediction

In this study, we use ESM-IF embeddings to construct a database and retrieve Multi-Structure-Motifs (MSM) to support mutation effect predictions. We demonstrate that using only the top-1 neighbor provides interpretability benefits, as retrieval with ESM-IF embeddings effectively identifies information valuable for mutation effect prediction.

To assess how varying the number of retrieved neighbors impacts predictive performance, we conducted experiments on the s669 dataset. Results indicate that increasing the number of neighbors improves the model's predictive ability. This finding suggests that the top-ranked neighbors contain diverse, biologically relevant information that enhances prediction accuracy.

C.2 Ablation Study on Retrieval and Pre-training

The table below presents an ablation study comparing the performance differences when retrieval is omitted, pre-training is omitted, or both are omitted. We observed that pre-training an IPA module is crucial. Without pre-training, the model lacks a proper initial distribution for the 20 types of amino acids at masked positions, which negatively impacts subsequent tasks.

Table 6: Performance of MSM-Mut with varying numbers of neighbors on the S669 dataset.

Method	Pearson	RMSE
MSM-Mut (w/o retrieval)	0.45	1.73
MSM-Mut (1 neighbor)	0.49	1.62
MSM-Mut (2 neighbors)	0.51	1.57
MSM-Mut (4 neighbors)	0.53	1.53
MSM-Mut (8 neighbors)	0.53	1.55
MSM-Mut (16 neighbors)	**0.54**	**1.51**
MSM-Mut (32 neighbors)	0.54	1.52
MSM-Mut (1024 neighbors)	0.51	1.63

Table 7: Ablation study on the impact of retrieval and pre-training on the S669 dataset.

Method	Pearson	RMSE
MSM-Mut (w/o retrieval, w/o pretrain)	0.37	2.82
MSM-Mut (w/o pretrain, 16 neighbors)	0.43	2.15
MSM-Mut (w pretrain, w/o retrieval)	0.45	1.73
MSM-Mut (w pretrain, 16 neighbors)	**0.54**	**1.51**

C.3 Impact of Retrieval Dataset Size on Model Performance

To illustrate the impact of retrieval dataset size on model performance, we present results based on random selections of approximately one-tenth, one-hundredth, and one-thousandth of the database. Specifically, we randomly selected 16 motifs from the top 100, 1000, and 10,000 motifs, simulating a reduction of the database to approximately one-tenth, one-hundredth, and one-thousandth of its original size. Our observations indicate that within the existing high-quality structure database, the size of the dataset is indeed critical, as it significantly influences the availability of high-quality data.

Table 8: Impact of database size on model performance for the S669 dataset.

Method	Pearson	RMSE
MSM-Mut (w/o retrieval)	0.45	1.73
MSM-Mut (random in top 10000)	0.43	2.03
MSM-Mut (random in top 1000)	0.52	1.57
MSM-Mut (random in top 100)	0.53	1.59
MSM-Mut (top 16 neighbors)	**0.54**	**1.51**

D Comparison of Continuous Backbone Angle Embedding (CBAE) and ESM-IF Embeddings

To provide a better comparison, we implemented a simpler encoding method called Continuous Backbone Angle Embedding (CBAE). For this approach, we defined the ϕ and ψ angles corresponding to amino acid i as ϕ_i and ψ_i. For each amino acid, we created an 18-dimensional embedding vector $[\phi_{i-4}, \psi_{i-4}, \ldots, \phi_{i+4}, \psi_{i+4}]$, where each value ranges from $[-\pi, \pi]$. This vector consists of the angles for the amino acid itself and the four consecutive amino acids before and after it in the sequence. We calculated the distance between two embeddings using the Manhattan distance, ensuring that retrieved structures have similar local sequence backbone structures.

Using this retrieval method, we tested model performance on the SKEMPI2.0 and S669 datasets, as shown in Tables 9 and 10. Results indicate that, while structures retrieved using ESM-IF embeddings tend to perform better, CBAE still maintains reasonable backbone similarity.

Our experiments showed that structures retrieved using ESM-IF embeddings tend to perform better than CBAE in capturing relevant mutation information, as ESM-IF embeddings can capture non-adjacent backbone atom positions to some extent. However, the differences between these methods

Table 9: Performance of MSM-Mut with and without CBAE retrieval on the SKEMPI2.0 dataset.

Method	Pearson (P.S.)	Spearman (P.S.)	Pearson	Spearman	RMSE
MSM-Mut (w/o retrieval)	0.4325	0.4031	0.6233	0.4954	1.6076
MSM-Mut (CBAE retrieval)	0.4619	0.4262	0.6524	0.5158	1.5531
MSM-Mut	**0.4736**	**0.4354**	**0.6814**	**0.5786**	**1.4703**

Table 10: Performance of MSM-Mut with and without CBAE retrieval on the S669 dataset.

Method	Pearson	RMSE
MSM-Mut (w/o retrieval)	0.45	1.73
MSM-Mut (CBAE retrieval)	0.52	1.55
MSM-Mut	**0.54**	**1.51**

are limited, as both embedding methods ensure a high degree of backbone similarity in retrieved structures.

E Visualization of Retrieved Structures

To further illustrate the retrieval process, we provide a figure with two subfigures detailing the alignment and visualization of highly similar local antibody structures obtained through Multiple Structure Motif Alignment.

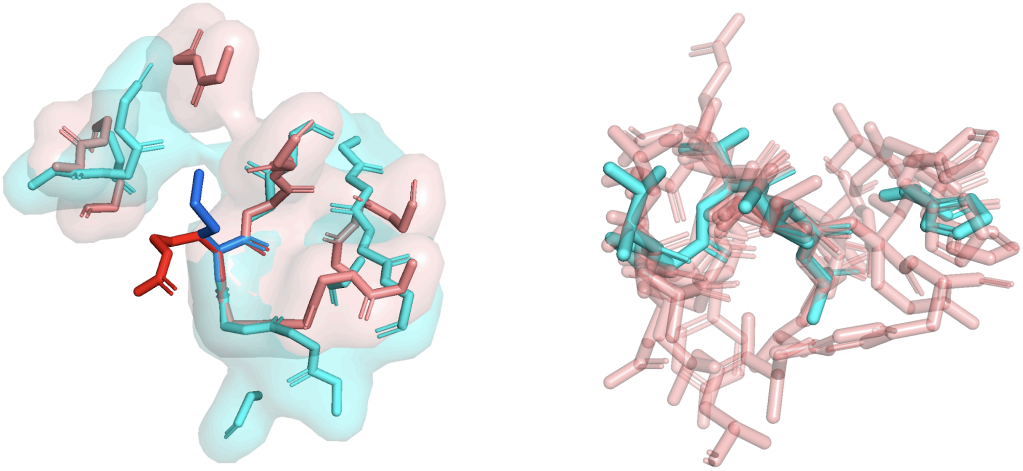

Figure 4: (a) Alignment centered on T chain 576 from 5KOV and H chain 103 from 7FAE, illustrating a high degree of similarity in their local structure motifs. (b) Visualization of the deduplicated retrieved local structure motifs, highlighting that similar structures are not limited to continuous chain segments; backbone atoms that are spatially close, even if not sequentially adjacent, also exhibit structural similarity.

F Details of Algorithms

Algorithm 1 MSM-IPA Information Fusion Algorithm

```
Require: Raw motif $\mathcal{M}_{\text{raw}}$, Retrieved motifs $\{\mathcal{M}_i\}_{i=1}^{L}$
Ensure: Merged motif representation $\tilde{\mathrm{s}}_i$
Combine raw and retrieved motifs:
    $\mathcal{M}_{\text{comb}} = \{\text{residue} \in \mathcal{M}_i, \forall i \in \{1, \ldots, L\}\} \cup \{\text{residue} \in \mathcal{M}_{\text{raw}}\}$
Encode single representations:
    $\boldsymbol{s}_{\text{raw}} = \text{Enc}_s(\mathcal{M}_{\text{raw}})$
    $\boldsymbol{s}_{\text{comb}} = \text{Enc}_s(\mathcal{M}_{\text{comb}})$
Encode pair representation:
    $\boldsymbol{z} = \text{Enc}_z(\mathcal{M}_{\text{raw}}, \mathcal{M}_{\text{comb}})$
Linear transformations:
    $q_i^h, \vec{q}_i^h = \text{LinearNoBias}(\boldsymbol{s}_{\text{raw}})$
    $k_i^h, v_i^h, \vec{k}_i^h, \vec{v}_i^h = \text{LinearNoBias}(\boldsymbol{s}_{\text{comb}})$
    $b_{ij}^h = \text{LinearNoBias}(\mathbf{z}_{ij})$
Compute attention weights:
    $a_{ij}^h = \text{softmax}_k \left( w_L \left( \frac{1}{\sqrt{c}} \mathbf{q}_i^{h^\top} \mathbf{k}_j^h + b_{ij}^h - \frac{\gamma^h w_C}{2} \sum_p \left\| T_i \circ \vec{\mathbf{q}}_i^{hp} - T_j \circ \vec{\mathbf{k}}_j^{hp} \right\|^2 \right) \right)$
Compute outputs:
    $\tilde{\mathbf{o}}_i^h = \sum_j a_{ij}^h \mathbf{z}_{ij}$
    $\mathbf{o}_i^h = \sum_j a_{ij}^h \mathbf{v}_j^h$
    $\vec{\mathbf{o}}_i^{hp} = T_i^{-1} \circ \sum_j a_{ij}^h \left( T_j \circ \vec{\mathbf{v}}_j^{hp} \right)$
Compute final representation:
    $\mathbf{s}_{\text{updated-raw}} = \text{Linear} \left( \text{concat}_{h,p} \left( \tilde{\mathbf{o}}_i^h, \mathbf{o}_i^h, \vec{\mathbf{o}}_i^{hp}, \left\| \vec{\mathbf{o}}_i^{hp} \right\| \right) \right)$
```

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present our method details in section 3 and domenstrate its effectness in section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Our algorithm procedure includes several third-party submodules, such as ESM-IF and CUHNSW, which may bound the performance of our method.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include any theory results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All datasets we used in evaluation are open-sourced.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets we used in evaluation are open-sourced, and we will release our model upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details of training and evaluation procedures are included in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We plot the complete distribution in Figure 2 to justify the design of our method.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We state the amount of computation resources to use when we introducing our method in section 3.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: As we discussed in the introduction section, our work may have positive societal impact since the proposed method can be applied to drug discovery and enzyme engineering.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: All resources used in this paper are free for academic usage.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.