Multimodal Protein Language Models for Flexibility Prediction and Loop Design

Tyler Verhaar [1] **Samuel W.K. Wong** [1]

Abstract

In this paper we use multimodal embeddings from the ESM-3 protein language model to predict protein flexibility. ESM-3's multi-modality allows for simultaneous integration of structural, sequence-based, and functional information into a residue-level embedding, yielding informative protein representations that can be used to predict protein flexibility. We demonstrate improved RMSF prediction compared to established methods such as CABS-flex. Additionally, we present a practical application of ESM-3 embeddings for loop design by sampling novel loop sequences conditioned on fixed structural and functional contexts. Such an approach may efficiently identify potential loop designs with increased rigidity.

1. Introduction

Protein loops are known for their structural flexibility and functional roles (Fiser et al., 2000; Malabanan et al., 2010). Loops are connecting regions between regular secondary structure and are often located on the protein surface. In the post-AlphaFold era (Jumper et al., 2021), it is still challenging to model long loops accurately (Stevens & He, 2022; Wang et al., 2024a), if treating crystal (static) structures from the Protein Data Bank (PDB, Berman et al., 2000) as the ground truth. However, loops may exhibit dynamic movement and have the flexibility to adopt multiple distinct conformations, and these may not be represented among PDB structures or be adequately sampled via prediction methods (Marks et al., 2018; Barozet et al., 2021). Accounting for loop dynamics has thus been recognized to be a key ingredient of emerging applications in protein design (Corbella et al., 2023).

A generally accepted approach to quantify the flexibility of protein backbones *in silico* is via root-mean-square fluctuations (RMSFs) from molecular dynamics (MD) simulation trajectories (Karplus & McCammon, 2002). Such MD-based RMSFs tend to more informative for protein dynamics compared to indirect measures from experimental techniques (e.g., B-factors) or plDDT outputs from AlphaFold (Ma et al., 2023; Vander Meersche et al., 2024). However, one hindrance to widely using MD for screening large numbers of protein sequences, e.g., as necessary in design applications, is its computational cost. Hence, various faster methods have been developed to predict per-residue RMSF, whether from sequence only (Narwani et al., 2019), from a given structure (Kurcinski et al., 2019; Nithin et al., 2024), or with the help of experimental data (e.g., cryo-EM maps, Song et al., 2024). With the advent of protein language models (pLMs) and structure encoders, RMSFs have also been predicted using different structure tokenizers (Yuan et al., 2025).

RMSFs of loop residues tend to be higher than those of regular secondary structure elements, e.g., as seen in the ATLAS database of MD simulations (Vander Meersche et al., 2024). However, this pattern is not uniform, as there are also many rigid loops (having low RMSFs) with structures that are easier to predict (Feng et al., 2021). Loop rigidity and stability are often associated, and hence loop rigidity is an important consideration in protein engineering (Nestl & Hauer, 2014), e.g., to increase the stability of an active site (Xie et al., 2014; Yu et al., 2017).

Computational design of protein loops to achieve specific structure and function has a long history (e.g., Hu et al., 2007; Kundert & Kortemme, 2019; Schmitz et al., 2021; Jiang et al., 2024), based on the idea of optimizing energy functions and was often laborious. Inverse folding based on pLMs (e.g., ESM-IF, Hsu et al., 2022) and deep learning models (e.g., ProteinMPNN, Dauparas et al., 2022) have demonstrated potential to accelerate the generation of plausible protein sequences, but are limited to a single modality, namely the sequence-structure relationship. The recent development of multimodal pLMs such as ESM-3 (Hayes et al., 2025), which can simultaneously condition on sequence, structure, and function (among other input tracks), may help further streamline this process. The specific applications of ESM-3 to predict loop flexibility (in terms of RMSFs) and to rigidify loops with a given structure and function, appear to be open directions for exploration.

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada. Correspondence to: Samuel W.K. Wong <samuel.wong@uwaterloo.ca>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

Thus, the contributions of this paper are two-fold. First, we leverage the multimodal embeddings of ESM-3 to develop a more accurate per-residue predictive model for RMSF. Via ablations, we examine the relative importance of the input modalities for prediction accuracy. Second, we apply our RMSF model to screen ESM-3-generated sequences (conditioned on structure and function) for rigidity, thereby providing a preliminary case study for the potential multimodal design of rigid loops.

2. Methods

2.1. Benchmark Dataset

We consider the main ATLAS database of MD simulations (Vander Meersche et al., 2024), using the latest version from November 18, 2024. ATLAS consists of 1,938 protein chains, representing the diversity of the PDB, for which standardized MD simulations have been run. RMSFs are reported for each protein residue, and we treat the average RMSFs of the three MD replicates as the ground truth labels.

2.2. Protein Annotations

For each PDB structure in the ATLAS dataset, we generated 8-class secondary structure (SS8) annotations via DSSP (Joosten et al., 2010) and solvent accessibility surface area (SASA) annotations via the algorithm implementation in biotite (Shrake & Rupley, 1973). We defined a loop to be four or more residues without regular secondary structures (DSSP classifications 'T', 'S', or '-'), as in Wang et al. (2024b). To annotate specific functional regions, we obtained residue numbers of binding sites by querying the PDBe-KB (Varadi et al., 2020).

2.3. ESM-3

ESM-3 (Hayes et al., 2025) is a multimodal pLM that employs a transformer encoder to process multiple input tracks: (1) the primary amino acid sequence, (2) 3D structure coordinates, (3) SS8, (4) per-residue SASA, and (5) functional annotations. These inputs are used to generate embeddings for each residue that leverage the context provided by the various input tracks. We leverage these residue-level embeddings as features to predict protein backbone flexibility, as measured by RMSF. We use the ESM-3 variant `esm3-sm-open-v1` to tokenize the input tracks and produce embeddings. This model comprises 1.4 billion parameters and is the smallest and most computationally efficient member of the ESM-3 family.

2.4. RMSF predictive model

We frame RMSF prediction as a per-residue regression task given multimodal ESM-3 embeddings, in order to quantify local disorder relevant for loop modelling.

For a protein of length L we obtain a feature matrix $\mathbf{x} \in \mathbb{R}^{L \times d_{\text{in}}}$[1] from the ESM-3 model (after removing the ESM-3 prefix (suffix) BOS (EOS) tokens). A linear projection maps each residue vector into the model space:

$$\mathbf{h}_0 = \mathbf{W}\mathbf{x} + \mathbf{b}, \qquad \mathbf{W} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{in}}}$$

Sinusoidal positional encodings $\mathrm{PE}(\cdot)$ (Vaswani et al., 2017) are added to preserve absolute indices $\mathbf{h}_0 \leftarrow \mathbf{h}_0 + \mathrm{PE}(L)$. The sequence passes through a 4-layer Transformer encoder; for $\ell = 0, \ldots, 3$,

$$\mathbf{h}_{\ell+1} = T_\ell(\mathbf{h}_\ell)$$
$$T_\ell = \mathrm{MHSA}(n_{\text{heads}}) \circ \mathrm{FFN}(d_{\text{ff}}) \circ \mathrm{DO}(p)$$

where **MHSA** denotes *multi-head self-attention*, **FFN** a position-wise *feed-forward network*, and **DO** the dropout operation. The Transformer encoder yields contextualised representations $\mathbf{h}_4 \in \mathbb{R}^{L \times d_{\text{model}}}$, where each residue embedding is informed by the entire sequence capturing long-range dependencies and structural context across the protein. A final two-layer MLP with hidden size d_{mlp}, ReLU activations and dropout p produces scalar predictions:

$$\hat{\mathbf{y}} = \mathrm{MLP}(\mathbf{h}_4) \in \mathbb{R}^L$$

The network contains $\approx$ 36M trainable parameters; ESM-3's 1.4B parameters remain frozen, making training much faster compared to a finetuning approach of ESM-3 and mitigating catastrophic forgetting (Kirkpatrick et al., 2017). The architecture parameters are listed in Table 1.

Table 1. Architecture hyper-parameters.

Component	Symbol	Value
Input dim.	d_{in}	1536
Model dim.	d_{model}	1024
FFN dim.	d_{ff}	2048
# Transformer layers	N_T	4
# Heads	n_{heads}	8
Dropout	p	0.20
MLP hidden dim.	d_{mlp}	512

RMSF values are strictly positive and left-skewed, posing a challenge from a learning perspective (Thompson & Fransson, 2016). We fit a Box–Cox power transform $\mathrm{BC}_\lambda(\cdot)$ given by $z = \left(y^\lambda - 1\right)/\lambda$ on the N training samples and train our model with a mean-squared error loss function in the transformed domain:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{z}_i - z_i\right)^2$$
$$z_i = \mathrm{BC}_\lambda(y_i), \qquad \hat{z}_i = \mathrm{BC}_\lambda(\hat{y}_i).$$

[1] Our ESM-3 configuration outputs $d_{\text{in}} = 1536$.

At inference time the inverse transform $\mathrm{BC}_\lambda^{-1}(\cdot)$ is used to recover the Å-scaled prediction.

The proteins in ATLAS (Vander Meersche et al., 2024) are randomly partitioned into training and validation sets in an 80:20 ratio, while ensuring that sequence similarity between the two sets remains minimal. Specifically, we aligned each pair of sequences using the pairwise2 module in BioPython (Cock et al., 2009). Pairs of proteins sharing at least 40% sequence similarity are kept together, and these pairs are then randomly assigned entirely to either the training or validation set, to help avoid data leakage from sequence homology. Once created, this train-validation split remains fixed across all experimental configurations. The model was trained over 50 epochs, which is sufficient for the validation loss to stabilize. Parameters are learned via the AdamW-optimizer (learning rate $\eta = 1 \times 10^{-4}$, weight decay 1×10^{-4}) (Loshchilov & Hutter, 2017). The learning rate is dynamically adjusted via the 'reduce on plateau' scheduler that reduces η by a factor $\gamma = 0.5$ when validation loss fails to improve for a specified number of epochs (scheduler patience $p_{\text{sch}} = 5$). To mitigate the risk of exploding gradients, gradient norms are clipped to $\|\nabla\theta\|_2 \leq 1.0$ (Pascanu et al., 2013). A fixed batch size of 32 is used throughout all experiments. The model hyper-parameters are summarized in Table 2.

Table 2. Optimizer and training hyper-parameters.

Description	Value
Initial learning rate η	1×10^{-4}
Weight decay	1×10^{-4}
LR scheduler	$\gamma = 0.5, p_{\text{sch}} = 2$
Batch size	32
Epochs	50
Gradient clipping	$\|\nabla\theta\|_2 \leq 1.0$

3. Results

3.1. RMSF prediction

We present the results of our RMSF prediction model evaluated on a held-out validation set comprised of 388 proteins (i.e., 20% of the ATLAS dataset), using the train-validation split described in Section 2.4.

As a performance benchmark, we use CABS-flex 2.0 (Kuriata et al., 2018), a coarse-grained simulator that employs Monte Carlo dynamics to efficiently estimate flexibility without requiring full MD simulations. While CABS-flex is significantly faster than traditional atomistic MD, it remains relatively computationally intensive for large proteins (up to ∼1 hr of CPU time). In contrast, our model incurs a one-time training cost, but inference is substantially more efficient (<1s per protein given ESM-3 embeddings), making it well-suited for large-scale screening once trained.

Table 3 presents the RMSF prediction performance across various combinations of ESM-3 input tracks: primary sequence (Seq), backbone coordinates (Struc), secondary structure (SS8), functional annotations (Func), and solvent accessibility (SASA). Prediction accuracy is quantified using root mean squared error (RMSE) and mean absolute error (MAE) between true and predicted RMSF values. To see the contributions of each input modality, we trained identical models—matched in architecture, training data, and hyperparameters—on ESM-3 embeddings derived from different subsets of input tracks. The fully-informed model (incorporating all available tracks) achieves the highest accuracy (RMSE = 1.184 Å, MAE = 0.562 Å). Omitting secondary structure (SS8) information (while retaining all other inputs) yields a moderate increase in RMSE to 1.279 Å, highlighting the benefit of explicit secondary structure context. Further removing functional annotations results in an additional modest error increase (RMSE = 1.308 Å); conversely, excluding solvent accessibility (SASA) yields a slightly larger degradation (RMSE = 1.339 Å), indicating solvent exposure information is potentially more influential than functional context in the absence of explicit secondary structure information. When examining the impact of removing a single input modality from the ESM-3 embeddings, Table 3 suggests that the relative importance of each track (from most to least impactful) is structural coordinates, sequence, functional annotations, and secondary structure (SS8). Generally, prediction accuracy decreases as fewer input modalities are included. In the simplest case, analogous to traditional single-modality approaches, we consider sequence-only and structure-only models. These yield the poorest performance among the ESM-3-based models, most notably under the sequence-only modality (RMSE = 1.555 Å, MAE = 0.782 Å); the corresponding structure-only model only provides a modest increase in performance, illustrating that these two modalities have comparable predictive power when used in isolation. These results thus highlight the benefit of multimodal embeddings. Overall, each model evaluated here outperforms the established baseline CABS-flex (RMSE = 1.769 Å, MAE = 0.982 Å).

The consistently large gap between RMSE and MAE—observed across all models (including CABS-flex)—indicates that residues with high RMSF values remain particularly difficult to predict, in line with previous findings (Narwani et al., 2019; Feng et al., 2025).

Using our model with all available input tracks, Figure 1 presents two representative examples from the validation set. Figure 1(a) illustrates RMSF predictions across a protein with relatively rigid loops (PDB ID: 1BX7). In this case, CABS-flex tends to overestimate flexibility while our model

Table 3. RMSF prediction accuracy (RMSE and MAE in Å) across 388 proteins under different ESM-3 input track configurations. Each row corresponds to a model trained using a specific combination of input tracks. The final row corresponds to CABS-flex predictions.

Seq	Struc	SS8	Func	SASA	RMSE	MAE
✓	✓	✓	✓	✓	1.184	0.562
✓	✓	×	✓	✓	1.279	0.641
✓	✓	✓	×	✓	1.301	0.642
✓	✓	×	×	✓	1.308	0.642
×	✓	✓	✓	✓	1.320	0.634
✓	✓	×	✓	×	1.339	0.707
✓	×	✓	✓	✓	1.340	0.640
✓	✓	✓	×	×	1.378	0.661
✓	✓	×	×	×	1.473	0.699
×	✓	×	×	×	1.536	0.704
✓	×	×	×	×	1.555	0.782
CABS-flex					1.769	0.982

more closely follows the MD-derived ground-truth RMSF. As a further example, Figure 1(b) shows RMSF predictions for a protein containing both flexible and rigid loops (PDB ID: 2FB5). Contrary to the previous case, CABS-flex tends to underestimate flexibility overall. Among loop regions, its predicted RMSFs can be too low (e.g., loops at residues 35–42 and 71–76) or too high (e.g., loop at residues 127–136), whereas our model accurately characterizes both low- and high-flexibility regions. These examples illustrate our model's superior predictive performance over the benchmark CABS-flex, both in terms of capturing the overall flexibility of the protein and the variations in flexibility across regions within a protein.

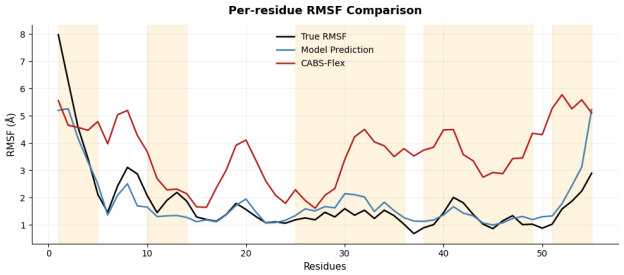
Furthermore, as the flexibility of loop regions tends to be more variable than in regular secondary structures, we segment predictive performance for loop versus non-loop regions (see Table 4). Our model consistently outperforms CABS-flex in both categories. For both methods, lower prediction errors are observed in non-loop residues, which is expected given the inherently higher flexibility (and thus prediction difficulty) of loop regions.

Table 4. RMSE and MAE comparison of the proposed model vs. CABS-flex across loop and non-loop regions in the validation set.

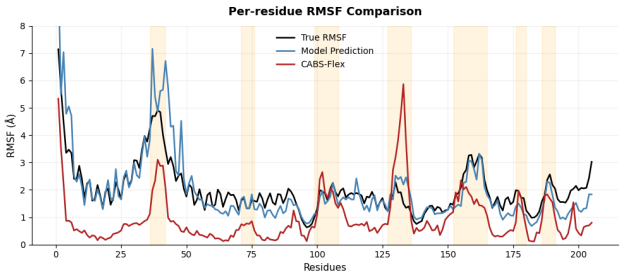
Region	Proposed model		CABS-flex	
	RMSE	MAE	RMSE	MAE
Non-loop	1.184	0.522	1.564	0.883
Loop	1.599	0.796	2.130	1.185

3.2. Rigid loop design

Next, we explore the feasibility of rigidifying loops through generative sampling with ESM-3. We select a representative



(a) Example 1 (PDB ID: 1BX7)



(b) Example 2 (PDB ID: 2FB5)

Figure 1. Per-residue RMSF predictions for two representative proteins from the validation set with varying levels of loop flexibility. Black curves show MD-derived ground-truth RMSF values, blue curves our model's predictions, and red curves CABS-flex predictions. Orange-shaded spans mark loop regions.

case study protein (PDB ID: 2VMC), focusing on the loop spanning residues 35–51. This loop was chosen due to its functional relevance (containing binding interface residues) and exhibiting elevated flexibility (higher RMSFs) relative to its surrounding context of regular secondary structures.

We implement a generative sampling procedure by masking only the amino acid sequence within the selected loop region, preserving all other contextual annotations (structural coordinates, SS8, SASA, and functional sites). ESM-3 is then prompted to generate novel sequences conditioned on these fixed contextual tracks which ensures structural and functional consistency.

For each of the $N = 500$ sampled sequences, we integrate the newly generated loop embeddings into the original protein context and predict loop flexibility using our RMSF model. The resulting RMSF distributions are shown in Figure 2. The generated loop sequences exhibit substantial variation in predicted RMSF with multiple sequences exhibiting reduced flexibility compared to the original MD-based RMSF profile. To assess the fidelity of the sampled sequences to the target loop structure, we also performed a forward folding pass with ESM-3; five candidates exhibited both increased rigidity and a reconstructed loop RMSD <3

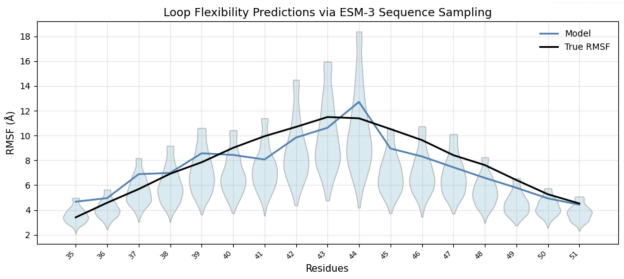


Figure 2. Loop flexibility distributions for the 2VMC protein region (residues 35–51). Light blue violins show per-residue distribution of RMSF predictions across 500 generated loop sequence variants (right tail clipped, 0–95th percentile shown). For the original structure (before loop design), the dark blue curve is the RMSF prediction of the model, and the black curve is the true RMSF (from MD simulations).

Å to the original one.[2] Results suggest that sequence-level sampling from multimodal embeddings is a promising strategy for rationally designing loops with targeted flexibility characteristics.

4. Discussion

In this study we have shown that the multimodal embeddings from the ESM-3 (Hayes et al., 2025) model can be leveraged effectively to predict protein backbone flexibility (quantified via residue-level RMSF values). Results suggest that incorporating sequence, structural coordinates, functional annotations, secondary structure, and solvent accessibility into the ESM-3 contextualized embeddings significantly enhances predictive performance relative to sequence-only or single-modality approaches. The case study on rigid loop design illustrates the potential of conditioning ESM-3 sequence generation on fixed additional structural, functional, solvency contexts to sample novel sequences. Preliminary results suggest this approach may be beneficial for designing loops, potentially leading to improved protein stability.

Multiple promising directions for future research remain. Firstly, incorporating additional outputs from ESM-3, such as structural logits or entropy measures of the distribution of sampled tokens may yield a more informative feature set from which a more accurate RMSF prediction model can be trained. Secondly, to improve loop design one could utilize a more sophisticated generative scheme such as including iterative refinement (Lin et al., 2024) or structure-conditioned sequence sampling integrated directly into the training procedure (Krapp et al., 2024).

[2]We find this reduction in loop accuracy to be a byproduct of ESM-3's structure encoder and decoder, which is limited in its ability to preserve detailed local structure, see Yuan et al. (2025).

Acknowledgements

This work was partially supported by Discovery Grant RGPIN-2019-04771 from the Natural Sciences and Engineering Research Council of Canada.

Impact Statement

This paper contributes methods aimed at advancing research in protein design by leveraging multimodal embeddings for residue-level flexibility prediction.

References

Barozet, A., Bianciotto, M., Vaisset, M., Siméon, T., Minoux, H., and Cortés, J. Protein loops with multiple meta-stable conformations: a challenge for sampling and scoring methods. *Proteins: Structure, Function, and Bioinformatics*, 89(2):218–231, 2021.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

Corbella, M., Pinto, G. P., and Kamerlin, S. C. Loop dynamics and the evolution of enzyme activity. *Nature Reviews Chemistry*, 7(8):536–547, 2023.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Feng, H., Zhao, J. Y., and Wei, G.-W. Multiscale differential geometry learning for protein flexibility analysis. *Journal of Computational Chemistry*, 46(7):e70073, March 2025. doi: 10.1002/jcc.70073. First published: 12 March 2025.

Feng, J.-J., Chen, J.-N., Kang, W., and Wu, Y.-D. Accurate structure prediction for protein loops based on molecular dynamics simulations with rsff2c. *Journal of Chemical Theory and Computation*, 17(7):4614–4628, 2021.

Fiser, A., Do, R. K. G., and Šali, A. Modeling of loops in protein structures. *Protein science*, 9(9):1753–1773, 2000.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M.,

et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Hu, X., Wang, H., Ke, H., and Kuhlman, B. High-resolution design of a protein loop. *Proceedings of the National Academy of Sciences*, 104(45):17668–17673, 2007.

Jiang, H., Jude, K. M., Wu, K., Fallas, J., Ueda, G., Brunette, T., Hicks, D. R., Pyles, H., Yang, A., Carter, L., et al. De novo design of buttressed loops for sculpting protein functions. *Nature chemical biology*, 20(8):974–980, 2024.

Joosten, R. P., Te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C., and Vriend, G. A series of pdb related databases for everyday needs. *Nucleic acids research*, 39(suppl_1):D411–D419, 2010.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Karplus, M. and McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature structural biology*, 9(9):646–652, 2002.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Krapp, L. F., Meireles, F. A., Abriata, L. A., Devillard, J., Vacle, S., Marcaida, M. J., and Dal Peraro, M. Context-aware geometric deep learning for protein sequence design. *Nature Communications*, 15(1):6273, 2024. doi: 10.1038/s41467-024-50571-y. URL https://doi.org/10.1038/s41467-024-50571-y.

Kundert, K. and Kortemme, T. Computational design of structured loops for new protein functions. *Biological chemistry*, 400(3):275–288, 2019.

Kurcinski, M., Oleniecki, T., Ciemny, M. P., Kuriata, A., Kolinski, A., and Kmiecik, S. Cabs-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics*, 35(4):694–695, 2019.

Kuriata, A., Gierut, A. M., Oleniecki, T., Ciemny, M., Kolinski, A., Kurcinski, M., and Kmiecik, S. Cabs-flex 2.0: a web server for fast simulations of flexibility of protein structures. *Nucleic Acids Research*, 46 (W1):W338–W343, 05 2018. ISSN 0305-1048. doi: 10.1093/nar/gky356. URL https://doi.org/10.1093/nar/gky356.

Lin, Z., Verkuil, R., Hayes, A., Meier, J., Carbonneau, M.-A., Gao, K., and Rives, A. Esmdiff: Conditional generation of protein structures from language models. *arXiv preprint arXiv:2410.18403*, 2024. URL https://arxiv.org/abs/2410.18403.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ma, P., Li, D.-W., and Brüschweiler, R. Predicting protein flexibility with alphafold. *Proteins: Structure, Function, and Bioinformatics*, 91(6):847–855, 2023.

Malabanan, M. M., Amyes, T. L., and Richard, J. P. A role for flexible loops in enzyme catalysis. *Current opinion in structural biology*, 20(6):702–710, 2010.

Marks, C., Shi, J., and Deane, C. M. Predicting loop conformational ensembles. *Bioinformatics*, 34(6):949–956, 2018.

Narwani, T. J., Etchebest, C., Craveur, P., Léonard, S., Rebehmed, J., Srinivasan, N., Bornot, A., Gelly, J.-C., and de Brevern, A. G. In silico prediction of protein flexibility with local structure approach. *Biochimie*, 165:150–155, 2019.

Nestl, B. M. and Hauer, B. Engineering of flexible loops in enzymes. *Acs Catalysis*, 4(9):3201–3211, 2014.

Nithin, C., Fornari, R. P., Pilla, S. P., Wroblewski, K., Zalewski, M., Madaj, R., Kolinski, A., Macnar, J. M., and Kmiecik, S. Exploring protein functions from structural flexibility using cabs-flex modeling. *Protein Science*, 33 (9):e5090, 2024.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR, 2013. URL https://arxiv.org/abs/1211.5063.

Schmitz, S., Ertelt, M., Merkl, R., and Meiler, J. Rosetta design with co-evolutionary information retains protein function. *PLoS Computational Biology*, 17(1):e1008568, 2021.

Shrake, A. and Rupley, J. A. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology*, 79(2):351–371, 1973.

Song, X., Bao, L., Feng, C., Huang, Q., Zhang, F., Gao, X., and Han, R. Accurate prediction of protein structural flexibility by deep learning integrating intricate atomic

structures and cryo-em density information. *Nature Communications*, 15(1):5538, 2024.

Stevens, A. O. and He, Y. Benchmarking the accuracy of alphafold 2 in loop structure prediction. *Biomolecules*, 12(7):985, 2022.

Thompson, W. H. and Fransson, P. On stabilizing the variance of dynamic functional brain connectivity time series. *arXiv preprint arXiv:1603.00201*, 2016. URL https://arxiv.org/abs/1603.00201.

Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.-C., and Galochkina, T. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

Varadi, M., Berrisford, J., Deshpande, M., Nair, S. S., Gutmanas, A., Armstrong, D., Pravda, L., Al-Lazikani, B., Anyango, S., Barton, G. J., et al. Pdbe-kb: a community-driven resource for structural and functional annotations. *Nucleic Acids Research*, 48(D1):D344–D353, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5998–6008, 2017. URL https://arxiv.org/abs/1706.03762.

Wang, T., Wang, L., Zhang, X., Shen, C., Zhang, O., Wang, J., Wu, J., Jin, R., Zhou, D., Chen, S., et al. Comprehensive assessment of protein loop modeling programs on large-scale datasets: prediction accuracy and efficiency. *Briefings in Bioinformatics*, 25(1):bbad486, 2024a.

Wang, T., Zhang, X., Zhang, O., Chen, G., Pan, P., Wang, E., Wang, J., Wu, J., Zhou, D., Wang, L., et al. Highly accurate and efficient deep learning paradigm for full-atom protein loop modeling with karmaloop. *Research*, 7:0408, 2024b.

Xie, Y., An, J., Yang, G., Wu, G., Zhang, Y., Cui, L., and Feng, Y. Enhanced enzyme kinetic stability by increasing rigidity within the active site. *Journal of Biological Chemistry*, 289(11):7994–8006, 2014.

Yu, H., Yan, Y., Zhang, C., and Dalby, P. A. Two strategies to engineer flexible loops for improved enzyme thermostability. *Scientific reports*, 7(1):41212, 2017.

Yuan, X., Wang, Z., Collins, M., and Rangwala, H. Protein structure tokenization: Benchmarking and new recipe. *arXiv preprint arXiv:2503.00089*, 2025.