LEARNING DYNAMIC PROTEIN REPRESENTATIONS AT SCALE WITH DISTOGRAMS

Nicolas Portal[1], Wissam Karroucha[2,3,4], Vincent Mallet[2,3,4,†], Massimiliano Bonomi[1,†]

[1]Institut Pasteur, Université Paris Cité, CNRS UMR 3528, Computational Structural Biology Unit, Paris, France;
[2]Mines Paris, PSL Research University, CBIO, Paris, France;
[3]Institut Curie, PSL Research University, Paris, France;
[4]INSERM, U1331, Paris, France;

[†] co-corresponding authors.
vincent.mallet@minesparis.psl.eu massimiliano.bonomi@pasteur.fr

ABSTRACT

Protein function and other biological properties often depend on structural dynamics, yet most machine-learning predictors rely on static representations. Physics-based molecular simulations can describe conformational variability but remain computationally prohibitive at scale. Generative models provide a more efficient alternative, though their ability to produce accurate conformational ensembles is still limited. In this work, we bypass expensive simulations by leveraging residue–residue distance probability distributions (distograms) from structure predictors such as AlphaFold2. Our approach provides a scalable way to encode dynamic information into protein representations, aiming to improve function prediction without explicit conformational sampling. All code required to reproduce the experiments presented in this work is publicly available at https://github.com/nicolas1805961/DistoDyn.

1 INTRODUCTION

Proteins perform a wide range of functions within cells. Recently, Machine Learning (ML) approaches have been developed to predict protein function, particularly their interactions with small molecules, RNA, DNA, and other macromolecules. Some methods leverage information from a protein's three-dimensional structure, while others rely on its amino acid sequence to take advantage of the more abundant sequence data. However, since structure encodes function more directly, structure-based methods generally outperform those based solely on sequence Yan et al. (2023).

Structure alone is often not enough to understand protein functions. Biological systems populate a variety of conformational states, and their functions often emerge from the interplay between structural and dynamic properties. For example, conformational transitions Nussinov et al. (2023), allosteric regulation Wodak et al. (2019), and ligand binding site flexibility Alghamedy et al. (2018) play a crucial role to achieve specific functions. Experimental techniques such as nuclear magnetic resonance spectroscopy and cryo-electron microscopy provide valuable insights into biomolecular dynamics but remain limited in their ability to fully characterize complex conformational landscapes at atomistic resolution. Consequently, simulation-based and integrative approaches currently represent the most effective strategies to characterize protein dynamics Hoff et al. (2024).

Simulation methods used to model protein conformational ensembles can be broadly categorized into physics-based approaches and generative models. The main limitation of physics-based methods, such as molecular dynamics (MD) simulations, lies in the prohibitive computational cost required to exhaustively sample complex conformational landscapes. Generative models are also not guaranteed to explore all relevant conformational states, and furthermore their accuracy ultimately depends on the quantity and quality of the data they have been trained on.

ML models aiming at predicting protein function ultimately need to account for the dynamic nature of proteins. While state-of-the-art ML approaches are not built to support multi-conformation in-

puts Hu & Ohue (2025); Cao et al. (2025), recent studies have explored the possibility to encode multiple conformations directly into protein representations. A recent approach aggregates different conformations into pairwise residue correlations, ultimately improving static representations Guo et al. (2025). While promising, the method depends on the availability of extensive MD datasets.

In this work, we take a different approach, completely sidestepping the explicit generation of protein conformations. Specifically, we leverage the probability distributions over residue–residue pairwise distances (distograms) predicted by modern structure prediction methods such as AlphaFold2 Jumper et al. (2021) and Boltz2 Passaro et al. (2025). These distributions have recently been shown to capture prediction uncertainty as well as structural dynamics Brotzakis et al. (2025); Schnapka et al. (2025); Sen et al. (2025); Savaş et al. (2025). Notably, distograms are obtained as byproducts of the structure prediction pipeline and are therefore orders of magnitude cheaper to compute than MD-derived correlation features (Figure 1).

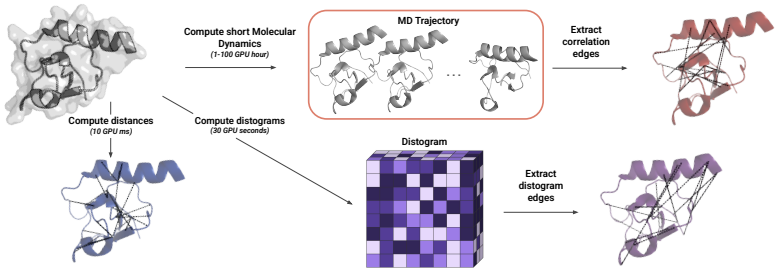

Figure 1: Overview of the protein representations used in this study. A protein can be represented as a static structure, computed on the fly. To capture dynamics, previous work often relies on compute-intensive MD simulations, encoding the results as additional edges. In contrast, we propose bypassing simulations by leveraging predicted distograms to incorporate dynamic information.

2 Related work

Generating protein conformational ensembles Currently, the most popular approach to generate conformational ensembles of proteins and other biomolecules is MD, which draws samples from the Boltzmann distribution given a model of the physico-chemical interactions, or force field. Due to the high computational cost of sampling complex conformational landscapes, several different strategies have been developed. Enhanced sampling techniques Hénin et al. (2022), such as metadynamics Laio & Parrinello (2002), have been developed to accelerate the exploration of conformational landscapes. More recently, large-scale MD datasets have been released Amaro, R. E. et al. (2025); Korlepara et al. (2024); Siebenmorgen et al. (2024); Vander Meersche et al. (2024), opening the door to training generative models directly on MD data.

Due to the expensive cost of MD, generative models have grown attention as a convenient alternative to generate conformational diversity Klein et al. (2023); Jing et al. (2024a); Costa et al. (2024); Lombard et al. (2025); Wolf et al. (2025); Jing et al. (2024b); Cheng et al. (2025), in some cases approximating the Boltzmann distribution Noé et al. (2019); Mardt et al. (2018); Lewis et al. (2025); Tan et al. (2025); Akhound-Sadegh et al. (2025); Lu et al. (2025); Roney et al. (2025). Another class of approaches modify modern structure prediction models to generate conformational diversity, for example by manipulating the multiple sequence alignment del Alamo et al. (2022); Wayment-Steele et al. (2024); Kalakoti & Wallner (2025; 2026) or by steering diffusion for diversity Richman et al. (2025). In some cases, these generative models have been shown to reproduce experimental observables such as NMR order parameters and small-angle X-ray scattering profiles, as well as temperature-dependent ensemble properties Lewis et al. (2025); Janson et al. (2025).

Protein representation learning Geometric deep learning encoders have been used for protein representation learning Isert et al. (2023), using various representations and architectures, such as 3D convolutional networks Jiménez et al. (2017); Weiler et al. (2018), sequence Rao et al. (2021), surfaces Gainza et al. (2020), graphs Aumentado-Armstrong (2018) and equivariant discrete networks Jing et al. (2021). In addition, some methods were developed ad-hoc to handle protein structure,

where protein properties are baked into the network Zhang et al. (2022); Hermosilla et al. (2020); Fan et al. (2022); Wang et al. (2022; 2025).

Multi-modal protein representations can encode different biological and computational priors. A well-studied combination is the use of sequence information along with a graph representation of the structure Hermosilla et al. (2020); Fan et al. (2022); Wu et al. (2023); Zhang et al. (2023). Some approaches include information derived from protein structures in the training of protein language models Bepler & Berger (2019); Heinzinger et al. (2024); Su et al. (2023). More recently, approaches combining different structure representations have demonstrated strong performances Somnath et al. (2021); Mallet et al. (2025); Zhang et al. (2024).

A few methods have emerged to incorporate MD simulations in ML-based representations. Some methods consider different conformations similarly as data augmentation for input protein structures Wu et al. (2022); Min et al. (2024); Libouban et al. (2025). Other approaches adopt a multi-instance learning framework Ilse et al. (2018), encoding each conformation independently and grouping their outputs Zankov et al. (2021); Kleiman et al. (2025). This increases inference time for a limited performance gain Criscuolo et al. (2024). Finally, some approaches directly aggregate the different conformations into a composite graph Chiang et al. (2022); Kalifa et al. (2025); Guo et al. (2025). All these approaches rely on datasets of MD trajectories.

3 Motivation and contributions

To encode dynamic information obtained from MD trajectories, Guo et al. (2025) proposed to use residue-residue motion correlation as a pairwise relationship. They enrich the radius graph traditionally used in graph-based protein representation learning, with this additional relationship. Training relational graph neural networks with enriched graphs enhances performance across various tasks.

Here, we follow a similar strategy by extracting information about dynamics from distograms. Distograms are distance distributions between C_β atoms (or C_α for glycines). These distributions, first introduced in methods presented at CAPRI13 Senior et al. (2020); Xu & Wang (2019), are now provided by most state-of-the-art structure predictors, such as AlphaFold2 and Boltz2, as a set of equally spaced bins spanning a distance range from 0.2 to 2.2 nm, with the last bin also capturing distances beyond the upper limit. Distograms can be generated at scale for training and deployed at inference time. By leveraging the full probability distribution over inter-node distances, they can capture both spatial proximity and structural uncertainty Brotzakis et al. (2025); Schnapka et al. (2025); Sen et al. (2025); Savaş et al. (2025).

Our key contributions can be summarized as follows:

- We enrich distance-based residue graphs, by extracting edges from distograms, as well as edge features, and encode these graphs with relational graph neural networks.
- We compare our enriched graphs to ones encoding MD trajectories, with enhanced results.
- We successfully apply our protocol to protein and RNA tasks without MD data.

In the following, we present the construction of our graphs and their processing in Section 4. We then compare our approach to the previously proposed approach to encode dynamics using correlations extracted from MD simulations, in cases where MD trajectories are available (Section 5). Finally, we apply our approach to the prediction of protein stability and RNA properties in Section 6. Our method opens the door to function prediction beyond static structures at limited computational cost.

4 Graph construction

Protein structures are represented by a graph $\mathcal{G}_\mathbf{P} = (\mathcal{V}, \mathcal{E})$. Depending on the task, $\mathcal{V}$ corresponds to either atoms or residues represented by their C_α atom. Node features are defined as one-hot encoding of the amino acid type or atom type. Traditionally, edges connect residues close in three-dimensional space, as determined by k-nearest neighbors or a radius cutoff. In this work, we used radius graph edges, defined as $\mathcal{E}_{dist} = \{(v_i, v_j) \mid d(v_i, v_j) < \tau_{dist}\}$, where $d(v_i, v_j)$ represents the distance between nodes v_i and v_j, and τ_{dist} is a distance threshold.

When an MD trajectory is available, we follow the approach of Guo et al. (2025) and compute the correlation between the motions of residues v_i and v_j, denoted as $|C_{ij}|$. We define correlation-based edges $\mathcal{E}_{corr} = \{(v_i, v_j) \mid |C_{ij}| > \tau_{corr}\}$, where τ_{corr} is a correlation threshold. Following Guo et al. (2025), τ_{dist} and τ_{corr} are set to 10Å and 0.3 respectively, (4.5Å and 0.6 for atomic-graphs). Finally, Chroma's developers Ingraham et al. (2023) suggested adding random edges in protein graphs. These edges enable long-range message passing and represent an important negative control for our approach. We sampled edges uniformly, taking as many samples as the number of distogram edges, to obtain $\mathcal{E}_{rand} \sim \mathcal{U}(\mathcal{V} \times \mathcal{V}), \quad \text{s.t. } |\mathcal{E}_{rand}| = |\mathcal{E}_{dist}|$.

4.1 DISTOGRAM-BASED EDGE FEATURES

Distograms $\mathcal{D}$ encode the probability distribution of the distance between each pair of residues (v_i, v_j), effectively representing prediction uncertainty and, possibly, variability due to the underlying dynamics Brotzakis et al. (2025); Schnapka et al. (2025); Sen et al. (2025); Savaş et al. (2025). In practice, these probabilities are discretized into B bins, $(b_1, b_2, \cdots, b_B)$, with $b_1 = 0$ and the convention that $b_{B+1} = \infty$. Distograms are therefore tensors of shape (N, N, B), where N is the number of residues in the graph. The value at position (i, j, k) corresponds to the predicted probability that the distance between nodes v_i and v_j falls into bin k, i.e. $\mathcal{D}(i, j, k) = \mathbb{P}[b_k \leq d(v_i, v_j) < b_{k+1}]$.

Based on these probabilities, we can define a distogram-based edge set composed of pairs predicted to be close with sufficient probability,

$$\mathcal{E}_{\text{disto}} = \{(v_i, v_j) \mid \mathbb{P}[d(v_i, v_j) \leq \delta] > \tau_{\text{disto}}\}, \tag{1}$$

where τ_{disto} is a probability threshold and δ is a distance cutoff for the distogram-based neighborhood. The value of δ is determined from the distribution of distances between neighboring residues as measured in the Protein Data Bank (PDB), and depends on the specific pairs (u, v) of amino-acid types as $\delta_{u,v} = \mu_{u,v} + 1.645\sigma_{u,v}$ where $\mu_{u,v}$ and $\sigma_{u,v}$ are tabulated for each amino-acid pairKamisetty et al. (2013). To be used in the discrete setting of distograms, $\delta_{u,v}$ is translated into a cutoff bin defined as the bin closest to the cutoff distance, $b_\delta^{uv} = \text{argmin}_{b \in \{b_1, \ldots, b_B\}} d(b, \delta_{uv})$.

The probability on the left hand size of Eq. 1 is computed as $\mathbb{P}[d(v_i, v_j) \leq \delta_{v_i, v_j}] = \sum_{k \leq b_\delta^{v_i, v_j}} \mathcal{D}(i, j, k)$. In our experiments, distograms were extracted from the confidence head of Boltz2 Passaro et al. (2025) after softmax normalization. Our results were obtained with $\tau_{\text{disto}} = 10^{-4}$, which corresponds to roughly the number of edges computed from MD correlations. In addition to enriching graphs with additional edges, we can use the probability distributions $\mathcal{D}(i, j) \in \Delta^B$ as edge features. This allows the model to capture both the expected distances and the dynamic variability between nodes.

4.2 GRAPH ANALYSIS

We analyze the different edge types introduced above using the MISATO dataset Siebenmorgen et al. (2024), a collection of thousands of short MD simulations of protein-ligand complexes. A detailed description of this dataset is provided in Appendix A.2. We first compute the intersection of the different edge types, counted as a fraction of the total number of possible edges (Figure 2A). There is a fair agreement between edge types, and notably distance edges (3.1% of possible pairs) are covered by the distogram and correlation edges (only 0.1% are specific to distance). We notice a moderate overlap between correlation edges and distogram edges.

To illustrate their difference, we use the `6nvd` protein, an enzyme involved in biotin synthesis in the bacteria responsible for tuberculosis. First, we show the edge lengths for each type in Figure 2C, and observe increased means for $\mathcal{E}_{corr}$ (9.63Å) and $\mathcal{E}_{disto}$ (9.57Å) compared to $\mathcal{E}_{dist}$ (7.16Å). This trend is even stronger on the whole dataset, with means values of 17.1Å, 19.2Å and 7.2Å, respectively. Hence, correlation-based and distogram-based edges both enable long-distance message passing.

Moreover, we notice that these edges have distinct patterns (Figure 2D). Notably, edges specific to correlation were mapped to the structure (Figure 2B, in red). These edges correspond to a helix and a beta-sheet motif that move in a correlated way. Therefore, the corresponding edges do not correspond to proximity in alternative conformations. On the other hand, distogram edges are found around distance edges, encoding fuzziness. We also identified a set of distogram-based edges (Figure 2B, in purple), which corresponds to a flexible loop with a highly dynamic secondary structure.

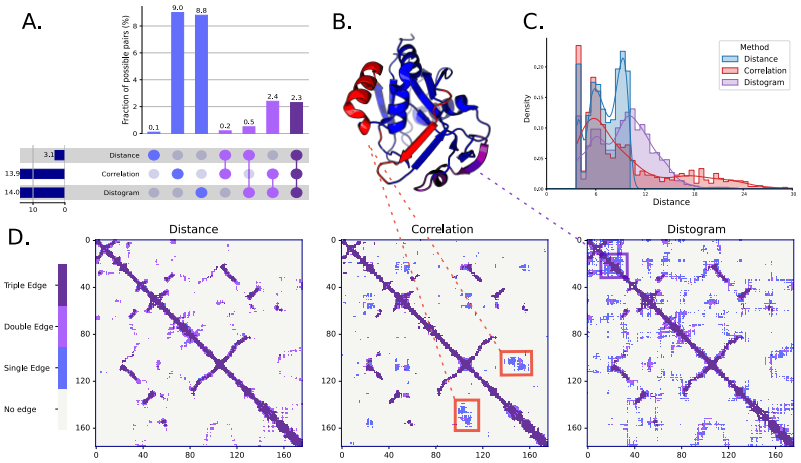

Figure 2: **A.** Upset plot showing the overlap between our different edge types ($\mathcal{E}_{dist}, \mathcal{E}_{corr}$, and $\mathcal{E}_{disto}$) across the ligand binding site task test set. Rows represent individual edge types, while columns indicate specific combinations. **B.** 3D structure of the `6nvd` protein (from our test set). Red residues correspond to the red box in the correlation adjacency matrix, and purple residues correspond to the box in the distogram adjacency matrix. **C.** Distance distributions for the different edge types introduced for the `6nvd` protein. **D.** Adjacency matrices for each edge type in the `6nvd` protein. Colors indicate whether a residue pair appears as a single, double, or triple edge—that is, whether it is present in one, two, or all three edge sets ($\mathcal{E}_{dist}, \mathcal{E}_{corr}, \mathcal{E}_{disto}$).

4.3 Relational graph neural networks

To effectively leverage the diverse types of edge features described above (distance, correlation, and distogram-based), we employ *relational graph neural networks* (R-GNNs) that are designed to handle graphs with multiple relation types. In this framework, each edge type corresponds to a distinct relation, allowing the model to learn edge-specific message passing rules and therefore capture the different structural and dynamic properties encoded by each edge feature.

We experiment with three R-GNN variants: Relational Graph Convolutional Networks (R-GCN) Schlichtkrull et al. (2018), Relational Graph Attention Networks (R-GAT) Busbridge et al. (2019), and a Relational version of the Equivariant Graph Neural Network (R-EGNN) Satorras et al. (2021). A detailed description of these models is provided in Appendix A.1. By using these relational GNN architectures, the model can capture the complementary information encoded in different edge types, while simultaneously leveraging node features to predict properties at the residue or atom level.

5 Validation on MD datasets

We now investigate the impact of encoding distogram information with relational graph networks. We start with scenarios where explicit dynamic information is available in the form of short MD trajectories, to compare to approaches encoding dynamics using $\mathcal{E}_{corr}$.

Experimental setup Experiments are conducted on two separate tasks: ligand binding site and ligand binding affinity prediction, as proposed in Guo et al. (2025). Following their work, binding sites are composed of protein residues closer than 10Å to any non-hydrogen atom of the ligand. Entire protein structures are used, resulting in graphs with an average number of nodes equal to 443.

For the binding affinity prediction task, only atoms belonging to the binding pocket are considered, resulting in much smaller graphs (47 residues on average). Moreover, since this task requires representing a ligand, we need to adapt the distogram edge set $\mathcal{E}_{disto}$ introduced in the previous section. Namely, we complement $\mathcal{E}_{disto}$ with atomic radius graph edges on the ligand side, and we connect protein and ligand nodes if one atom of the protein is closer than 8Å to an atom of the ligand.

The three models described above are trained and evaluated on each task using distance, correlation, and distogram-based edges, either individually or in combination as distinct relation types. When using distogram-based edges, the R-GAT and R-EGNN models are also trained with all distograms used as edge features. For both tasks, model architectures and hyperparameters are kept identical across different relation combinations, except for the dropout rate, which is independently tuned for each model to achieve optimal performance. Details about the learning and architecture hyperparameters are provided in Appendix A.3. The results of the R-EGNN model are presented in Table 1 while performance for the R-GCN and R-GAT models is available in the Appendix A.5.

Binding Site Prediction (Mean number of nodes = 443)				
Graph Type	Accuracy	Precision	Recall	F1 score
Distance	0.832	0.282	0.444	0.345
Distance + Correlation	**0.882**	**0.393**	0.350	0.370
Distance + Random Edges	0.830	0.316	**0.607**	0.416
Distance + Distogram	<u>0.861</u>	<u>0.376</u>	<u>0.606</u>	**0.464**
Distance + Distogram + Features	0.859	0.371	0.602	<u>0.459</u>
Binding Affinity Prediction (Mean number of nodes = 47)				
Graph Type	MAE	RMSE	Pearson R	Spearman R
Distance	1.296	1.623	0.666	0.642
Distance + Correlation	1.357	1.713	0.611	0.576
Distance + Random Edges	<u>1.211</u>	<u>1.502</u>	<u>0.721</u>	<u>0.698</u>
Distance + Distogram	1.275	1.560	0.699	0.674
Distance + Distogram + Features	**1.208**	**1.479**	**0.736**	**0.725**

Table 1: Results on the binding site prediction task (Top, average of 443 nodes) and binding affinity prediction task (Bottom, 47 nodes on average) for the R-EGNN model. We compare various dynamic-encoding approaches using $\mathcal{E}_{dist}$ alone or combined with $\mathcal{E}_{corr}$, $\mathcal{E}_{random}$, or $\mathcal{E}_{disto}$. Distogram-based edge features are also incorporated where compatible. Best-performing models are shown in bold, and second-best are <u>underlined</u>.

Overall performance of our approach Incorporating distogram-based edges and features consistently and significantly improves performance across tasks and architectures. Across tasks and models, using distogram always ranks first (16/24 settings) or second. Importantly, when ranking second, it comes as a close second, trailing only 2 accuracy points for R-GCN, but when ranking first it can induce significant boosts (such as 9.3 F1 points or 4 Spearman points on affinity for R-GAT).

Impact of distogram edges On the binding site task, which involves a larger number of nodes, introducing additional edges beside those based on distance brings significant benefits. Moreover, selecting the right edges plays an important role and we observe the following performance (informal) ordering $\mathcal{E}_{random} < \mathcal{E}_{corr} < \mathcal{E}_{disto}$. However, on the binding affinity task the impact of adding distogram edges is more nuanced. While our approach always represents an improvement over using $\mathcal{E}_{dist}$ only, this improvement is comparable to incorporating random edges. Both approaches outperform graphs that use $\mathcal{E}_{corr}$. We attribute this result to the limited size of the binding pockets (47 residues on average), which are already well-connected.

Impact of distogram edge features Incorporating distograms as edge features often results in clear performance improvements (Affinity, R-EGNN, +5 Spearman points), but in some case it can be negligible, or even detrimental. R-GAT benefits from including distogram edge features on the binding site task and suffers on the binding affinity task, while the opposite is true for R-EGNN. However, being able to use one or the other results in a clear improvement in all cases (their performance is not equivalent). We advise users to test both approaches for their specific problem.

Comparison to MD-correlation based approaches Finally, we compare our method to the correlation-based approach proposed by Guo et al. (2025), which relies on compute-intensive MD simulations. Our distogram-based approach clearly outperforms theirs, even without edge features. On the binding site task, distogram edges outperform correlation edges 9 out of 12 times, in some

cases with a substantial gap (*R-EGNN Recall*, 60 vs 35). On the binding affinity task, distogram edges were superior *across all models and metrics considered*.

It should be noted that we do not fully reproduce the results of Guo et al. (2025) that reported consistent improvement by adding correlation edges. In our experiments, the performance of distance-based models was much higher than the reported one, and adding correlation edges did not always improve performance. This was observed on the recall of the binding site task, and more generally on the binding affinity task. Moreover, while adding correlation edges was useful to predict binding sites, we found it to be less efficient than adding random edges to the graph. Overall, our results indicate that distograms capture complementary structural information that is not encoded by neither static distance thresholds nor correlation-based neighborhoods. This added information brings benefits to both residue-level classification and atom-level regression tasks.

6 Applications on generic datasets

In this section, we evaluate the performance of our approach in predicting the effects of protein mutations and various RNA properties. In both tasks, dynamics play a key role. However, MD simulations datasets are not available, which makes these tasks particularly challenging and highlights the need for accurate prediction methods that do not rely on such data.

6.1 Prediction of mutation effects

Protein stability is measured as the change in free energy ΔG between its folded and unfolded states. Upon mutations in the amino acid sequence, stability can be significantly affected. This variation is quantified by $\Delta\Delta G$, which can be used to compare the effect of different mutations, notably for tasks like protein design. As for other biological properties, dynamics in both folded and unfolded states plays an important role in determining protein stability Frellsen et al. (2025).

We applied our approach to the classic ThermoMPNN architecture Dieckhaus et al. (2024), which is designed to predict $\Delta\Delta G$ from the native structure and mutated sequence. This model is trained on the mega-scale dataset Tsuboyama et al. (2023), a large dataset of experimentally measured ΔG for relatively small proteins. ThermoMPNN relies on ProteinMPNN Dauparas et al. (2022), a popular graph-based structure encoder, to extract structure embeddings. In our experiments, we sought to enrich the ProteinMPNN graphs with distograms. Data and splits were held constant. Details about the training datasets, architectures, and procedures are reported in Appendix A.3.

Given the small size of the proteins studied here (56 nodes on average) and the high number of neighbors used by ProteinMPNN (48 nearest neighbors), little room is left for adding edges. Therefore, we only investigate adding edge features to the baseline model. In addition, we present an ablation experiment where only 16 neighbors are considered, so that additional edges can be introduced in the graph. We present the results of our experiments in Table 2.

Connectivity	Distogram Features	R^2	RMSE	Spearman R	Pearson R
Baseline (Topk = 48)		0.518	0.727	0.726	0.761
Baseline	✓	**0.577**	**0.681**	**0.749**	**0.788**
Ablation (Topk = 16)	✓	0.550	0.702	0.727	0.770
Ablation + 8 Random Edges	✓	<u>0.556</u>	<u>0.698</u>	<u>0.733</u>	<u>0.773</u>
Ablation + 8 Distogram Edges	✓	**0.557**	**0.697**	**0.740**	**0.778**

Table 2: Protein stability results. Best models in bold, second-best <u>underlined</u>.

Our results show that incorporating distogram features clearly improves performance across all metrics. The R^2 increases from 0.518 to 0.577, RMSE decreases from 0.7266 to 0.6808, and both Pearson and Spearman correlations improve. This indicates that distograms capture structural variability relevant for stability prediction, providing informative edge features that complement ProteinMPNN. Moreover, when using a reduced set of edges in the ablation experiment, introducing additional edges improve the results. This particularly holds for the Spearman correlation that increases from 0.727 to 0.733 with random edges, and to 0.740 with edges derived from distograms.

6.2 PREDICTION OF RNA PROPERTIES

Finally, we apply our approach to the prediction of various RNA properties. RNA molecules typically display greater flexibility than proteins due to their less hydrophobic nature. We first adapt the graph construction introduced for proteins, such that nodes represent nucleotides instead of amino acids. We apply our approach to the RNA-CM and RNA-Site tasks, introduced in the RNAglib benchmark Wyss et al. (2025). RNA nucleotides can be chemically modified, subtly altering their shape but crucially affecting their functions. The RNA-CM task aims to predict these chemical modifications from an RNA structure. The RNA-Site task is similar to the aforementioned ligand binding site prediction in proteins. These tasks have 57 and 64 nodes on average, respectively.

Considering the limited size of the RNA graphs, we only investigate the impact of adding distogram edge features to RNA graphs. To isolate the impact of dynamics captured by the distogram from static distance information, we systematically report the metrics obtained using distance features (a Gaussian radial basis function encoding of the pairwise residue distances using the same distance bins as those used in the distogram). We report the results across three different graph constructions: distance-based graph (a graph in which each residue node is connected to the nodes of its neighbors in Euclidean space), 2D+ graph (encoding backbone and canonical base pairs), and 2.5D graph (encoding backbone, canonical, and noncanonical base pairs, as defined in Leontis & Westhof (1998)). The results are presented in Table 3, and additional details are provided in Appendix A.4.

Connectivity	RNA-CM			RNA-Site		
	No Features	Distance Features	Distogram Features	No Features	Distance Features	Distogram Features
Distance	52.8	52.2	**54.8**	59.2	**61.9**	59.3
2D+	64.5	**65.5**	**65.5**	61.1	59.8	**62.1**
2.5D	66.7	64.6	**67.3**	60.7	60.4	**60.8**

Table 3: Results on RNA-CM and RNA-Site tasks (metric reported: balanced accuracy).

Adding distogram edge features results in consistent improvements across different graph constructions and tasks. In particular, the superior performance of distogram edge features over distance edge features underscores the ability of distograms to capture valuable information regarding RNA conformational variability. These results are particularly interesting given that state-of-the-art structure predictors are generally less accurate for RNA molecules than for proteins Kretsch et al. (2026), partly owing to the difficulty of tackling RNA flexibility.

7 DISCUSSION AND CONCLUSIONS

In this paper, we propose an efficient approach to encode proteins beyond a static structure by using distograms derived from Boltz2, a state-of-the-art protein structure predictor. We propose to enrich graph-based models representing protein structures with edges between residues predicted to be in close proximity according to their corresponding distogram. When the model is compatible, we also include the full distogram distribution as edge features.

For ligand binding site and binding affinity predictions, our approach outperforms methods that rely on computationally expensive molecular dynamics MD simulations. For the prediction of protein stability and RNA properties, where MD data were not available, our method also consistently improved performance. Most importantly, across all tasks, the overall best-performing method among different architectures is the one incorporating distograms.

These results highlight the potential of incorporating distogram-derived information to enhance protein as well as RNA representation learning. Nonetheless, our approach presents a few limitations. First, it relies on high-quality distograms, whose accuracy ultimately depends on the strength of coevolutionary signals in the multiple sequence alignment. Second, computing distograms for large proteins or macromolecular complexes can be computationally demanding; consequently, our analysis was restricted to relatively small systems. Finally, because distograms represent marginalized pairwise distance distributions, they provide an entangled, collective description of conformational variability and may not explicitly capture higher-order, multi-residue dynamic correlations.

An interesting future direction is the application of our method to systems lacking experimental structural information. This setting is particularly challenging, as the performance of structure prediction methods strongly depends on coevolutionary signals and the availability of related structures in the PDB. Consequently, it remains uncertain whether reliable and informative features can be extracted for systems with low-accuracy structural predictions. Nevertheless, our approach represents a promising step toward bridging the gap between the abundance of sequence data and the capabilities of dynamic structural modeling.

MEANINGFULNESS STATEMENT

Protein and RNA are inherently dynamic, yet most computational predictors of their function rely on static structures due to the high cost of molecular simulations. We present a scalable alternative that leverages distograms produced by modern structure predictors to encode dynamic information without explicit simulations. By integrating this information into graph-based models, our approach improves predictions across multiple biological tasks while remaining computationally efficient. This work lowers the barrier to incorporating dynamics into biomolecular machine learning and broadens access to function prediction methods beyond settings where molecular dynamics data are available.

ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 101086685 – bAIes). V.M. is supported by a Junior Springboard Prairie program, funded by the ANR project ANR-23-IACL-0008. W.K. is supported by Fondation pour la Recherche Médicale (FRM) with the following grant number: ECO202406019160. This work was performed using the Maestro cluster at Institut Pasteur and HPC resources from GENCI–IDRIS (Grant AD010315435R1).

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024.

Tara Akhound-Sadegh, Jungyoon Lee, Avishek Joey Bose, Valentin De Bortoli, Arnaud Doucet, Michael M Bronstein, Dominique Beaini, Siamak Ravanbakhsh, Kirill Neklyudov, and Alexander Tong. Progressive inference-time annealing of diffusion models for sampling from boltzmann densities. *arXiv:2506.16471*, 2025.

Fatemah Alghamedy, Jeevith Bopaiah, Derek Jones, Xiaofei Zhang, Heidi L Weiss, and Sally R Ellingson. Incorporating protein dynamics through ensemble docking in machine learning models to predict drug binding. *AMIA Summits on Translational Science Proceedings*, 2018:26, 2018.

Amaro, R. E. et al. The need to implement fair principles in biomolecular simulations. *Nature Methods*, 22:641–645, 2025.

Tristan Aumentado-Armstrong. Latent molecular optimization for targeted therapeutic design. *arXiv:1809.02032*, 2018.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv:1902.08661*, 2019.

Z Faidon Brotzakis, Shengyu Zhang, Mhd Hussein Murtada, and Michele Vendruscolo. Alphafold prediction of structural ensembles of disordered proteins. *Nature Communications*, 16:1632, 2025.

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. Relational graph attention networks. *arXiv:1904.05811*, 2019.

Duanhua Cao, Mingan Chen, Runze Zhang, Zhaokun Wang, Manlin Huang, Jie Yu, Xinyu Jiang, Zhehuan Fan, Wei Zhang, Hao Zhou, et al. Surfdock is a surface-informed diffusion generative

model for reliable and accurate protein–ligand complex prediction. *Nature Methods*, 22:310–322, 2025.

Kaihui Cheng, Ce Liu, Qingkun Su, Jun Wang, Liwei Zhang, Yining Tang, Yao Yao, Siyu Zhu, and Yuan Qi. 4d diffusion for dynamic protein structure prediction with reference and motion guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 93–101, 2025.

Yuan Chiang, Wei-Han Hui, and Shu-Wei Chang. Encoding protein dynamic information in graph representation for functional residue identification. *Cell Reports Physical Science*, 3, 2022.

Allan Dos Santos Costa, Ilan Mitnikov, Franco Pellegrini, Ameya Daigavane, Mario Geiger, Zhonglin Cao, Karsten Kreis, Tess Smidt, Emine Kucukbenli, and Joseph Jacobson. Equijump: Protein dynamics simulation via so (3)-equivariant stochastic interpolants. *arXiv:2410.09667*, 2024.

Emanuele Criscuolo, Rıza Özçelik, Derek van Tilborg, and Francesca Grisoni. The surprising ineffectiveness of molecular dynamics coordinates for predicting bioactivity with machine learning. *Chemrxiv-2024-rp81*, 2024.

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*, 378:49–56, 2022.

Diego del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife*, 11:e75751, 2022.

Henry Dieckhaus, Michael Brocidiacono, Nicholas Z Randolph, and Brian Kuhlman. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121:e2314853121, 2024.

Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2022.

Jes Frellsen, Maher M. Kassem, Tone Bengtsen, Lars Olsen, Kresten Lindorff-Larsen, Jesper Ferkinghoff-Borg, and Wouter Boomsma. Zero-shot protein stability prediction by inverse folding models: a free energy interpretation. *arXiv:12506.05596*, 2025.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17:184–192, 2020.

Pengkang Guo, Bruno Correia, Pierre Vandergheynst, and Daniel Probst. Boosting protein graph representations through static-dynamic fusion. *bioRxiv 2025.02.04.636233*, 2025.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6:lqae150, 11 2024.

Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *arXiv:2007.06252*, 2020.

S.E. Hoff, M. Zinke, N. Izadi-Pruneyre, and M. Bonomi. Bonds and bytes: The odyssey of structural biology. *Current Opinion in Structural Biology*, 84:102746, 2024.

Wenxing Hu and Masahito Ohue. Spatialppiv2: Enhancing protein–protein interaction prediction through graph neural networks with protein language models. *Computational and Structural Biotechnology Journal*, 27:508–518, 2025.

Jérôme Hénin, Tony Lelièvre, Michael R. Shirts, Omar Valsson, and Lucie Delemotte. Enhanced sampling methods for molecular dynamics simulations. *Living Journal of Computational Molecular Science*, 4:1583, Dec. 2022.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623:1070–1078, 2023.

Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023.

Giacomo Janson, Alexander Jussupow, and Michael Feig. Deep generative modeling of temperature-dependent structural ensembles of proteins. *Communication Chemistry*, 8:354, 2025.

José Jiménez, Stefan Doerr, Gerard Martínez-Rosell, Alexander S Rose, and Gianni De Fabritiis. Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33:3036–3042, 2017.

Bowen Jing, Stephan Eismann, Pratham N. Soni, and Ron O. Dror. Equivariant graph neural networks for 3d macromolecular structure. *arXiv:2106.03843*, 2021.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv:2402.04845*, 2024a.

Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. *Advances in Neural Information Processing Systems*, 37:40534–40564, 2024b.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Yogesh Kalakoti and Björn Wallner. Afsample2 predicts multiple conformations and ensembles with alphafold2. *Communications Biology*, 8:373, 2025.

Yogesh Kalakoti and Björn Wallner. Afsample3: Generating and selecting multiple conformational states with alphafold3. *bioRxiv 2026.01.16.699904*, 2026.

Dan Kalifa, Eric Horvitz, and Kira Radinsky. Learning protein representations with conformational dynamics. *bioRxiv 2025.10.06.680789*, 2025.

Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110:15674–15679, 2013.

Diego E Kleiman, Jiangyan Feng, Zhengyuan Xue, and Diwakar Shukla. Esmdynamic: Fast and accurate prediction of protein dynamic contact maps from single sequences. *bioRxiv 2025.08.20.671365*, 2025.

Leon Klein, Andrew Foong, Tor Fjelde, Bruno Mlodozeniec, Marc Brockschmidt, Sebastian Nowozin, Frank Noé, and Ryota Tomioka. Timewarp: Transferable acceleration of molecular dynamics by learning time-coarsened dynamics. *Advances in Neural Information Processing Systems*, 36:52863–52883, 2023.

Divya B Korlepara, Vasavi CS, Rakesh Srivastava, Pradeep Kumar Pal, Saalim H Raza, Vishal Kumar, Shivam Pandit, Aathira G Nair, Sanjana Pandey, Shubham Sharma, et al. Plas-20k: Extended dataset of protein-ligand affinities from md simulations for machine learning applications. *Scientific Data*, 11:180, 2024.

Rachael C Kretsch, Alissa M Hummer, Shujun He, Rongqing Yuan, Jing Zhang, Thomas Karagianes, Qian Cong, Andriy Kryshtafovych, and Rhiju Das. Assessment of nucleic acid structure prediction in casp16. *Proteins: Structure, Function, and Bioinformatics*, 94:192–217, 2026.

Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99:12562–12566, 2002.

Neocles B Leontis and Eric Westhof. Conserved geometrical base-pairing patterns in rna. *Quarterly Reviews of Biophysics*, 31:399–455, 1998.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389:eadv9817, 2025.

Pierre-Yves Libouban, Camille Parisel, Maxime Song, Samia Aci-Sèche, Jose C Gómez-Tamayo, Gary Tresadern, and Pascal Bonnet. Spatio-temporal learning from molecular dynamics simulations for protein–ligand binding affinity prediction. *Bioinformatics*, 41:btaf429, 2025.

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of Chemical Research*, 50: 302–309, 2017.

Valentin Lombard, Sergei Grudinin, and Elodie Laine. Petimot: A novel framework for inferring protein motions from sparse data using se (3)-equivariant graph neural networks. *arXiv:2504.02839*, 2025.

Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Aurélie Lozano, Vijil Chenthamarakshan, Payel Das, and Jian Tang. Aligning protein conformation ensemble generation with physical feedback. *arXiv:2505.24203*, 2025.

Vincent Mallet, Yangyang Miao, Souhaib Attaiki, Bruno Correia, and Maks Ovsjanikov. Atomsurf: Surface representation for learning on protein structures. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature Communications*, 9:5, 2018.

Yaosen Min, Ye Wei, Peizhuo Wang, Xiaoting Wang, Han Li, Nian Wu, Stefan Bauer, Shuxin Zheng, Yu Shi, Yingheng Wang, et al. From static to dynamic structures: Improving binding affinity prediction with graph-based deep learning. *Advanced Science*, 11:2405404, 2024.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365:eaaw1147, 2019.

Ruth Nussinov, Yonglan Liu, Wengang Zhang, and Hyunbum Jang. Protein conformational ensembles in function: roles and mechanisms. *RSC Chem. Biol.*, 4:850–864, 2023.

Carlos Oliver, Vincent Mallet, Roman Sarrazin Gendron, Vladimir Reinharz, William L Hamilton, Nicolas Moitessier, and Jérôme Waldispühl. Augmented base pairing networks encode rna-small molecule binding preferences. *Nucleic Acids Research*, 48:7690–7699, 2020.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv 2025.06.14.659707*, 2025.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.

Daniel D Richman, Jessica Karaguesian, Carl-Mikael Suomivuori, and Ron O Dror. Unlocking hidden biomolecular conformational landscapes in diffusion models at inference time. *arXiv:2512.03312*, 2025.

James P Roney, Chenxi Ou, and Sergey Ovchinnikov. Protein diffusion models as statistical potentials. *bioRxiv 2025.12.09.693073*, 2025.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Büşra Savaş, Ayşe Berçin Barlas, and Ezgi Karaca. Exploring the potential of alphafold distograms for predicting binding-induced hinge motions. *bioRxiv 2025.07.25.666757*, 2025.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Vincent Schnapka, Tatiana I. Morozova, Samiran Sen, and Massimiliano Bonomi. Atomic resolution ensembles of intrinsically disordered proteins with alphafold. *bioRxiv 2025.06.18.660298*, 2025.

Samiran Sen, Samuel E. Hoff, Tatiana I. Morozova, Vincent Schnapka, and Massimiliano Bonomi. Advancing in silico drug design with bayesian refinement of alphafold models. *bioRxiv 2025.06.25.661454*, 2025.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:706–710, 2020.

Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Kieran Didi, André Santos Dias Mourão, Radosław Kitel, Pietro Liò, Stefan Kesselheim, Marie Piraud, et al. Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature Computational Science*, 4:367–378, 2024.

Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems*, 34:25244–25255, 2021.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv 2023.10.01.560349*, 2023.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of Chemical Information and Modeling*, 59:895–913, 2018.

Charlie B Tan, Avishek Joey Bose, Chen Lin, Leon Klein, Michael M Bronstein, and Alexander Tong. Scalable equilibrium sampling with sequential boltzmann generators. *arXiv:2502.18462*, 2025.

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*, 620:434–444, 2023.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52:D384–D392, 2024.

Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3d graph networks. *arXiv:2207.12600*, 2022.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48:4111–4119, 2005.

Zhiyu Wang, Arian Jamasb, Mustafa Hajij, Alex Morehead, Luke Braithwaite, and Pietro Liò. Topotein: Topological deep learning for protein representation learning. *arXiv:2509.03885*, 2025.

Hannah K. Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M. Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature*, 625:832–839, 2024.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31, 2018.

Shoshana J. Wodak, Emanuele Paci, Nikolay V. Dokholyan, Igor N. Berezovsky, Amnon Horovitz, Jing Li, Vincent J. Hilser, Ivet Bahar, John Karanicolas, Gerhard Stock, Peter Hamm, Roland H. Stote, Jerome Eberhardt, Yassmine Chebaro, Annick Dejaegere, Marco Cecchini, Jean-Pierre Changeux, Peter G. Bolhuis, Jocelyne Vreede, Pietro Faccioli, Simone Orioli, Riccardo Ravasio, Le Yan, Carolina Brito, Matthieu Wyart, Paraskevi Gkeka, Ivan Rivalta, Giulia Palermo, J. Andrew McCammon, Joanna Panecka-Hofman, Rebecca C. Wade, Antonella Di Pizio, Masha Y. Niv, Ruth Nussinov, Chung-Jung Tsai, Hyunbum Jang, Dzmitry Padhorny, Dima Kozakov, and Tom McLeish. Allostery in its many disguises: From theory to applications. *Structure*, 27:566–578, 2019.

Nicolas Wolf, Leif Seute, Vsevolod Viliuga, Simon Wagner, Jan Stühmer, and Frauke Gräter. Learning conformational ensembles of proteins based on backbone geometry. *arXiv:2503.05738*, 2025.

Fang Wu, Shuting Jin, Yinghui Jiang, Xurui Jin, Bowen Tang, Zhangming Niu, Xiangrong Liu, Qiang Zhang, Xiangxiang Zeng, and Stan Z Li. Pre-training of equivariant graph matching networks with conformation flexibility for drug binding. *Advanced Science*, 9:2203796, 2022.

Fang Wu, Lirong Wu, Dragomir Radev, Jinbo Xu, and Stan Z Li. Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, 6:876, 2023.

Luis Wyss, Vincent Mallet, Wissam Karroucha, Karsten Borgwardt, and Carlos Oliver. A comprehensive benchmark for rna 3d structure-function modeling. *arXiv:2503.21681*, 2025.

Jinbo Xu and Sheng Wang. Analysis of distance-based protein structure prediction by deep learning in casp13. *Proteins: Structure, Function, and Bioinformatics*, 87:1069–1081, 2019.

Tian-Ci Yan, Zi-Xuan Yue, Hong-Quan Xu, Yu-Hong Liu, Yan-Feng Hong, Gong-Xing Chen, Lin Tao, and Tian Xie. A systematic review of state-of-the-art strategies for machine learning-based protein function prediction. *Computers in Biology and Medicine*, 154:106446, 2023.

Dmitry V Zankov, Mariia Matveieva, Aleksandra V Nikonenko, Ramil I Nugmanov, Igor I Baskin, Alexandre Varnek, Pavel Polishchuk, and Timur I Madzhidov. Qsar modeling based on conformation ensembles using a multi-instance learning approach. *Journal of Chemical Information and Modeling*, 61:4913–4923, 2021.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv:2203.06125*, 2022.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. A systematic study of joint representation learning on protein sequences and structures. *arXiv:2303.06275*, 2023.

Zuobai Zhang, Pascal Notin, Yining Huang, Aurelie C Lozano, Vijil Chenthamarakshan, Debora Marks, Payel Das, and Jian Tang. Multi-scale representation learning for protein fitness prediction. *Advances in Neural Information Processing Systems*, 37:101456–101473, 2024.

A Appendix

A.1 Detailed model definition

Relational Graph Convolutional Network (R-GCN) R-GCN extends the standard GCN by introducing relation-specific weight matrices. Let $h_i^{(l)}$ be the representation of node i at layer l. This representation is modulated by relation-specific weight matrices $W_r^{(l)}$, with the convention that $r = 0$ corresponds to self-loops. The message sent by a node i across relation r is defined as

$$m_{r,i}^{l} = W_r^{(l)} h_i^{(l)}. \tag{2}$$

Message passing updates the node representation as:

$$h_i^{(l+1)} = \sigma\left(m_{0,i}^{(l)} + \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \frac{1}{c_{i,r}} m_{r,j}^{(l)} \right), \tag{3}$$

where $\mathcal{N}_i^r$ is the set of neighbors of node i under relation r, $c_{i,r}$ is a normalization constant, and σ is a non-linear activation function.

Relational Graph Attention Network (R-GAT) R-GAT introduces relation-specific attention coefficients to weight messages from different neighbors:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \alpha_{ij}^r m_{r,j}^{(l)} \right), \tag{4}$$

where the attention coefficients α_{ij}^r are computed as

$$\alpha_{ij}^r = \frac{\exp\left(\text{LeakyReLU}\left(a_r^\top [m_{r,i}^{(l)} \,||\, m_{r,j}^{(l)}]\right)\right)}{\sum_{k\in\mathcal{N}_i^r} \exp\left(\text{LeakyReLU}\left(a_r^\top [m_{r,i}^{(l)} \,||\, m_{r,k}^{(l)}]\right)\right)}, \tag{5}$$

with a_r as a learnable attention vector for relation r and $||$ denoting vector concatenation.

Relational Equivariant Graph Neural Network (R-EGNN) To extend EGNN to handle multiple edge types, we create a relational variant (R-EGNN) by processing each relation type separately. Specifically, for each edge type r (e.g., distance-based, correlation-based, or distogram-based), a separate EGNN processes the corresponding subgraph, producing relation-specific node and coordinate updates. The outputs from all relations are then combined via summation to obtain the final node embeddings and coordinate updates.

Formally, the node update for relation r follows the standard EGNN message passing:

$$m_{ij}^r = \phi_e(h_i^{(l)}, h_j^{(l)}, \|x_i^{(l)} - x_j^{(l)}\|^2, e_{ij}^r), \tag{6}$$

$$\Delta x_i^r = \frac{1}{|\mathcal{N}_i^r|} \sum_{j\in\mathcal{N}_i^r} (x_i^{(l)} - x_j^{(l)}) \phi_x(m_{ij}^r), \tag{7}$$

$$h_i^{(l+1),r} = \phi_h\Big(h_i^{(l)}, \sum_{j\in\mathcal{N}_i^r} m_{ij}^r\Big), \tag{8}$$

where $h_i^{(l)}$ and $x_i^{(l)}$ are the feature vector and coordinates of node i at layer l, e_{ij}^r is the edge feature vector for relation r, and ϕ_e, ϕ_x, ϕ_h are learnable MLPs.

The final node representation and coordinates are obtained by summing the contributions across all relations:

$$h_i^{(l+1)} = \sum_{r\in\mathcal{R}} h_i^{(l+1),r}, \tag{9}$$

$$x_i^{(l+1)} = x_i^{(l)} + \sum_{r\in\mathcal{R}} \Delta x_i^r. \tag{10}$$

This design allows the model to leverage relation-specific information while maintaining the equivariance properties of EGNN: the predictions are invariant to translations and equivariant to rotations of the input coordinates. By incorporating multiple edge types (distance, correlation, and distogram), R-EGNN captures complementary structural and dynamic information while directly processing 3D node coordinates.

A.2 DATASETS

The MISATO dataset Siebenmorgen et al. (2024) is used for the binding site and binding affinity prediction task. This dataset contains 19443 protein-ligand complexes extracted from PDBbind Su

et al. (2018); Liu et al. (2017); Wang et al. (2005). In the binding site prediction task, the data is partitioned following the split provided with the MISATO dataset. In the binding affinity prediction task, models are trained on the PDBbind 2020 refined set and tested on the core set which contain respectively 5318 and 285 protein–ligand complexes retained after extensive filtering to ensure the quality of both binding affinity data and crystal structures.

Experiments for the protein stability prediction task are conducted on the mega-scale dataset Tsuboyama et al. (2023). This dataset contains 776000 high-quality folding stability measurements spanning all single amino acid variants and selected double mutants across 331 natural and 148 de novo–designed protein domains, each 40–72 amino acids long. After removing duplicate and unreliable stability measurements, we end up with 577313 samples. The data are split into training, validation, and test sets following the partitioning scheme used by Dieckhaus et al. (2024). After removing double mutants, the final dataset contains 298 wild types, accounting for 443,906 protein sequences.

A.3 Implementation details

The code necessary to run our experiments and reproduce our results can be found at `https://anonymous.4open.science/r/DistoDyn-55AB/README.md`.

Our distograms were generated using Boltz with default parameters. The MSAs were generated on the fly. For RNA, Boltz does not recommend using MSAs. Alignments were generated with the following command line:

```
boltz predict "${INPUT_PATH}" --use_msa_server --model boltz2
    --out_dir "${OUTPUT_DIR}"
```

For the binding site and binding affinity prediction, we used the models as defined by Guo et al. (2025), but we grid-searched dropout rates over the following values : $\{0.0, 0.1, 0.2, 0.3, 0.4\}$ for all models.

For the finetuning of the protein stability tasks, all layers are kept frozen except the initial edge embedding layers, light attention layers and heads. We chose to fine-tune ThermoMPNN instead of training it from scratch. To do so, we kept the ProteinMPNN encoder layers frozen in early epochs and gradually unfroze them. Out of fairness, we also tried to fine-tune their model (allowing for more computations), which ultimately did not affect performance. Different learning rate scheduling are used for these layers. We tried training our models with 6 different learning rate strategies detailed in Table 4. We report the best test performance. In addition, the first encoder layer of ProteinMPNN is unfrozen after epoch 10. Then between epoch 10 and 20 the second encoder layer of ProteinMPNN is also unfrozen. Finally, the last encoder layer is unfrozen from epoch 20 until the end of training.

Learning rate setup		A	B	C	D	E	F
Encoder	Initial	10^{-4}	10^{-4}	10^{-5}	10^{-4}	10^{-3}	10^{-3}
	Final	10^{-5}	10^{-5}	10^{-6}	10^{-5}	10^{-4}	10^{-5}
Edge	Initial	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-3}	10^{-3}
	Final	10^{-5}	10^{-4}	10^{-5}	10^{-5}	10^{-4}	10^{-4}
Head	Initial	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-3}	10^{-3}
	Final	10^{-5}	10^{-4}	10^{-5}	10^{-5}	10^{-4}	10^{-4}
Light Attention	Initial	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-3}	10^{-3}
	Final	10^{-5}	10^{-4}	10^{-5}	10^{-5}	10^{-4}	10^{-4}

Table 4: Setups used for the protein stability prediction task

A.4 RNA SETTINGS

A.4.1 DETAILS ABOUT GRAPH CONSTRUCTION

We follow the graph nomenclature by Wyss et al. (2025). Three families of graph representations are being benchmarked: distance-based graphs, 2D+ graphs and 2.5D graphs. In all graph representations used, a graph denotes a connected component of an RNA, and a node denotes a residue.

In distance-based graphs, each residue is connected to all residues located within a neighborhood of radius 8.0Å (self-loops are removed). In order to compute residue coordinates, following Boltz2 Passaro et al. (2025) and AlphaFold 3 Abramson et al. (2024), we choose the C2 atom of the base as representative atom of residues with pyrimidine bases and the C4 atom of the base as representative atom of residues with purine bases. This parameterization has the advantage of aligning with the atoms used in the distogram computation by Boltz2 and capturing information regarding base pairing and base stacking, two driving forces of RNA structure and interactions with proteins and small molecules. The 8.0Å cutoff was chosen after a careful examination of graph connectivities for various possible values. In our datasets, when using an 8.0Å cutoff for graph construction, each node has on average 7.7 neighbors, which is reasonable given the relatively small size of our graphs (57 and 64 nodes on average for RNA-CM and RNA-Site Wyss et al. (2025), respectively). In this setting, we note that there is only one edge type.

In 2D+ graphs, three distinct edge types are added. A first edge type is created for 5' to 3' backbone connections (phosphodiester bonds), a second edge type is added for 3' to 5' backbone connections, and a third edge type (bidirectional) for canonical base pairs. These graphs are therefore directed.

In 2.5D graphs, 20 edge types are created: 5' to 3' backbone connections, 3' to 5' backbone connections, and one for each of the base pair geometries as defined by the Leontis-Westhof nomenclature of base pairs Leontis & Westhof (1998). These edges were shown to improve the performance of machine learning approaches for drug design applications Oliver et al. (2020), and on a recent benchmark for RNA 3D structure-function modeling Wyss et al. (2025).

For each of these settings, we run experiments without any edge features, with distogram-based edge features, and with distance-based edge features. In the setting without edge features, the RGAT acts upon edge types and node features only. In the setting with edge features, the RGAT processes edge features, in addition to node features and edge types. Edge features are processed within the attention computation mechanism of the RGAT layer.

For each edge of the graph connecting any two residues, its distogram-based edge features are the probabilities of the pairwise residue distances being in each of the 64 distance bins of the distogram. Its distance-based edge features are the encoding of the pairwise residue distances through Gaussian radial basis functions using the exact same distance bins as Boltz's distograms.

A.4.2 ARCHITECTURE AND TRAINING HYPERPARAMETERS

The experiments were carried out using relational graph attention networks (RGAT). This architecture was chosen for its ability to natively handle simultaneously distinct edge types and continuous edge features. For a fair comparison across settings, we also used the RGAT architecture for the experiments carried out without distance- or distogram-based edge features. The implementation was performed using RNAglib Tasks Wyss et al. (2025).

The hyperparameters chosen are reported below:

The loss function is weighted by the relative occurrences of the classes in our datasets.

All results reported in Table 3 represent the mean performance computed across three independent trials with distinct random seeds.

Table 5: Hyperparameter settings used for the experimental evaluation on RNA structures

Hyperparameter	**Value**
Number of Layers	3
Hidden Channels	128
Training Epochs	200
Batch Size	8
Dropout Rate	0.5
Loss Weights	Ratio-based
Learning Rate	$\{10^{-3}, 10^{-4}\}$ (Tuned via validation)

A.5 FULL MD RESULTS

Binding Site Prediction (Mean number of nodes = 443)					
Model	Graph Type	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 score (↑)
R-GCN	Distance	0.761	0.187	<u>0.420</u>	0.259
	Distance + Correlation	**0.809**	<u>0.217</u>	0.357	<u>0.270</u>
	Distance + Random Edges	0.773	0.193	0.405	0.261
	Distance + Distogram	<u>0.782</u>	**0.236**	**0.535**	**0.328**
R-GAT	Distance	0.753	0.196	0.478	0.278
	Distance + Correlation	0.791	0.229	0.469	0.308
	Distance + Random Edges	0.760	0.217	0.546	0.311
	Distance + Distogram	<u>0.814</u>	<u>0.289</u>	<u>0.598</u>	<u>0.390</u>
	Distance + Distogram + Features	**0.816**	**0.298**	**0.627**	**0.404**
R-EGNN	Distance	0.832	0.282	0.444	0.345
	Distance + Correlation	**0.882**	**0.393**	0.350	0.370
	Distance + Random Edges	0.830	0.316	**0.607**	0.416
	Distance + Distogram	<u>0.861</u>	<u>0.376</u>	<u>0.606</u>	**0.464**
	Distance + Distogram + Features	0.859	0.371	0.602	<u>0.459</u>
Binding Affinity Prediction (Mean number of nodes = 47)					
Model	Graph Type	MAE (↓)	RMSE (↓)	Pearson R (↑)	Spearman R (↑)
R-GCN	Distance	1.244	1.562	0.689	0.656
	Distance + Correlation	1.299	1.601	0.673	0.637
	Distance + Random Edges	**1.171**	**1.453**	**0.744**	**0.718**
	Distance + Distogram	<u>1.195</u>	<u>1.487</u>	<u>0.731</u>	<u>0.694</u>
R-GAT	Distance	1.226	1.562	0.691	0.658
	Distance + Correlation	1.280	1.568	0.686	0.656
	Distance + Random Edges	1.236	1.542	0.701	0.685
	Distance + Distogram	**1.176**	**1.487**	**0.740**	**0.713**
	Distance + Distogram + Features	<u>1.179</u>	<u>1.492</u>	<u>0.722</u>	<u>0.693</u>
R-EGNN	Distance	1.296	1.623	0.666	0.642
	Distance + Correlation	1.357	1.713	0.611	0.576
	Distance + Random Edges	<u>1.211</u>	<u>1.502</u>	<u>0.721</u>	<u>0.698</u>
	Distance + Distogram	1.275	1.560	0.699	0.674
	Distance + Distogram + Features	**1.208**	**1.479**	**0.736**	**0.725**

Table 6: Results on the binding site prediction task (Top, average of 443 nodes) and binding affinity prediction task (Bottom, 47 nodes on average). We compare various dynamic-encoding approaches using $\mathcal{E}_{dist}$ alone or combined with $\mathcal{E}_{corr}$, $\mathcal{E}_{random}$, or $\mathcal{E}_{disto}$. Distogram-based edge features are also incorporated where compatible. Best-performing models are shown in bold, and second-best are <u>underlined</u>.

Binding Site Prediction (Mean number of nodes = 443)					
Model	Graph Type	Accuracy (↑)	Precision (↑)	Recall (↑)	F1 score (↑)
R-GCN	Distance + Distogram	0.782	0.236	**0.535**	**0.328**
	Distance + Distogram + correlation	**0.803**	**0.246**	0.474	0.324
R-GAT	Distance + Distogram	0.814	0.289	**0.598**	0.390
	Distance + Distogram + correlation	**0.819**	**0.293**	0.584	**0.391**
R-EGNN	Distance + Distogram	**0.861**	**0.376**	0.606	**0.464**
	Distance + Distogram + correlation	0.843	0.339	**0.610**	0.436
Binding Affinity Prediction (Mean number of nodes = 47)					
Model	Graph Type	MAE (↓)	RMSE (↓)	Pearson R (↑)	Spearman R (↑)
R-GCN	Distance + Distogram	**1.195**	**1.487**	**0.731**	**0.694**
	Distance + Distogram + correlation	1.303	1.601	0.676	0.655
R-GAT	Distance + Distogram	**1.176**	**1.487**	**0.740**	**0.713**
	Distance + Distogram + correlation	1.319	1.630	0.661	0.630
R-EGNN	Distance + Distogram	**1.275**	**1.560**	**0.699**	**0.674**
	Distance + Distogram + correlation	1.359	1.672	0.632	0.614

Table 7: Effect of adding correlation edges to distance and distogram edges

Table 7 indicates that incorporating correlation edges in addition to distance and distogram edges yields little to no improvement in either binding site or binding affinity prediction, suggesting that the correlation information is largely captured by the existing distance and distogram relationships. This effect could be explained by the overlap between distograms and correlation edges, since both capture related aspects of the system's dynamic behavior.