Learning Biophysical Dynamics with Protein Language Models

Chao Hou[1], Haiqing Zhao[4,5], Yufeng Shen[1,2,3,*]

1 Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032

2 Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032

3 JP Sulzberger Columbia Genome Center, Columbia University, New York, NY 10032

4 Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555

5 Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX 77555

* Corresponding author: ys2411@cumc.columbia.edu

Abstract

Structural dynamics are fundamental to protein functions and mutation effects. Current protein deep learning models are predominantly trained on sequence and/or static structure data, which often fail to capture the dynamic nature of proteins. To address this, we introduce SeqDance and ESMDance, two protein language models trained on dynamic biophysical properties derived from molecular dynamics simulations and normal mode analyses of over 64,000 proteins. SeqDance, trained from scratch, learns both local dynamic interactions and global conformational properties for ordered and disordered proteins. SeqDance predicted dynamic property changes reflect mutation effect on protein folding stability. ESMDance, built upon ESM2 outputs, substantially outperforms ESM2 in zero-shot prediction of mutation effects for designed and viral proteins which lack evolutionary information. Together, SeqDance and ESMDance offer a new framework for integrating protein dynamics into language models, enabling more generalizable predictions of protein behavior and mutation effects.

Keywords: protein language model, molecular dynamics, normal mode analysis, mutation effects.

Introduction

Protein deep learning models have made significant progresses in predicting 3D structure (e.g., AlphaFold[1, 2], RoseTTAfold[3], and ESMFold[4]), protein function[5], stability[6], localization[7], and interactions[8]. A central challenge in developing effective models lies in the selection of information sources that models can interpret and learn from. Broadly, these information can be classified into two main types: evolutionary and biophysical[9].

Evolutionary information is derived from protein homologous sequences across species that have been shaped by similar selection pressures. Multiple sequence alignments (MSAs) of homologs contain information of conserved motifs and co-evolutionary pairs. Recently, protein language models (pLMs) like ESM1,2[4, 10], ProtTrans[11], and ProGen[12], trained on large-scale protein sequence datasets, have emerged as powerful tools to implicitly learn evolutionary information[13-17]. MSAs and pLMs have been successfully applied to predict 3D structures[1, 2, 4], predict pathogenic mutations[18, 19], predict signal peptides[6], and even generate novel proteins[12]. However, the effectiveness of evolutionary information relies on the quality and quantity of available homologous sequences. For instance, the confidences of structure predictions correlate to the number of homologous sequences[1, 4]. As a result, evolutionary information is less effective for intrinsically disordered regions (IDRs), designed proteins, rapidly evolving viral proteins, immune proteins, and those from under-studied species like extremophiles[13], where homologous sequences are either sparse or highly divergent. Moreover, evolutionary patterns in homologs are the outcomes of protein functions rather than the underlying causes (Figure 1A). Overreliance on evolutionary information may bias models toward correlational patterns rather than learning fundamental principles governing protein behaviors.

Biophysical information, primarily derived from protein structures, circumvents the limitations of evolutionary information. Predicting protein behaviors from biophysical information aligns with the central paradigm that sequence determines structure, and structure governs function (Figure 1A), thereby enabling models to generalize more broadly across the protein universe. Biophysical information has been applied to a wide range of biological questions. For example, predicted structures have been used to predict mutation effects and functional sites[20]. MaSIF[8] utilized biophysical and geometric features on protein surface to study protein interactions. PScore[21] was

trained to predict Pi-interactions derived from PDB[22] structures and can be used to identify phase-separating proteins[23]. Additionally, there have been initiatives to integrate biophysical information into pLMs. ProSE[9] was trained to predict masked residues, contacts within static structures, and structural similarities, while METL[24] was developed to predict biophysical properties derived from static structures.

However, current protein models use biophysical information from static structures. These structural snapshots lack crucial information about thermodynamics, topological landscape of folding, allostery, and dynamic interactions. Moreover, static structures cannot describe the dynamic nature of IDRs, which constitute more than 30% of the human proteome[25] and use their inherent flexibility to mediate essential biological processes[26, 27]. Protein dynamics are fundamental to understand protein behaviors and mutation effects. Although evolutionary information encodes aspects of protein dynamics[28, 29], it depends on the quality and quantity of homologous sequences and cannot be directly extracted. Molecular dynamics (MD) simulations are widely used to study protein dynamics, generating ensembles of structures based on defined physical forces over the simulation time. All-atom MD simulations are often computationally intensive to reach a converged equilibrium state[30]. Coarse-grained models simplify protein residues into beads and use specialized force fields to reduce computation cost[25]. Additionally, normal mode analysis[31, 32] (NMA) can describe protein vibrations around the equilibrium conformation, with low-frequency modes capturing large-scale functionally relevant motions. While these methods are wildly used to study protein dynamics, the produced data are often high-dimensional and irregularly shaped. Thus, effectively and efficiently integrating these dynamic data into protein deep learning models remains a significant challenge.

In this study, we present a deep learning framework to integrate dynamics data into protein language models. Using dynamics profiles from 64,403 proteins, we developed two models: SeqDance and ESMDance. SeqDance is trained directly on dynamics data without incorporating prior structural or evolutionary information, while ESMDance is built upon ESM2 outputs. We demonstrated that these models capture local dynamic interactions and global conformational properties in unseen proteins. We also applied these models for zero-shot prediction of mutation effects, achieving superior performance on designed and viral proteins.

Results

Pre-training Protein Language Models on Dynamics Properties of over 64,000 Proteins.

To establish the protein dynamics dataset, we curated both high-resolution and low-resolution data sources. High-resolution dynamics data includes experimental data and all-atom molecular dynamics (MD) simulation trajectories from mdCATH[33], ATLAS[34], GPCRmd[35], PED[36]. Since high-resolution data is limited, we further used low-resolution dynamics data, including coarse-grained MD trajectories of human intrinsically disordered regions (IDRs) from IDRome[25] and Normal mode analysis (NMA)[31, 32, 37] for representative structures in the PDB. Overall, we collected 7,799 high-resolution data and 56,604 low-resolution data, covering ordered domains, IDRs, membrane proteins, antibodies, and protein complexes[38] (see Table 1 and Methods for details).

Table 1: Protein dynamic datasets.

	Source	Description	Count	Method
High-resolution	PED[36]	Disordered regions	382	Experiment and others
	GPCRmd[35]	Membrane proteins	509	All-atom MD, ~500 ns × 3
	mdCATH[33]	CATH domains	5,392	All-atom MD, ~464 ns × 5
	ATLAS[34]	Ordered structures in PDB	1,516	All-atom MD, 100 ns × 3
Low-resolution	IDRome[25]	Disordered regions	28,058	Coarse-grained MD
	Proteinflow[38]	Ordered structures in PDB	28,546	Normal mode analysis

We extracted residue-level and pairwise dynamic properties that describe the distribution of properties in structure ensembles (Figure 1C, Table S1). Residue-level properties include normalized root mean square fluctuation (RMSF), solvent accessible surface area (SASA), eight-class secondary structures, and dihedral angles (*phi*, *psi*,

chi1) which describe the rotations around bonds in the protein backbone and side chains. Pairwise properties include the correlation of Cα movements and frequencies of hydrogen bonds, salt bridges, Pi-cation, Pi-stacking, T-stacking, hydrophobic interactions, and van der Waals interactions. For NMA data, we categorized normal modes of each structure into three frequency-based clusters. For each cluster, we calculated residue fluctuation and pairwise movement correlation (see Table S1 and Methods for details).

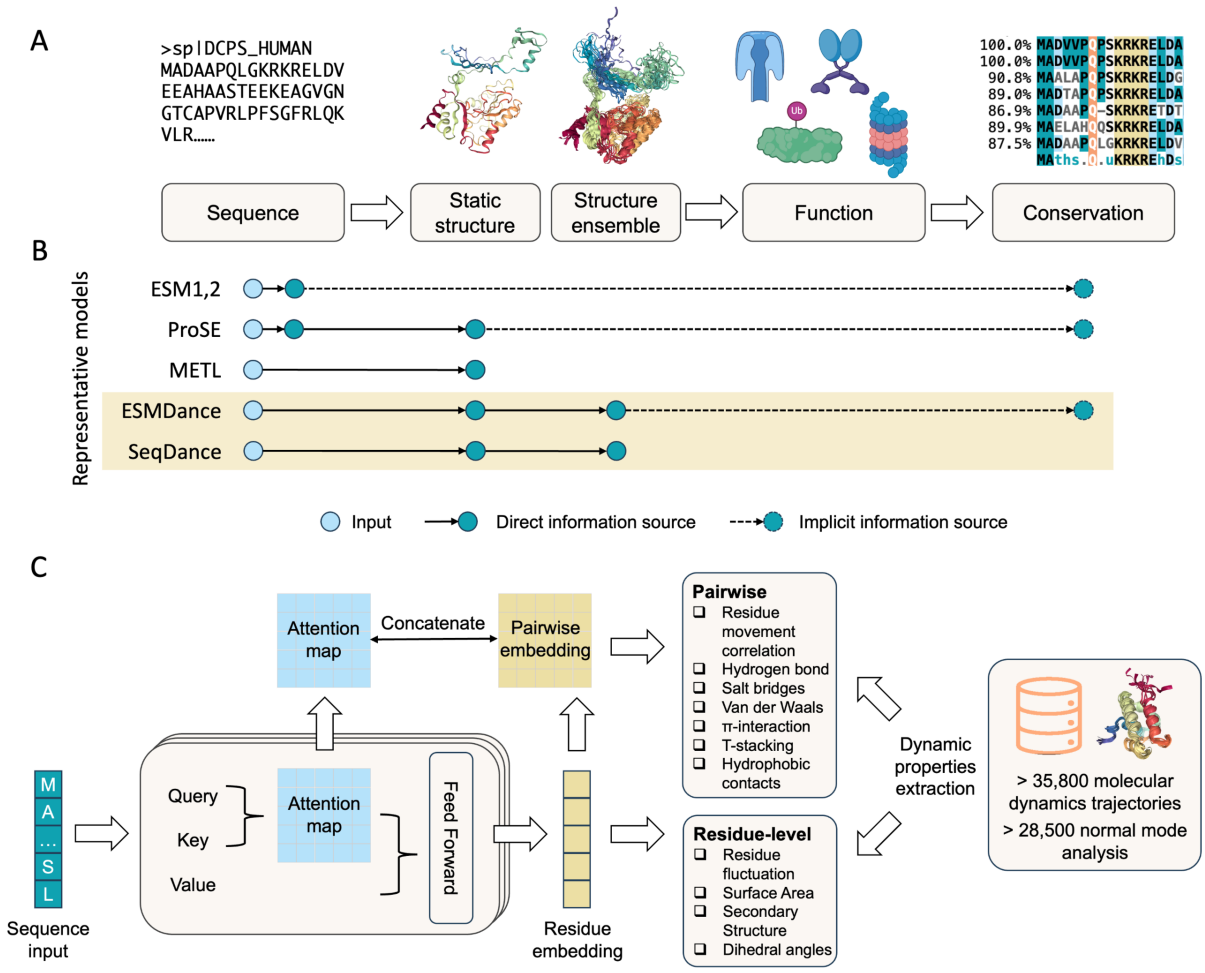

Figure 1: Information flow in protein study, representative protein language models, and model pre-training.

A. Illustration of the "sequence - structure ensemble - function - evolution" paradigm. Sequences are the basic elements of proteins that fold into structural ensembles to perform functions. Functionally important regions exhibit conserved patterns across homologs. **B**. Representative protein language models (pLMs) and their information sources. ESM1 and 2 were trained to predict masked residues and implicitly learned evolutionary information. ProSE was trained to predict masked residues, pairwise contact in static structures, and structure similarity. METL was trained to predict biophysical terms calculated from static structures. SeqDance and ESMDance were trained on protein dynamic properties from molecular dynamics (MD) simulations, experimental data, and normal mode analysis (NMA) of static structures. **C**. Diagram of the pre-training process. SeqDance and ESMDance take a protein sequence as input and predict residue-level and pairwise dynamic properties, which are extracted from structures ensembles and NMA. Both models use a Transformer encoder architecture identical to ESM2-35M, consisting of 12 layers with 20 heads per layer and an embedding dimension of 480. Linear layers are applied to the residue embeddings to predict residue-level properties. For pairwise property prediction, pairwise embeddings—constructed from residue embeddings—are concatenated with attention maps, a linear layer is then applied to predict pairwise properties.

We developed two models, SeqDance and ESMDance, both consisting of Transformer encoders and dynamic property prediction heads. Both models take protein sequences as input, use the encoder's final layer embeddings and attention maps to predict residue-level and pairwise dynamic properties (Figure 1C). The Transformer encoder follows the same architecture as ESM2-35M with 12 layers and 20 heads per layer, the dynamic property prediction heads contain 1.2 million trainable parameters in total. The key difference between the two models lies in their initialization and use of prior knowledge. In SeqDance, all parameters were randomly initialized, allowing the model to learn dynamic properties entirely from scratch. In contrast, ESMDance retains and freezes all

parameters from ESM2-35M, enabling it to leverage the evolutionary information already encoded in ESM2 to predict dynamic properties (see Methods for details).

To balance contributions of different properties from different data sources, we adjusted their respective weights in the loss function (see Methods for details). We clustered sequences using a 50% sequence identity cutoff (same to ESM2 training)[39], then randomly chose 95% of the clusters as the training set, reserving the remaining 5% as the test set. The training and test losses are shown in Supplementary Figure 1. SeqDance and ESMDance achieve comparable overall test losses, with SeqDance performing better on the IDRome test set and NMA test set, while ESMDance performing better on the high-resolution test set. This result is consistent with the fact that evolutionary information is less effective for IDRs. After pre-training, we observed strong correlations between the weights of linear layers used to predict co-movement in MD simulations and those used for co-movement in NMA, in both models (Figure S2). This result conforms that NMA effectively captures dynamic motions similar to those observed in MD simulations, supporting its use as a complementary source for augmenting model pre-training.

In the following, we first assessed how much dynamic information can SeqDance learn from scratch without relying on evolutionary information, we analyzed its attention mechanism, embeddings, and predicted dynamic properties. In the last section, we examined the advantage of ESMDance compared to ESM2 and SeqDance.

SeqDance's Attentions Capture Dynamic Interactions and Co-movement.

The Transformer model employs self-attention mechanism[40] to aggregate information from other tokens, with attention values representing the relationship between tokens (in protein context, amino acids). Here, we evaluated whether SeqDance's attention mechanism captures dynamic interactions and residue co-movement. The analysis was conducted on the test dataset, which included 405 high-resolution proteins, 1,360 IDRome proteins, and 1,362 proteins with NMA. Additionally, we evaluated an independent test set comprising 29 Dynamic PDB proteins. All test proteins share less than 50% sequence identity with the SeqDance training sequences, allowing us to assess the model's ability to generalize to unseen proteins. The Dynamic PDB test set was generated using a different MD force field to those used in pre-training, enabling us to further evaluate SeqDance's generalization ability across different simulation conditions. For all analyses, only residue pairs farther than five amino acids were considered as recommended in CASP[41].

We compared the dynamic properties (interaction frequency and movement correlation) of residue pairs with attention values from each attention head. As these properties are continuous rather than binary, we calculated Spearman correlations for all 240 attention heads and highlighted the top five with the highest median correlations in Figure 2. For interaction frequency in the MD ensemble, the best attention heads of SeqDance achieve median Spearman correlations of 0.41 on the high-resolution test set, 0.52 on the IDRome test set, and 0.52 on the Dynamic PDB test set (Figure 2A-C; the best attention head for different data may be different). For residue pairs with positive movement correlation in the MD ensemble, SeqDance performs particularly well on the IDRome test set, achieving a median Spearman correlation of 0.74 for the best attention head. In contrast, median Spearman correlations of best attention heads are just 0.22 for both ordered test set (high-resolution and Dynamic PDB; Figure 2D-F). This may be because the co-movement patterns of IDRs often involve nearby residues and are therefore easier to capture. SeqDance's attention do not capture residue pairs with negative movement correlation in the MD ensemble well (Figure S3A-C). For NMA data, the best attention heads achieve median Spearman correlations of 0.35, 0.29, and 0.23 for positively correlated pairs in low-, medium-, and high-frequency modes, respectively (Figure 2G-I), and Spearman correlations of –0.04, –0.26, and –0.35 for negatively correlated pairs in three frequency modes (Figure S3D-F; negative correlation means better capturing this property).

We further evaluated ESM2-35M, as ESMDance was trained with ESM's parameters frozen and thus shares identical attentions. We note that the test proteins are dissimilar to SeqDance's training data, but appear in, or have homologs in, the ESM2 training data. Previous studies showed that ESM's attention heads capture both native interactions observed in PDB structures[42] and non-native interactions revealed in MD simulations[28] via learning evolutionary information in the training set. Across all evaluations, ESM2-35M and SeqDance show comparable performances (Figure 2; Figure S3). SeqDance performs slightly better on four evaluations (interaction frequencies, negatively correlated pairs in the MD ensemble, and both positively and negatively correlated pairs in NMA), while ESM2-35M performs slightly better on one evaluation (positively correlated pairs in the MD ensemble). Overall, these results demonstrated that our pre-training strategy enables SeqDance's attention to capture dynamic interactions and residue co-movement directly from sequence, performing comparable to the state-of-the-art ESM model on unseen proteins and different simulation conditions. ESM2's

ability to capture dynamic properties enabled us to train ESMDance, achieving comparable test loss to SeqDance with far fewer trainable parameters (Figure S1).

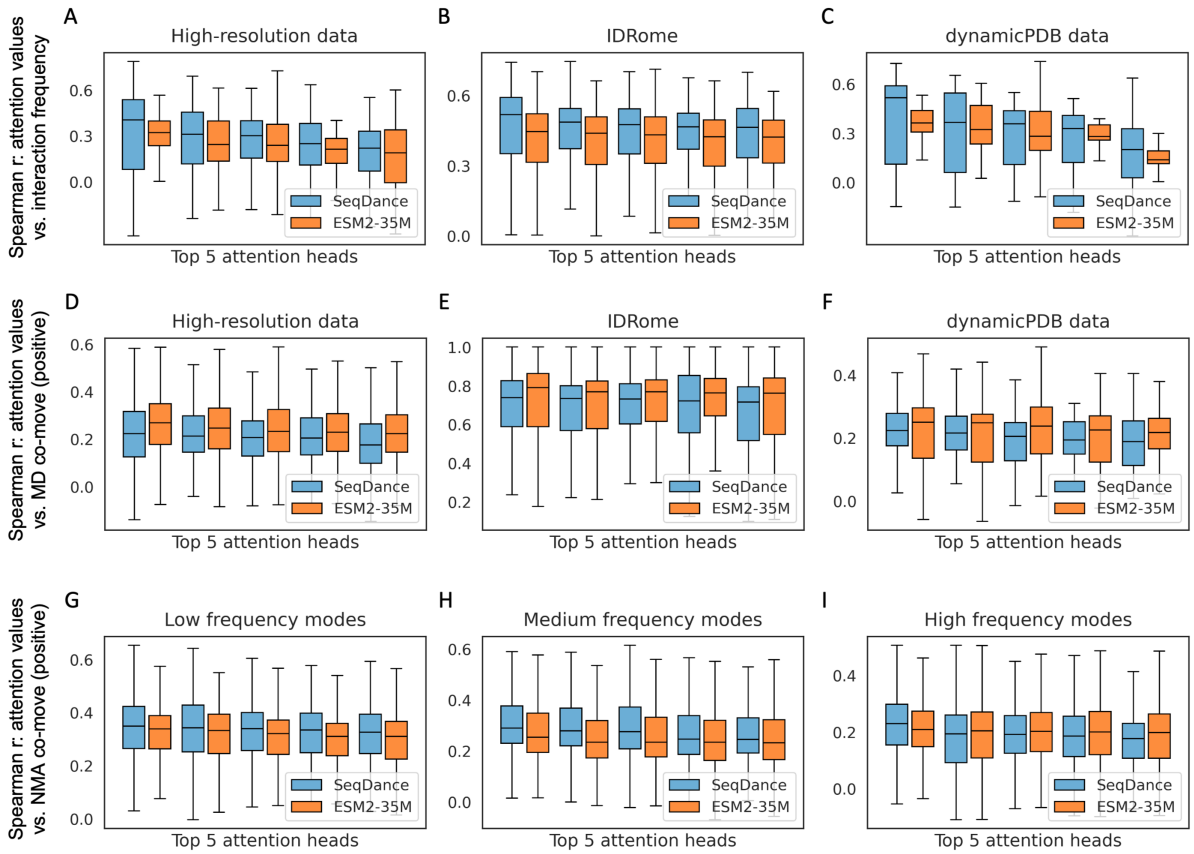

Figure 2. SeqDance's attention mechanism captures dynamic interactions in test sets.

For each attention head, Spearman correlation was calculated between attention values and pairwise dynamic properties, top five heads of SeqDance and ESM2-35M are shown. **D-I** show results for positively correlated residue pairs. Boxplots show the distribution of Spearman correlations of test data, The box extends from the first quartile to the third quartile of the data, with a line at the median. The whiskers extend from the box to the farthest data point lying within 1.5 times the inter-quartile range from the box.

SeqDance's Embeddings Encode Global Protein Conformation Properties.

Next, we investigated whether SeqDance learns protein conformational properties not included in the pre-training tasks—an important aspect for assessing the model's generalization capability. Specifically, we considered end-to-end distance, asphericity, and radius of gyration (R_g), which quantify motion range, deviation from a spherical shape, and compactness, respectively. Since these properties cannot be directly obtained from the model, we trained linear regression models using its embeddings in a supervised setting. To ensure a fair comparison with other pLMs in Figure 1B (ESMDance embedding is identical to ESM2-35M), we applied principal component analysis to the mean-pooled embeddings from each method and used the top 200 components as input features.

We first evaluated these pLMs on structure ensembles of IDRs[43] generated using a different coarse-grained force field compared to SeqDance training set (IDRome). After removing IDRs with over 20% sequence identity and 50% coverage to any SeqDance training sequence, 19,810 IDRs remained (see Methods for data filtering). We used mean end-to-end distance, asphericity, and R_g in the ensemble and normalized these values to account for protein length. The training and test sets were split in a 6:4 ratio using a 20% sequence identity cutoff to prevent information leakage. Linear regression models were trained for each pLM, and the experiment was repeated 20 times. The distributions of test loss (mean squared error, Figure 3A-C) show that SeqDance significantly outperforms the randomly initialized model prior to pre-training and other pLMs. SeqDance reduces the test losses of ESM2-650M, METL, and ProSE by 35%, 44%, and 57% on average, respectively.

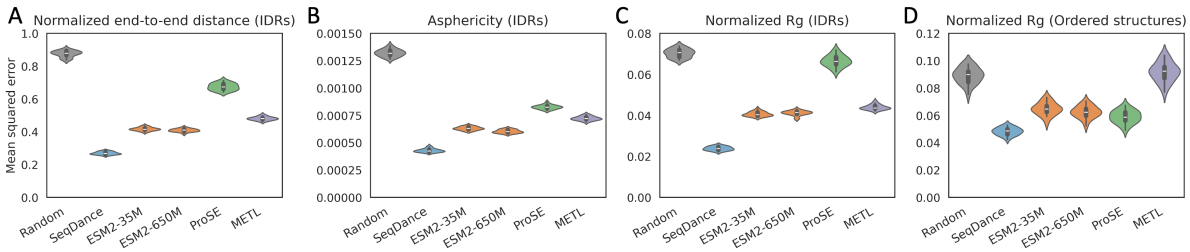

Figure 3. SeqDance's embeddings encode global protein conformational properties.

Performance comparison of SeqDance (35M parameters), METL, ProSE, and ESM2 (ESMDance embedding is identical to ESM2-35M) in predicting the normalized end-to-end distance of disordered proteins (**A**), asphericity of disordered proteins (**B**), normalized radius of gyration (R_g) of disordered proteins (**C**) and ordered proteins (**D**). The training and test split was 6:4 with a 20% sequence identity cutoff. Linear regression model was trained to predict conformational properties using the first 200 principal components of mean-pooled embeddings from each method. Results presented are the distributions of the test loss (mean squared error) of 20 independent repeats.

For ordered proteins, obtaining conformational properties from structure ensembles is challenging. Therefore, we used normalized R_g values (see Methods for details) of over 11,000 static monomer structures in the PDB from the paper[44]. Since SeqDance was trained on NMA of nearly all representative PDB structures[38], we did not exclude sequences with homologs in SeqDance training set. Using the same evaluation method described above, SeqDance reduces the test losses of the random model, ESM2-650M, ProSE, and METL by 46%, 23%, 18%, and 47%, respectively (Figure 3D).

We evaluated the performance using Pearson correlation, where SeqDance achieves the highest correlations across all tasks as well (Figure S4). Notably, even using the full embeddings instead of the first 200 principal components, SeqDance still outperforms other pLMs in all tasks (Table S2), despite having the shortest embedding size. Overall, these results demonstrate that SeqDance learns both local dynamic interactions and global conformation properties for both ordered proteins and IDRs.

SeqDance Predicted Dynamic Property Changes Reflect Mutation Effect on Protein Stability.

To evaluate whether SeqDance captures information about biophysics of protein folding, we compared its output with mutation effects on protein folding stability. We hypothesized that destabilizing mutations would induce significant changes in protein dynamics, whereas mutations with minimal impact would induce smaller dynamic changes. To test this hypothesis, we used SeqDance to predict dynamic properties for both wild-type and mutated sequences, computed their relative changes, and compared these to experimentally measured ΔΔG values (Figure 4A, see Methods for details). This approach was framed as a zero-shot prediction as the model was not specifically trained on mutation data. We applied this approach to the mega-scale protein folding stability dataset[45], which includes ΔΔG values for 379,865 single mutations and 147,926 double mutations across 412 proteins, encompassing both natural and designed proteins.

To evaluate the zero-shot predictive power of individual dynamic properties, we computed the Spearman correlation between the relative change of each property and the corresponding ΔΔG values for mutations in each protein. While the randomly initialized model prior to pre-training shows Spearman correlations around 0, many SeqDance-predicted properties achieve median Spearman correlations above 0.2 (Figure 4B, S5), confirming that pre-training enables SeqDance to capture meaningful biophysical dynamics. Among individual properties, solvent-accessible surface area (SASA mean) emerges as one of the top-performing properties, consistent with the fact that destabilizing mutations disrupt core packing and increase solvent exposure. Dihedral angles *phi* and *psi* also rank highly, highlighting the role of backbone conformational flexibility in protein stability. Additionally, pairwise movement correlations derived from MD simulations and NMA demonstrate relatively high performance, further emphasizing the importance of collective dynamics in predicting mutation effects. To integrate these dynamic properties, we used the geometric mean of quantile-normalized relative changes (see Table S3 and Methods for details). The unified score yields a median Spearman correlation of 0.24 across all 412 proteins.

To assess whether SeqDance's performance depends on the presence of homologs in its training set, we first analyzed its performance relative to the most similar protein in the training set. SeqDance performs better on

proteins with close homologs of over 95% sequence identity in its training set (Figure 4C). For the 125 proteins with such close homologs, the median Spearman correlation is 0.29, whereas for the 191 proteins with no homologs in the training set, the median correlation is 0.20. Proteins with homologs of 20–95% sequence identity show similar performance to those with no homologs. We also examined SeqDance's performance as a function of the number of homologs (over 20% sequence identity and 50% coverage) and observed only a minor correlation (Figure 4D).

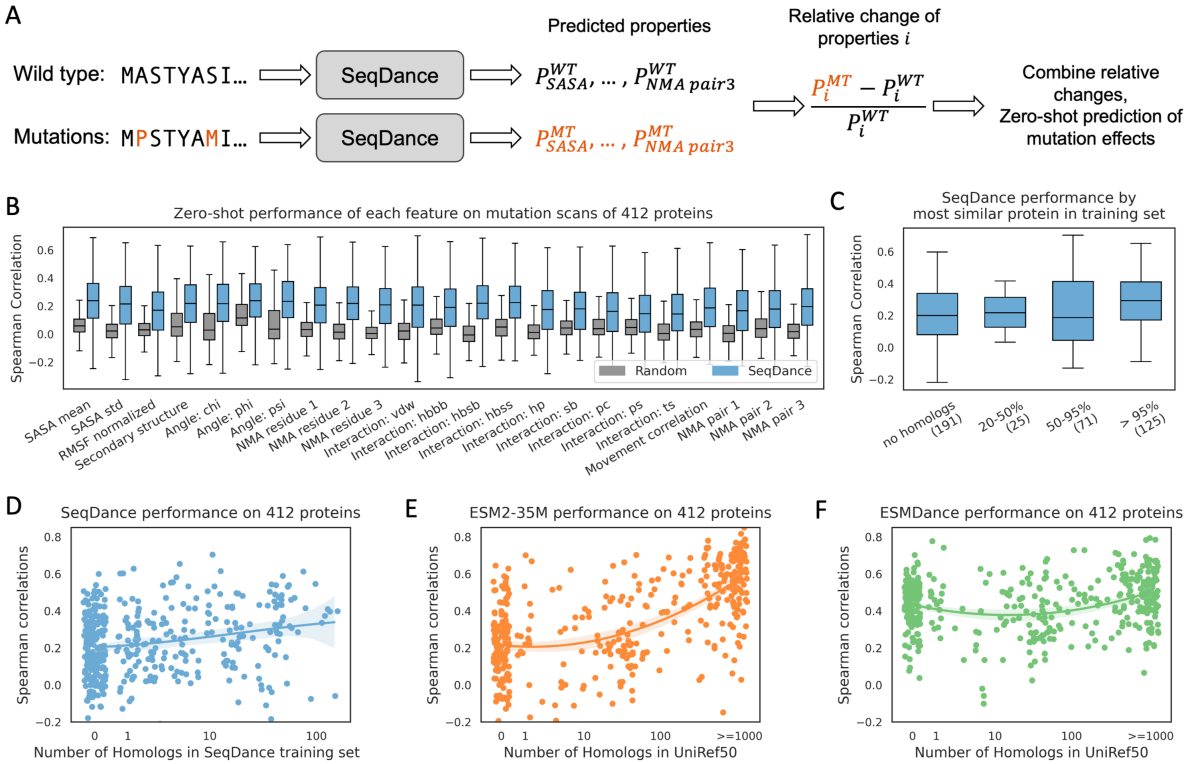

Figure 4. Zero-shot prediction of mutation effect on protein folding stability.

A. Framework of using SeqDance (35M parameters) or ESMDance (35M parameters) to perform zero-shot prediction of mutation effects. **B**. Distribution of zero-shot performance (Spearman correlation) on 412 proteins of different dynamic properties; the random model (color grey) represents the randomly initialized model prior to pre-training. SASA mean, std: mean and standard deviation of solvent-accessible surface area. NMA properties 1, 2, 3: properties calculated from low-, median-, and high-frequency normal modes. vdw: van der Waals interactions; hbbb: backbone-to-backbone hydrogen bonds; hbsb: side-chain-to-backbone hydrogen bonds; hbss: side-chain-to-side-chain hydrogen bonds; hp: hydrophobic interactions; sb: salt bridges; pc: Pi-cation interactions; ps: Pi-stacking interactions; ts: T-stacking interactions. **C**. SeqDance's performance split by most similar proteins in SeqDance training set. We used cutoff of 95% sequence identity together with 95% coverage, 50% sequence identity together with 80% coverage, 20% sequence identity together with 50% coverage. **D-F**. Relationship between zero-shot performance and the number of homologs (20% sequence identity and 50% coverage) in training set. The line represents a polynomial regression of order two, with the shaded area indicating the 95% confidence interval. The x-axis shows the log-scaled number of homologs, with small random noises added to the x values to reduce overlap.

Furthermore, we compared SeqDance with ESM2, which can perform zero-shot prediction of mutation effects by estimating the probability of 20 amino acids based on the sequence context[19], leveraging evolutionary information learned during pre-training. ESM2-35M achieves a median Spearman correlation of 0.33, outperforming SeqDance. This is consistent with the fact that evolutionary information is highly effective in predicting mutation effects. However, the performance of ESM2 is strongly dependent on the number of homologs in its training set, UniRef50, particularly for proteins with over 100 homologs (Figure 4E, ESM2-650M and -15B in Figure S6A, B). To isolate the effect of homolog abundance, we evaluated a subset of 257 proteins that have a similar number of homologs (between 0 and 146) in both UniRef50 and the SeqDance training set. In this subset, ESM2-35M achieves a median Spearman correlation of 0.22, comparable to SeqDance's performance of 0.24 (Figure S6C). These results suggest that, given similar model size and training data abundance, SeqDance and ESM2 perform similarly in predicting mutation effects on protein stability. Overall, despite being pre-trained exclusively on wild-

type sequences without reliance on evolutionary information, SeqDance is able to make meaningful predictions about mutation effects—even for proteins with no homologs in the training set.

ESMDance Enhances Zero-Shot Prediction of Mutation Effects in Designed and Viral Proteins.

While both ESM2-35M, which leverages evolutionary information, and SeqDance, which captures protein dynamic properties, can predict mutation effects, their individual performance remains modest at a model size of 35 million parameters. To harness the strengths of both approaches, we developed ESMDance that integrates evolutionary and dynamic information, also with 35 million parameters. Using the zero-shot approach illustrated in Figure 4A, ESMDance achieves a median Spearman correlation of 0.46 across 412 proteins (Figure 4F), significantly outperforming both ESM2-35M (0.33) and SeqDance (0.24). Remarkably, ESMDance also matches or surpasses much larger models, including ESM2-650M (0.46; Figure S6A) and ESM2-15B (0.43; Figure S6B). Moreover, ESMDance demonstrates robust generalization: its performance is independent of the presence of homologs in its training set or in UniRef50 (Figure 4F, S6D, S6E; We note that proteins without homologs in UniRef50 also lack homologs in the ESMDance training set. Thus, the performance gains observed for these proteins are not attributable to the presence of their dynamic data in the ESMDance training.).

One limitation of ESM2 is that the evolutionary information it learns from natural proteins may fail to generalize to de novo designed proteins. To evaluate whether ESMDance can address this limitation, we analyzed a subset of 135 designed proteins, which lack homologs in UniRef50 or in the SeqDance/ESMDance training set. ESM2-35M, -650M, and -15B achieve median Spearman correlations of 0.21, 0.32, and 0.29, respectively—significantly lower than their average performances (Figure 5A, 4E, S6A, S6B). In contrast, ESMDance achieved a median Spearman correlation of 0.46, substantially outperforming SeqDance (0.20) and all ESM2 models (Figure 5A). Further analysis at the individual protein level revealed that the performances of different ESM2 models are highly correlated (Figure 5B). In contrast, both SeqDance and ESMDance exhibit orthogonal performance patterns to ESM2 (Figure 5C-D), with ESMDance showing significant improvements for most proteins compared to ESM2-35M (Figure 5D).

Furthermore, we visualized predictions for two designed proteins (highlighted in Figure 5B-D). The first, r10_572_TrROS_Hall (Figure 5E), was designed using TrRosetta hallucination[45]. Both SeqDance and ESMDance achieve significantly higher correlations (0.52 for both) compared to the ESM2 models (0.18, 0.25, and 0.20 for ESM2-35M, -650M, and -15B, respectively). The second, GG|run4_0284_0002 (Figure 5F), was designed using the EEHH method[45]. Here, both ESM2 and SeqDance perform poorly (0.02, 0.09, -0.02, and 0.18 for ESM2-35M, -650M, -15B, and SeqDance, respectively), whereas ESMDance achieves a markedly higher correlation of 0.53. These results suggest that dynamic properties provide additional information to evolutionary information in ESM2 and underscore the advantage of ESMDance in directly integrating dynamic properties and evolutionary information during model training, outperforming separate predictions from two models, each using only one type of information.

Besides designed proteins, we also examined their performances on viral proteins, which evolve rapidly and are underrepresented in UniProt. For deep mutational scanning experiments in ProteinGYM[19], ESM2 models consistently perform worse on viral proteins compared to the other proteins regardless of model size (Figure S7A). ESM2-35M achieves a median Spearman correlation around 0 on viral proteins (Figure S7A). For 23 viral proteins shorter than 1024 residues, encompassing 187,902 mutations, SeqDance and ESMDance achieve median Spearman correlations of 0.18 and 0.25, respectively, using the zero-shot approach in Figure 4A (which may not be the optimal strategy for mutational scanning beyond stability, see Discussion). They outperform ESM2-35M (0.03) but are not as good as ESM2-15B (0.36; Figure 5G). Among these viral proteins (Figure S7B), SeqDance and ESMDance perform best on the stability mutational scanning of porcine sapovirus viral protein genome-linked (VPg) protein, which acts as a protein primer for genome replication and as a protein cap for translation initiation[46]. SeqDance and ESMDance achieve Spearman correlations of 0.37 and 0.67, respectively (Figure 5H). Despite VPg having 42 homologs in UniRef50, all ESM2 models perform poorly, with the best Spearman correlation reaching 0.16. Moreover, ESMDance outperformed all methods evaluated in ProteinGYM on VPg, regardless of model size and input. Even though VPg is included in our training set (mdCATH: 2mxdA00, Figure 5H), ESMDance's performance is not sensitive to the presence of a protein in the training set (Figure S6E).

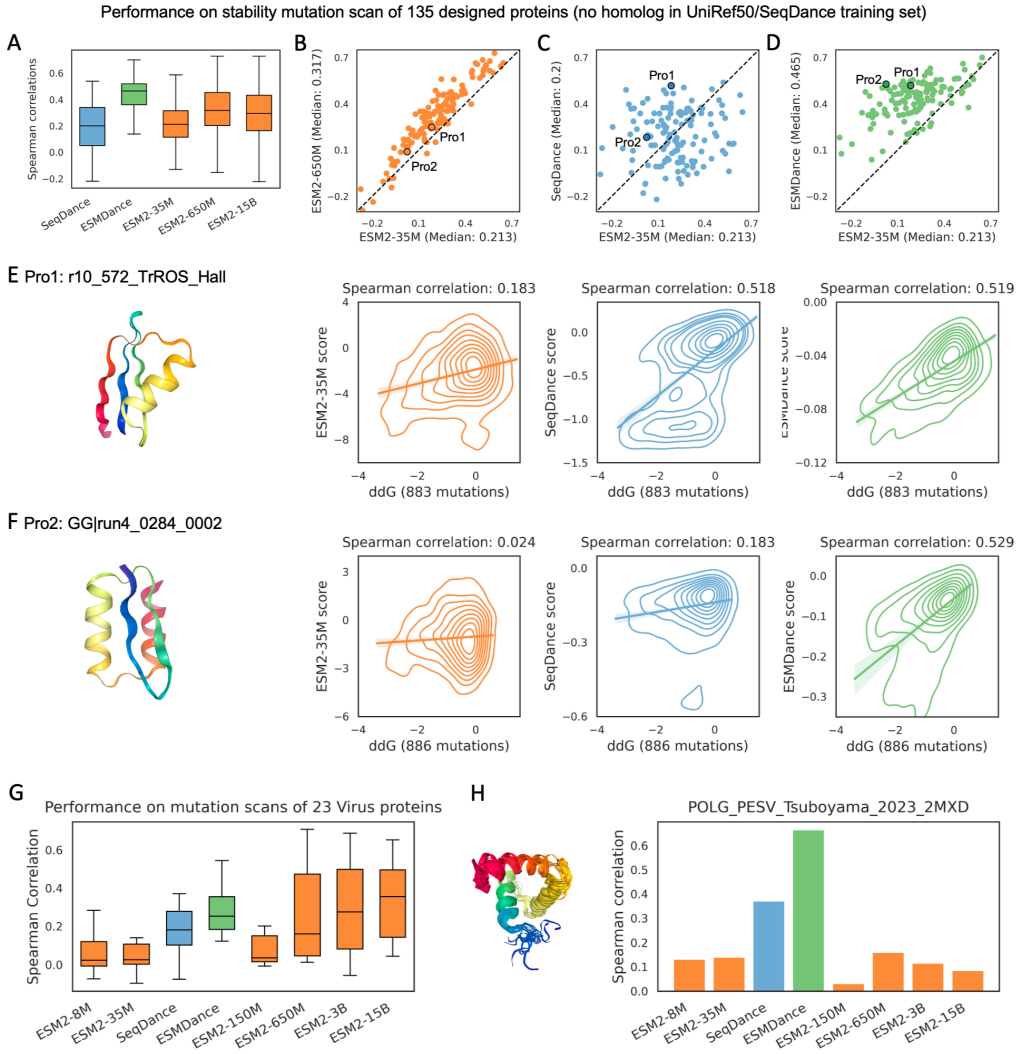

Figure 5. Zero-shot prediction of mutation effects for designed and viral proteins.

A-D. Performance comparison between ESM2, SeqDance (35M parameters), and ESMDance (35M parameters) for 135 designed proteins with no homologs in UniRef50 or the SeqDance training set. Two highlighted proteins are further analyzed in E-F. **E-F.** Structure and the relationship between predictions and ΔΔG values for two designed proteins. The kernel density estimate plot shows the distribution of experimentally measured folding ΔΔG values and zero-shot prediction values for the three methods. The line represents the linear regression, with the shaded area indicating the 95% confidence interval. **G.** Zero-shot performance of ESM2, SeqDance, and ESMDance on 23 viral proteins shorter than 1024 residues in ProteinGYM. **H.** Zero-shot Spearman correlations for the stability mutational scanning of porcine sapovirus VPg (named POLG_PESV_Tsuboyama_2023_2MXD in ProteinGYM). The structure shown here is based on ten random frames from mdCATH: 2mxdA00.

Discussion

In this work, we developed SeqDance and ESMDance, two protein language models (pLMs) pre-trained with dynamic properties derived from molecular dynamics (MD) simulations and normal mode analysis (NMA). SeqDance captures both local dynamic interactions, residue co-movement, and global conformational properties after pre-training, using only single sequence input and without relying on evolutionary information. ESMDance integrates the evolutionary information inherited from ESM2 and the newly learned dynamic properties, significantly outperforms SeqDance and ESM2-35M in zero-shot prediction of mutation effects for designed and viral proteins.

In the pre-training process, we did not directly predict full structure ensembles due to the immense size of the dataset (over 60 million frames) and the complexity of modeling entire ensembles. Instead, we used simplified dynamic property descriptors such as the mean, standard deviation, and interval distribution of ensemble-derived properties. Prior study has demonstrated that mean values from structure ensembles provide significantly more

information than those from static structures[47]. This strategy enables SeqDance and ESMDance to learn from simplified but informative properties of protein dynamics, without the overwhelming computational demand of full ensemble modeling. During pretraining, the test loss plateaued early (Figure S1), but we continued training to enable the model to learn dynamic information more effectively from the training set. As the training set was representative and relatively large, and extended training did not result in increased test loss. Using attention maps to predict pairwise properties constrains SeqDance to focus on interacting and co-moving residue pairs, reducing random or irrelevant attention. This approach is crucial for learning biophysically meaningful information and improves SeqDance's ability to extrapolate to unseen proteins. Our results show that our pre-training strategy similarly enables the transformer's attentions to effectively capture pairwise dynamic properties using only 1/1000th of the training sequences compared to ESM2.

SeqDance effectively learns global conformational properties of both ordered and disordered proteins, which are essential for understanding protein shape and flexibility. In comparison, METL[24] underperformed in predicting the radius of gyration (R_g) for ordered proteins (Figure 3D), despite having R_g prediction as a pre-training task. ProSE[9] performed well in predicting the conformational properties of ordered regions but struggled with IDRs (Figure 3), likely because its pre-training focused on ordered PDB structures. ESM2[4] performed well on both ordered and disordered proteins, as conformational properties are also conserved[43].

Current methods for zero-shot prediction of mutation effects use evolutionary information and static structure. These methods are less effective for proteins without sufficient homologs and structures (While some models can use predicted structures, the accuracy of structure prediction also depends on the availability of sufficient homologs.). We applied SeqDance and ESMDance to predict mutation effects by calculating changes of dynamic properties between wild-type and mutated sequences. These results highlight the effectiveness and efficiency of integrating both evolutionary information and dynamic properties for understanding mutation effects, particularly for proteins that cannot be accurately predicted using evolutionary information alone. While, to our knowledge, we are the first to apply this kind of zero-shot strategy to predict mutation effects on protein, analogous approaches like Enformer[48] and Borzoi[49] have been used to predict effect of non-coding mutations in DNA sequences, where evolutionary information is also less effective. Our current zero-shot strategy uses changes across the entire protein, which is reasonable for predicting mutation effects on protein stability but not optimal for mutation effects on function, binding, or activity, where dynamic changes at functional sites, binding interfaces, or catalytic centers may be more relevant. Additionally, our models were trained on protein dynamics without post-translational modifications, which are important for these effects. Moreover, for SeqDance, the presence of highly similar proteins in the training set leads to improved performance on mutation effect (Figure 4C). Thus, running MD simulations for a protein of interest and using the extracted features to fine-tune SeqDance may yield better predictions, but this may not work for ESMDance (Figure S6E).

We envision several directions to further improve the models. First, expanding protein dynamics data: while high-resolution data remain limited relative to the vast number of sequenced proteins, lower-resolution sources—such as coarse-grained MD simulations and NMA—can still provide valuable training signals. Second, incorporating more detailed dynamic properties that describe higher-order relationships and time-dependent behaviors can help the models to capture more nuanced aspects of protein motion. Third, scaling up model size: as demonstrated in many areas of deep learning, larger models can learn more complex patterns.

In conclusion, SeqDance and ESMDance advance the integration of protein dynamic properties—captured through molecular dynamics simulations and normal mode analysis—into deep learning frameworks. By complementing widely used evolutionary and static structure-based models, our models reveal novel insights into protein behavior and mutation effects, particularly for proteins lacking homologs, such as designed and viral proteins. This underscores the practical utility of dynamic properties as a critical, yet underexplored, dimension in computational protein design and studies, offering a powerful tool to bridge the gap between sequence, dynamics, and function.

Data availability

The extracted dynamic properties (approximately 100 GB in total), sequences (including train-test splits), loss weights for different properties, and all other data required to reproduce main results in this paper are available at: https://huggingface.co/datasets/ChaoHou/protein_dynamic_properties.

Code availability

The codes for feature extraction, model training and evaluations are available on GitHub: https://github.com/ShenLab/SeqDance. Model weights are available at: https://doi.org/10.5281/zenodo.15047777.

Author contributions

Y.S. and C.H. conceived the study. C.H. and H.Z. curated and processed the dataset, C.H. designed and implemented the methods, analyzed the data. C.H. drafted the manuscript. Y.S. and H.Z. critically revised the manuscript. All authors interpreted the results, reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgment

This work was supported by NIH grants (R35GM149527) and Simons Foundation SFARI #1019623. We thank Dr. Guojie Zhong and Aziz Zafar for helpful discussions and suggestions.

Methods

Protein Dynamic Data Collection and Processing

The Protein Ensemble Database[36] (PED) provides conformational ensembles for intrinsically disordered proteins, primarily derived from experiments, with some from molecular dynamics (MD) simulations or predictions. All available ensembles in PED were downloaded, and sequences shorter than 16 residues were excluded. GPCRmd[35] is a community-driven database of MD simulations of G-protein-coupled receptors (GPCRs), with most proteins simulated with the CHARMM force field for 500 ns in three replicates. mdCATH[33] provides MD simulations for 5,398 CATH domains ranging in length from 50 to 500 residues, using the CHARMM22 force field. Each domain was simulated at five temperatures (320 K, 348 K, 379 K, 413 K, and 450 K), with five replicas per temperature, averaging 464 ns per replica. Five replicas at 320 K were used in this study as this is close to human body temperature. The ATLAS[34] database contains MD simulations for 1,516 PDB structures, each conducted using the CHARMM36m force field for 100 ns with three replicates. The Dynamic PDB[50] dataset includes over 10,000 molecular dynamics (MD) simulations of PDB structures, generated using the Amber-ff14SB force field. IDRome[25] contains conformational ensembles of human disordered regions generated via the coarse-grained residue-level CALVADOS model. All coarse-grained trajectories were downloaded and converted to all-atom trajectories using cg2all (v1.5)[51]. While this conversion may result in structural artifacts and unrealistic conformations, we opted to use the converted all-atom trajectories as they provide more information, and deep learning models are generally robust to noise.

For IDRome, the first 10 frames were excluded as described in the original paper[25]; for mdCATH, no frame was discarded as the downloaded trajectories were recorded after the pre-equilibration phase; for the other trajectories, the first 20% of frames were discarded. All frames were aligned to the first frame based on Cα atoms using MDTraj[52], and trajectories of the same protein were merged. mdCATH data was obtained in February 2025. Dynamic PDB data was obtained in March 2025, as this dataset was not fully open and under uploading, we obtained 756 proteins with 100 ns simulations. Other data were obtained in January 2024.

Dynamic Properties Extraction from Structure Ensembles

GetContacts (https://getcontacts.github.io) was used to extract nine types of interactions from MD simulation trajectories: backbone-to-backbone hydrogen bonds, side-chain-to-backbone hydrogen bonds, side-chain-to-side-chain hydrogen bonds, salt bridges, Pi-cation, Pi-stacking, T-stacking, hydrophobic interactions, and van der Waals interactions. Default definitions of these interactions were used as described in https://getcontacts.github.io/interactions.html. For each residue pair, nine interaction frequencies were calculated, resulting in a matrix with the size of $L \times L \times 9$ for a protein of length L.

MDTraj (v1.9.9)[52] was used to extract residue-level properties. Surface area per residue was computed using mdtraj.shrake_rupley(mode='residue'), the mean and standard deviation of surface areas were calculated; Root mean square fluctuations (RMSF) were calculated using mdtraj.rmsf, as RMSF is related to simulation time, it was further normalized by dividing the max value in the protein; Eight-class secondary structures were determined using mdtraj.compute_dssp, the frequency of secondary structures was calcualted; For dihedral angles, mdtraj.compute_phi, mdtraj.compute_psi, and mdtraj.compute_chi1 were employed to extract the *phi*, *psi*, and *chi1* angles, respectively. Dihedral angles were partitioned into 12 bins (30° intervals), and the percentage of frames falling into each bin was calculated for each residue. For residues don't have a specific angle (like terminal residues don't have *phi* or *psi*, Glycine don't have *chi1*), 1/12 was used. Collectively, this yielded a residue-level property matrix of size $L\times(2+1+8+3\times12)$ for a protein of length L, the dimension corresponds to surface area (2), normalized RMSF (1), secondary structure (8), and three dihedral angles distributions (3×12), respectively.

For the calculation of pairwise residue movement correlations, we first computed the covariance matrix for the *x*, *y*, and *z* coordinates of all Cα atoms:

$$C_{3L} = \frac{1}{p-1}\sum_{i=0}^{p}(X_i - \bar{X})\,(X_i - \bar{X})^T \tag{1}$$

Where p is the number of frames in the ensemble, X_i represents the positions (x, y, z) of all Cα atoms in frame i, and $\bar{X}$ is the mean position of the Cα atoms of all frames. The matrix C_{3L} has a dimension of $3L\times3L$ where L is the number of Cα atoms (protein length).

To reduce this 3D covariance matrix to residue level, the trace over the spatial dimensions is taken:

$$C_L = Tr_{x,y,z}(C_{3L}) \tag{2}$$

Here, C_L is the reduced covariance matrix, and the diagonal of C_L corresponds to the squared fluctuation of L residues. The correlation matrix R_L is then computed by normalizing the covariance matrix:

$$\sigma_i = \sqrt{C_{L,ii}} \tag{3}$$

$$R_{L,ij} = \frac{C_{L,ij}}{\sigma_i\sigma_j} \tag{4}$$

where $C_{L,ij}$ is the covariance between residues i and j, σ_i and σ_j are their respective standard deviations. This correlation matrix describes the linear relationship between the displacements of residue pairs, independent of their absolute motion magnitude.

Normal Mode Analysis

For normal mode analysis (NMA), PDB structures in ProteinFlow[38] were used. Structures containing gaps (missing residues in the middle) or exceeding 5,000 residues were excluded. Terminal missing residues were removed. For the 20230102_stable dataset, MMseqs2[39] cluster representatives at 90% sequence identity were used. For the 20231221_sabdab dataset, MMseqs2 cluster representatives at 100% identity were used. '*X*' was added between chains of complexes for MMseqs2 clustering.

NMA was conducted using the Gaussian Network Model (GNM)[31] and the Anisotropic Network Model (ANM)[32], both implemented in ProDy (v2.4.0)[37]. These models represent macromolecules as elastic node-and-spring networks, where Cα atoms are nodes, and springs connect residues within a defined cutoff distance. A distance-dependent spring force constant was applied as recommended in ProDy website (http://www.bahargroup.org/prody/tutorials/enm_analysis/gamma.html): for Cα atoms 10-15 Å apart, a unit force constant was used; for atoms 4-10 Å apart, a force constant twice as strong was used; and for atoms within 4 Å (i.e., connected residue pairs), a force constant 10 times stronger was employed. GNM, which models isotropic motion, was used to compute residue-level properties, while ANM, which captures anisotropic motion, was employed to calculate pairwise properties.

After building the elastic network (Kirchhoff matrix for GNM or Hessian matrix for ANM), normal modes were computed by eigenvalue decomposition. Eigenvalues (λ_m) and eigenvectors (v_m) describe the collective motions of residues in mode m. The individual contribution of each mode is the proportion of the inverse eigenvalue to all modes:

$$\frac{1/\lambda_m}{\sum_{k=1}^{M} 1/\lambda_k} \tag{5}$$

Where M is the number of total modes (L-1 for GNM and $3L$-6 for ANM, L is protein length). Modes of GNM and ANM were first ranked by contribution, then partitioned into three ranges separately. The ranges were selected such that the first set of modes accounts for ~33% of the dynamics, the second set for ~33–66%, and the final set for ~66–100%. This ensures that the slow, intermediate, and fast modes are separated.

For each set of modes, the mean-square fluctuation (*MSF*) of each residue was calculated from GNM modes as:

$$MSF_i = \sum_m \frac{v_{mi}^2}{\lambda_m} \tag{6}$$

Where v_{mi} is the eigenvector component corresponding to mode m and residue i, and λ_m is the corresponding eigenvalue. This calculation was repeated for three mode ranges.

The residue covariance matrix was calculated using ANM modes. Firstly, the covariance matrix with the size of $3L \times 3L$ was first computed as:

$$C_{3L} = \mathbf{V}\Lambda^{-1}\mathbf{V}^T \tag{7}$$

Where $\mathbf{V}$ is the matrix of eigenvectors and Λ^{-1} is the inverse diagonal matrix of eigenvalues. To reduce the 3D covariance matrix to the residue-level, we used the same method as described in the calculation of pairwise residue movement correlations (equation 2-4). Readers can also read the paper[53] for the calculation of pairwise correlation in NMA.

Sequence Clustering

After addressing issues such as failed downloads, unreadable files, and processing errors (e.g., six trajectories from mdCATH had the error of atoms virtually overlapping reported by Mdtraj), we successfully extracted dynamic properties for 64,403 proteins (Table 1; excluding Dynamic PDB). To remove nearly duplicate proteins, we performed clustering using MMseqs2 with the parameters *--min-seq-id 1 -c 0.95 --cov-mode 0* ('*X*' was added between chains in complexes for clustering). We further selected sequences in each cluster with the following order: PED > GPCRmd > mdCATH > ATLAS > IDRome > ProteinFlow. This resulted in a final dataset of 61,959 proteins for model training and testing. We further clustered these sequences using MMseqs2 with the parameters *--min-seq-id 0.5, -c 0.8 --cov-mode 1*, yielding 43,193 clusters. We randomly selected 5% of the clusters as the test set, comprising 3,127 proteins, while the remaining 58,832 proteins formed the training set. The distribution of proteins across data sources is as follows: PED: 252 (train), 17 (test); GPCRmd: 107 (train), 5 (test); mdCATH: 5,067 (train), 312 (test); ATLAS: 1,361 (train), 71 (test); IDRome: 26,502 (train), 1,360 (test); ProteinFlow: 25,543 (train), 1,362 (test).

Dynamic PDB was used as an independent test set because its trajectories were generated using a different force field (Amber) compared to the other datasets (CHARMM and coarse-grained). Proteins in Dynamic PDB dataset were searched against SeqDance/ESMDance training set using MMseqs2 search with the parameters: *-s 7 -a 1*, only 29 proteins without homologs (over 50% identity and 80% coverage) in the training set were analyzed in Figure 2.

Model Architecture

The models consist of a transformer encoder and the dynamic property prediction heads. The transformer encoder uses the ESM2-35M architecture and the same sequence tokenizer. It comprises 12 layers with 20 attention heads each and employs rotational positional embeddings, with a final embedding dimension of 480. The choice of the 35M model is due to both the relatively small training dataset (compared to hundreds of millions of sequences) and our computational resources. Linear layers are added to predict residue-level properties from the residue embeddings of the last transformer layer. For pairwise feature prediction, pairwise embeddings are computed from residue embeddings as follows:

$$P_{ij} = concat(Q_i \odot Q_j, Q_i - Q_j) \tag{8}$$

$$Q = W_2\left(LayerNorm\left(GELU\left(W_1(E)\right)\right)\right) \quad (9)$$

where E represents the final residue embeddings, W_1 and W_2 are learnable linear matrices. i and j are residue indexes, $\odot$ represents element-wise multiplication, The pairwise embeddings P_{ij}, along with attention values from 240 heads, are passed through a linear layer to predict pairwise features. The dynamic property prediction heads contain 1.2 million parameters in total.

Using attention maps and pairwise embeddings allows the model to capture the distinct characteristics of two pairwise properties: the mean value of interaction frequency depends on the sequence length, as the maximum number of interactions a residue can form is constrained, while the mean value of movement correlation is length-independent. By combining length-dependent attention values (which sum to one after SoftMax operation in attention calculation) with length-independent pairwise embeddings, the model can capture both types of pairwise properties.

Different activation functions were used for different properties: For secondary structure and dihedral angles, where the values for each property sum to one, SoftMax functions were used. For RMSF, which is normalized to (0,1], Sigmoid functions were applied. For correlation properties, Tanh functions were used to maintain values within (−1,1). For other non-negative properties, SoftPlus functions were applied (ReLU led to the dead ReLU problem in our experiments).

Pre-training Procedure

Models were implemented and trained using PyTorch (v 2.5.1)[54] and the Transformers library (v4.48.2, https://huggingface.co/docs/transformers/index). Training was performed with a batch size of 128, parameters were updated every three batches, each from high-resolution data, IDRome, and the NMA dataset. Optimization was conducted using the AdamW optimizer, with a peak learning rate of 1×10^{-4}, an epsilon value of 1×10^{-8}, betas of (0.9, 0.98), and a weight decay of 0.01. The learning rate was linearly warmed up over the first 2,000 steps, followed by a linear decay to 1×10^{-5} over 90% of total training steps. To mitigate overfitting, a dropout rate of 0.1 was applied in both the transformer encoder and the dynamic property prediction heads.

SeqDance was trained for 200,000 steps, with a maximum sequence length of 256 for the first 160,000 steps and 1024 for the remaining 40,000 steps. ESMDance was trained for 60,000 steps, using a maximum sequence length of 256 for the first 40,000 steps and 1024 for the final 20,000 steps. ESMDance has fewer trainable parameters, both its training and test losses plateaued earlier (Figure S1). For sequences exceeding the maximum length, a region of the max length was randomly selected, the selected regions were different in each epoch. Two models were trained for ten days using four Nvidia L40s GPUs. Due to computational constraints, hyperparameter tuning was not performed, which means the current model configuration may not be the best.

Adjusting Weights of Dynamic Properties in the Loss Function

Root mean squared error (RMSE) was used as the loss function for pre-training. For interaction frequencies, some interacting pairs exhibit frequencies close to zero, which are meaningful but treated similarly to zero in RMSE. To address this, we applied a cube root transformation to scale up low-frequency interactions.

The models were pre-trained on multiple tasks of varying scales. To balance the losses across tasks, we used the standard deviations of dynamic properties to calculate task-specific weights. The standard deviation of each property for each protein in the training set was computed. For multi-class features, such as secondary structures, the losses were calculated for each class and then averaged. These values were used to adjust the task-specific weights. Given the higher complexity of pairwise properties compared to residue-level properties, we set the pairwise-to-residue loss ratio to 3:1. For NMA data, this ratio was adjusted independently for the three frequency ranges. The adjusted losses for SASA, RMSF, secondary structure, and dihedral angles were kept equal. The adjusted losses for the nine interactions were kept equal. The ratio between the summed interaction loss and the co-movement loss was set to 3:1. The final losses for high-resolution data, IDRome data, and NMA data were further adjusted to be equal. The final weights for each property are available at https://huggingface.co/datasets/ChaoHou/protein_dynamic_properties.

Attention analysis

The attention map and the interaction map are not symmetric (the interactions have direction), therefore, the means of the original matrixes and its transposes were used. Only residues farther than five amino acids in protein sequence are considered. The interaction frequencies of all nine interaction types were summed for analysis. For proteins longer than 1,024 residues, only the first 1,024 residues were analyzed.

Evaluation of Conformational Properties for IDRs and Ordered Structures

Conformational properties from coarse-grained MD trajectories of over 40,000 IDRs[43] were downloaded. IDRs with a normalized R_g value below 1 were excluded, as they were too compact to be truly disordered. Additionally, only IDRs ranging from 32 to 1024 residues in length were included, IDRs with over 20% sequence identity and 50% coverage (determined using MMseqs2 search with parameters: *-s 7 -a 1*) to any SeqDance training sequence were removed, resulting in a final set of 19,810 IDRs.

R_g values for PDB structures were downloaded from the paper[44], and normalized by fitting a scaling relationship as follows:

$$normalized\ R_g = \frac{R_g}{L^{0.366}} \tag{10}$$

Where L is the sequence length, 0.366 was obtained by fitting a linear regression between $\log(R_g)$ and $\log(L)$.

For both IDRs and ordered structures, MMseqs2 easy-cluster was used to cluster sequences with the parameters: *--min-seq-id 0.2 -c 0.5 --cov-mode 0*. Sequences from 60% of randomly selected clusters were used as the training set, while the remaining sequences formed the test set. SeqDance was compared with METL (METL-G-50M-1D), ESM2 models, and ProSE. The last layer embeddings of SeqDance, METL, and ESM2, and concatenated embeddings of all layers of ProSE (as suggested in their GitHub: https://github.com/tbepler/prose) were used. All embeddings were mean-pooled along protein sequences. Linear regression (scikit-learn v1.4.1) with default parameters was used to predict conformational properties, trained separately for the embeddings from each method. Principle component analysis was conducted with scikit-learn.

Zero-shot Prediction of Mutation Effects

For SeqDance and ESMDance, the dynamic properties of both wild-type and mutated sequences were predicted. For property i, the zero-shot score S_i is computed as:

$$S_i = mean\left(\left|\frac{P_i^{MT} - P_i^{WT}}{P_i^{WT}}\right|\right) \tag{11}$$

where P_i^{WT} and P_i^{MT} represent the predicted values of property i for the wild-type and mutated sequences, respectively. The relative change was computed element-wise. For example, in the case of secondary structure with eight classes, the resulting relative change matrix had a shape of $L\times 8$ for a protein of length L, and the mean of absolute values of this matrix was used as the zero-shot score of secondary structure.

To combine scores from different methods, multiple aggregation strategies were evaluated, including mean, max, weighted mean, and geometric mean, applied to both raw and quantile-normalized scores. Quantile normalization ensures that different score distributions are same by aligning their values as mean of raw values of each rank. Weighted mean and geometric mean were calculated as:

$$S_{\text{weighted}} = \frac{\sum_{i=1}^{M} S_i S_i}{\sum_{i=1}^{M} S_i} \tag{12}$$

$$S_{\text{geometric}} = \left(\prod_{i=1}^{M} S_i\right)^{\frac{1}{M}} \tag{13}$$

Where M is the number of all properties. All these zero-shot scores are non-negative. To compare them with negative ΔΔG values, the negative of the zero-shot scores was used. Various strategies to combine the relative changes of different dynamic properties were tried but not improvement over using individual properties such as mean SASA or dihedral angle *psi* was observed. Users may use these two individual properties for applications.

SeqDance and ESMDance are not suitable for cross-protein prediction in our experiment, as the effects of dynamic property changes vary between proteins.

For ESM2 models, we followed the method used in ProteinGYM. Each residue in a protein was sequentially masked, ESM2 was used to predict the probabilities of 20 amino acids for the masked position. For a protein of length L, ESM2 was run L times to obtain the wild-type and mutated residue probabilities at each site. The log-likelihood ratio (LLR) was then calculated for each mutation. For double mutations, the LLR scores of the two single mutations were summed to approximate the combined effect. For ESM2-15B, half-precision (float16) was used, as the model could not be loaded onto a single GPU with 48 GB of memory.

Mutation Datasets

For the evaluation of protein folding stability, the dataset "Tsuboyama2023_Dataset2_Dataset3_20230416.csv" was downloaded from https://zenodo.org/records/7992926. Insertions and deletions were removed, as well as substitution mutations without ddG_ML values. The mean ddG_ML values for the same sequence were used. Homologs in both the SeqDance/ESMDance model training set and UniRef50 (downloaded in February 2025) were identified using MMseqs2 search with the parameters: *-s 7 -a 1 --max-seqs 1000*.

ProteinGYM[19] data and ESM2 predictions for these mutations were downloaded from its website (https://proteingym.org/) and GitHub in June 2024.

References

1. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596 (2021).
2. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024).
3. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
4. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
5. Kulmanov, M. et al. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence* **6**, 220-228 (2024).
6. Hou, C., Li, Y., Wang, M., Wu, H. & Li, T. Systematic prediction of degrons and E3 ubiquitin ligase binding via deep learning. *BMC Biol* **20**, 162 (2022).
7. Kilgore, H.R. et al. Protein codes promote selective subcellular compartmentalization. *Science* **387**, 1095-1101 (2025).
8. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* **17**, 184-192 (2020).
9. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst* **12**, 654-669 e653 (2021).
10. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118** (2021).
11. Elnaggar, A. et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112-7127 (2022).
12. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* **41**, 1099-1106 (2023).
13. Ding, F. & Steinhardt, J. Protein language models are biased by unequal sequence sampling across the tree of life. *bioRxiv*, 2024.2003.2007.584001 (2024).
14. Zhang, Z. et al. Protein language models learn evolutionary statistics of interacting sequence motifs. *bioRxiv*, 2024.2001.2030.577970 (2024).

15. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024.2007.2001.600583 (2024).
16. Gordon, C., Lu, A.X. & Abbeel, P. Protein Language Model Fitness Is a Matter of Preference. *bioRxiv*, 2024.2010.2003.616542 (2024).
17. Hermann, L., Fiedler, T., Nguyen, H.A., Nowicka, M. & Bartoszewicz, J.M. Beware of Data Leakage from Protein LLM Pretraining. *bioRxiv*, 2024.2007.2023.604678 (2024).
18. Zhang, H., Xu, M.S., Fan, X., Chung, W.K. & Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nat Mach Intell* **4**, 1017-1028 (2022).
19. Notin, P. et al. ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *bioRxiv* (2023).
20. Akdel, M. et al. A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* **29**, 1056-1067 (2022).
21. Vernon, R.M. et al. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *eLife* **7**, e31486 (2018).
22. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
23. Hou, C. et al. PhaSepDB in 2022: annotating phase separation-related proteins with droplet states, co-phase separation partners and other experimental information. *Nucleic Acids Res* **51**, D460-D465 (2023).
24. Gelman, S. et al. Biophysics-based protein language models for protein engineering. *bioRxiv* (2024).
25. Tesei, G. et al. Conformational ensembles of the human intrinsically disordered proteome. *Nature* (2024).
26. Holehouse, A.S. & Kragelund, B.B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol* **25**, 187-211 (2024).
27. Shi, M. et al. Quantifying the phase separation property of chromatin-associated proteins under physiological conditions using an anti-1,6-hexanediol index. *Genome Biol* **22**, 229 (2021).
28. Fung, A., Koehl, A., Jagota, M. & Song, Y.S. The Impact of Protein Dynamics on Residue-Residue Coevolution and Contact Prediction. *bioRxiv*, 2022.2010.2016.512436 (2022).
29. Wayment-Steele, H.K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* (2023).
30. Piana, S., Lindorff-Larsen, K. & Shaw, D.E. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences* **110**, 5915-5920 (2013).
31. Haliloglu, T., Bahar, I. & Erman, B. Gaussian Dynamics of Folded Proteins. *Physical Review Letters* **79**, 3090-3093 (1997).
32. Atilgan, A.R. et al. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* **80**, 505-515 (2001).
33. Mirarchi, A., Giorgino, T. & De Fabritiis, G. mdCATH: A Large-Scale MD Dataset for Data-Driven Computational Biophysics. *Sci Data* **11**, 1299 (2024).
34. Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.C. & Galochkina, T. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Res* **52**, D384-D392 (2024).
35. Rodriguez-Espigares, I. et al. GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat Methods* **17**, 777-787 (2020).

36. Ghafouri, H. et al. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic Acids Res* **52**, D536-D544 (2024).
37. Bakan, A., Meireles, L.M. & Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **27**, 1575-1577 (2011).
38. Kozlova, E., Valentin, A., Khadhraoui, A. & Nakhaee-Zadeh Gutierrez, D. ProteinFlow: a Python Library to Pre-Process Protein Structure Data for Deep Learning Applications. *bioRxiv*, 2023.2009.2025.559346 (2023).
39. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
40. Vaswani, A. et al. in Advances in neural information processing systems 5998-6008 (2017).
41. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins* **87**, 1011-1020 (2019).
42. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020.2012.2015.422761 (2020).
43. Lotthammer, J.M., Ginell, G.M., Griffith, D., Emenecker, R.J. & Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequences. *Nat Methods* (2024).
44. Tanner, J.J. Empirical power laws for the radii of gyration of protein oligomers. *Acta Crystallogr D Struct Biol* **72**, 1119-1129 (2016).
45. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434-444 (2023).
46. Hwang, H.J. et al. Solution structure of the porcine sapovirus VPg core reveals a stable three-helical bundle with a conserved surface patch. *Biochem Biophys Res Commun* **459**, 610-616 (2015).
47. Gu, S. et al. Can molecular dynamics simulations improve predictions of protein-ligand binding affinity with machine learning? *Brief Bioinform* **24** (2023).
48. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**, 1196-1203 (2021).
49. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D.R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics* (2025).
50. Liu, C. et al. Dynamic PDB: A New Dataset and a SE(3) Model Extension by Integrating Dynamic Behaviors and Physical Properties in Protein Structures. (2024).
51. Pang, Y.T., Yang, L. & Gumbart, J.C. From simple to complex: Reconstructing all-atom structures from coarse-grained models using cg2all. *Structure* **32**, 5-7 (2024).
52. McGibbon, R.T. et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **109**, 1528-1532 (2015).
53. Chiang, Y., Hui, W.-H. & Chang, S.-W. Encoding protein dynamic information in graph representation for functional residue identification. *Cell Reports Physical Science* **3** (2022).
54. Paszke, A. et al. in Proceedings of the 33rd International Conference on Neural Information Processing Systems Article 721 (Curran Associates Inc., 2019).