GenUnfold: Rapidly Predict Protein Mechanical Unfolding Trajectory via a Physics-Guided Diffusion Model

Anonymous Authors[1]

Abstract

Many fundamental biological processes are governed by mechanical forces, with proteins acting as the key molecular mediators. Elucidating how protein unfolding responds to force is critical for understanding the mechano-pathologies, such as cardiomyopathy and muscular dystrophy. While the unfolding trajectories measured by Single-Molecule Force Spectroscopy (SMFS) map the instantaneous force response against molecular extension, its broader application is limited by time-consuming data collection and high operational costs. Here, we present the first scalable generative diffusion framework for full unfolding trajectory prediction, which integrates protein encoders for multi-scale conditioning. Beyond establishing the field's first systematic benchmark using existing models, we propose GenUnfold, a novel physics-guided diffusion model that combines global coevolutionary context with a local mechanical representation of the protein. The representation is derived from a novel physics-biased attention mechanism, which steers the generative diffusion process by modeling dynamic residue dependencies as a function of both structural topology and interaction stiffness. The benchmark for this task is built upon the biomolecule stretching database and several representative baseline models. Empirical results demonstrate that GenUnfold achieves state-of-the-art performance, reducing distributional error (FID) by 30% and 54% compared to pretrained Evolutionary Scale Model (ESM)-2 and standard transformer, respectively. Beyond statistical curve similarity, GenUnfold demonstrates superior physical consistency; in downstream mechanical property prediction, it reduces prediction errors for unfolding force and energy distributions by 6% and 36% over the ESM-2 baseline. These results indicate that while existing generative AI approaches can alleviate the need for predicting representative force curves, GenUnfold further improves performance by leveraging the synergy between protein structure and evolutionary information. By enabling proteome-wide screening to identify mechanical candidates before costly physical validation, our approach is promising to accelerate the discovery of force-targeted therapeutics.

1. Introduction

The controlled mechanical unfolding of protein molecules is a critical biological process in *vivo* (Bustamante et al., 2004; Beedle & Garcia-Manyes, 2023), particularly within load-bearing tissues such as muscle and the cytoskeleton (Marszalek et al., 1999). The stability and function of these mechanical proteins heavily depend on how forces are transmitted and dissipated through their topological structures (Fernandez & Li, 2004; Stacklies et al., 2009). For example, in the protein dystrophin, the central spectrin-like repeat domains with a three-helix bundle structure act as molecular shock absorbers. By unfolding during muscle contraction, these domains dissipate mechanical energy to protect the myofiber membrane from damage (Ervasti, 2007). Furthermore, understanding the relationship between protein structure and mechanical response is fundamental to modern therapeutic design. For instance, in Duchenne Muscular Dystrophy (a fatal disease affecting 1 in 5,000 births born in the U.S.), identifying mechanically resilient analogs to replace missing or defective dystrophin remains a critical strategy for effective therapy but is still under study (Mendell et al., 2012; Ervasti, 2007).

Single-molecule force spectroscopy (SMFS) currently serves as the main technique to probe the mechanical unfolding of protein molecules (Viljoen et al., 2021). Instruments such as Atomic Force Microscopy (AFM) and optical tweezers apply external force to the protein of interest, driving it from a native equilibrium state to probe non-equilibrium unfolding dynamics (Neuman & Nagy, 2008). The measured

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

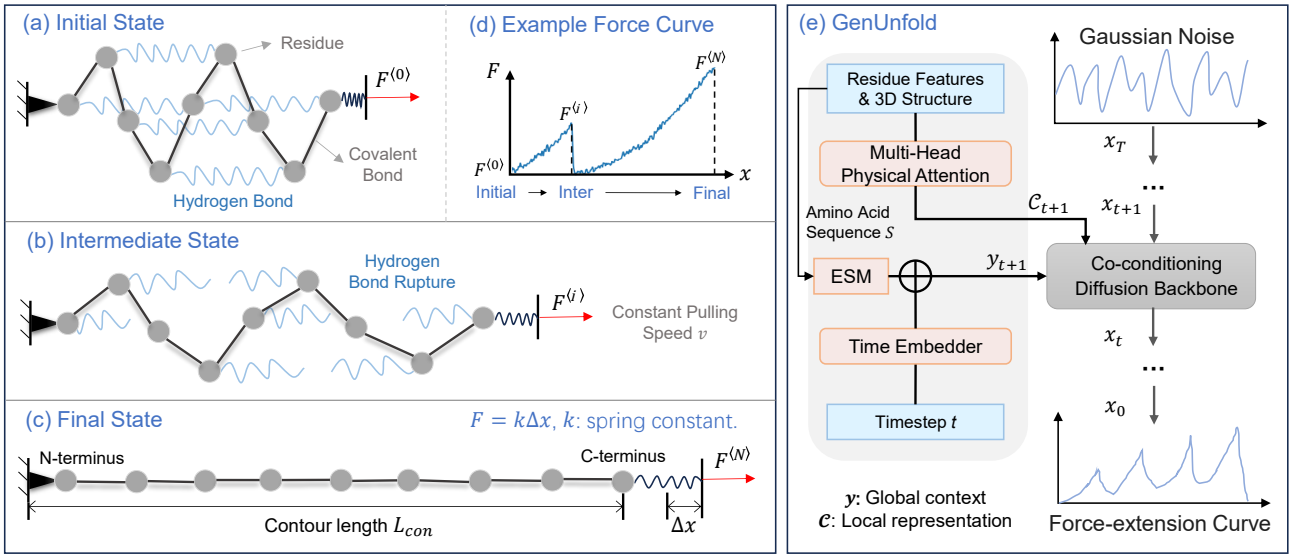

Figure 1. Overview of the protein unfolding process and the proposed modeling framework. Protein is represented as a spring network for visualization. (a-d) In the MD simulation, a tethered protein is mechanically stretched, yielding a F-x trajectory that encodes the unfolding dynamics. (e) GenUnfold extracts structure prior maps and residue features from the amino acid sequence and native structure of the protein as a conditional signal to predict the full unfolding trajectory.

force-extension (F-x) trajectories provide rich insights into the protein's free energy landscape (captured by total unfolding energy) and mechanical stability (characterized by the unfolding force) (Sengupta & Rief, 2021). However, the unfolding process is inherently stochastic and heterogeneous (Schönfelder et al., 2016), which necessitates thousands of F-x curves from repetitive pulling experiments to obtain statistically reliable ensembles (Hua et al., 2025). This requirement renders SMFS labor-intensive and time-consuming, creating a significant bottleneck in studying large-scale proteins.

Molecular Dynamics (MD) simulations offer a computational alternative, providing atomic-level resolution of unfolding dynamics and structural changes (Li & Makarov, 2003). The high-fidelity data generated by MD simulations can be exploited for data-driven, learning-based methodologies. However, the utility of MD for high-throughput research is constrained by the imperfections of current force fields (Lewis et al., 2025) and the prohibitive computational cost of sampling the thousands of trajectories required for statistical robustness. While coarse-grained alternatives offer improved computational efficiency, their ability to model complex unfolding pathways is often limited and remains an active area of development (Joshi & Deshmukh, 2021).

The emerging data-driven machine learning, particularly generative diffusion models, have been widely used in protein structure prediction and design (Jumper et al., 2021; Gruver et al., 2023). Recent works, such as BioEmu (Lewis et al., 2025), have demonstrated the capability to generate diverse structural ensembles and can be coupled with MD to mitigate sampling limitations. Moreover, Large-scale Protein Language Models (PLMs) such as Evolutionary Scaling Modeling (ESM)-2 (Lin et al., 2023) reveal that coevolutionary patterns in amino acid sequence data encode rich structural and functional information and provide identity of the protein. Despite these advances, current machine learning methods focus primarily on predicting equilibrium states of protein structures and fail to account for the directional, non-equilibrium dynamics driven by external force. Currently, there is no scalable, learning-based method capable of predicting full F-x trajectories and the specific mechanical properties they encode.

To address this gap, we present the first generative diffusion framework for predicting full protein unfolding trajectories conditioned on protein sequence and structure. We begin by building the benchmark with a suite of current State-of-the-Art deep learning approaches, e.g., standard Transformers and pretrained ESM-2. These models are used for encoding protein features as conditioning signals. Moreover, we propose GenUnfold, a novel physics-guided diffusion model that integrates multi-scale physical priors. The approach integrates coevolutionary information from the amino acid sequence for high-level semantic guidance, and structure-derived topology priors that explicitly capture residue-residue dependencies and force propagation pathway critical for modeling mechanical response. Our contributions include:

Evolution-Topology Co-Conditioning. We propose a diffusion backbone that effectively routes information through two parallel streams: high-level coevolutionary context via

Adaptive Layer Normalization (adaLN) and detailed mechanical representation via cross-attention. This architecture enables the model to generate trajectories that are both biologically plausible and mechanically consistent.

Physics-biased Attention Mechanism. We introduce a novel attention mechanism that fuses multi-scale structural topology priors into a learnable attention bias to model residue-residue dependencies and force propagation pathways. A lightweight CNN learns nonlinear combinations of these priors, while a temporal gating mechanism modulates their influence in different diffusion timesteps.

Empirical validation. We establish the first comprehensive benchmark for protein unfolding trajectory prediction. GenUnfold achieves state-of-the-art results, reducing distributional error by over 30% to 54% compared to strong baselines (ESM-2 and standard Transformers). Furthermore, it demonstrates high physical fidelity, reducing prediction error for unfolding force and energy distributions by 6% and 36%, respectively, offering a robust computational complement to physical SMFS.

2. Related Work

Generative Modeling of Molecular dynamics. Current generative modeling in biology mainly focuses on sampling 3D conformational ensembles or de novo design. Diffusion models, e.g., RFdiffusion (Watson et al., 2023), Chroma (Ingraham et al., 2023), and Flow Matching methods have achieved remarkable success in generating protein backbones by reversing a noise process on 3D coordinates. Similarly, Boltzmann Generators (Noé et al., 2019) and Torsional Diffusion (Jing et al., 2022) learn to sample equilibrium distributions to characterize metastable states. Whereas existing models generate 3D geometric coordinates to sample equilibrium states, our model aims to generate 1D functional signals (F-x trajectories) representing a stochastic, non-equilibrium process. We repurpose the diffusion framework not only to design or sample a 3D structure of a protein, but to simulate the dynamic response of a structure under external force.

Physics-Guided Inductive Biases. Injecting physical constraints into the learning process is a key strategy for improving generalization in scientific ML. Common approaches include enforcing E(n)-equivariance that model predictions ensure are invariant to global translations and equivariant to rotations and reflections (Garcia Satorras et al., 2021) or constraining outputs to satisfy Hamiltonian mechanics (Greydanus et al., 2019). In the context of protein mechanics, classical Elastic Network Models (ENM) (Atilgan et al., 2001) define harmonic interactions based on geometric proximity. We argue that geometric proximity alone is insufficient for modeling force propagation. GenUnfold introduces a novel physics-biased attention mechanism that integrates multi-scale physical priors directly into the attention bias. Our model routes information along stiff mechanical pathways, mimicking the physical transmission of stress through the polypeptide chain.

To the best of our knowledge, GenUnfold is the first generative framework designed to predict mechanical unfolding trajectories directly from protein sequence and structure. By bridging the gap between high-fidelity MD simulations and scalable generative deep learning, it offers a new framework for characterizing protein properties.

3. Preliminary

3.1. Problem statement

As shown in Fig. 1, in a typical steered MD simulation, the N-terminus (beginning end) of the studied protein molecule is anchored, and the C-terminus (end) is tethered to a moving harmonic spring with stiffness k (Sikora et al., 2010). The free end of this spring (dummy atom) moves away at a constant velocity v. The instantaneous pulling force is computed via Hooke's law based on the spring's extension Δx, $F = k(\Delta x)$. This process continues until the end-to-end extension, L_{ac}, reaches the protein's theoretical contour length, $L_{con} = L \times 3.6$ Å for a protein of L residues (Ainavarapu et al., 2007). The resulting F-x trajectories $\mathbf{F} = [F_1, F_2, \ldots, F_N]^\top \in \mathbb{R}^N$ (Fig. 1d) capture the mechanical unfolding dynamics, and the unfolding force F_i represents the mechanical stability of the protein.

In this study, a protein molecule is defined by its amino acid sequence S, and a coarse-grained structure $\mathbf{X} \in \mathbb{R}^{L\times 3}$ (one 3D coordinate per C_α atom of the residue). A key challenge is that the static structure $\mathbf{X}$ does not explicitly encode the mechanical pathways or force propagation networks that govern unfolding. To bridge this gap, we transform the raw static inputs (S and $\mathbf{X}$) into two mechanics-aware feature sets to serve as inductive biases: (1) Residue-level Features ($\mathbf{R_f}$): A set of 17 chemical features, e.g, residue type, hydrophobic and charge category, describing the local properties of each residue (details in Appendix A.2). (2) Physical Topology Priors ($\mathcal{G}$): A multi-channel graph representation derived from $\mathbf{X}$ that explicitly encodes long-range residue interaction by geometric adjacency, residue-residue interaction strength, and global mechanical coupling. (The derivation of $\mathcal{G}$ will be introduced in Section 4.2.) The resulting dataset denotes as $\mathcal{D} = \{\mathbf{F}_i, S_i, \mathbf{X}_i, \mathbf{R_f}_{,i}, \mathcal{G}_i\}_{i=1}^M$ with M samples.

Our objective is the conditional generative modeling of the full protein mechanical unfolding trajectories $\mathbf{F}$. Formally, we seek to learn the parameters θ of a neural network to approximate the conditional distribution $p_\theta(\mathbf{F} \mid S, \mathbf{X}, \mathbf{R_f}, \mathcal{G})$.

3.2. Diffusion model

Diffusion Probabilistic Model. For generative modeling, the diffusion model involves a forward process that incrementally adds noise to a data point $x_0 \sim q(x)$ until it reaches a noise-only state $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (Sohl-Dickstein et al., 2015). The transition at step t is governed by $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$, where constants β_t are hyperparameters. The core of the model lies in learning a reverse generative process $p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \sigma_\theta(x_t, t))$ by a neural network with parameters θ. By reparameterizing μ_θ as a noise prediction network ϵ_θ, the model can be trained using simple mean-squared error between the predicted noise $\epsilon_\theta(x_t, t)$ and the ground truth sampled Gaussian noise ϵ_t (Ho et al., 2020):

$$\mathcal{L}(\mathbf{x}_0) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(x_t|x_0)} \|\epsilon_\theta(x_t, t) - \epsilon_t\|_2^2 . \tag{1}$$

Transformer-based Diffusion Backbones. Diffusion Transformer (DiT) (Peebles & Xie, 2023) demonstrated the superior potential of Transformer-based architectures in terms of scalability and performance. To incorporate conditioning into the generation process, DiT proposed an efficient modulation mechanism through adaLN. It maps scalar or vector conditions (such as the diffusion timestep t and class labels y) through a small Multilayer Perceptron (MLP) to generate the parameters (γ, β) for an affine transformation. These parameters are then applied in the layer norm before self-attention, thereby modulating the network's behavior to guide the denoising process. However, the adaLN mechanism is primarily suited for global, single-vector conditions with compressed information (Podell et al., 2023). To handle more complex, sequential conditioning, such as textual descriptions, subsequent works, e.g., PIXART-α (Chen et al., 2023), introduced a key extension to the DiT architecture. They insert a cross-attention layer within each Transformer block, positioning it between the self-attention layer and the feed-forward (FFN) layer. This design allows the model to actively "query" the external condition sequence (e.g., sequential text embeddings) after processing its own tokens (i.e., the noisy x_t), enabling finer-grained guidance.

4. GenUnfold

As illustrated in Fig. 2, we introduce **GenUnfold**, a physics-guided diffusion framework tailored to predict dynamics protein mechanical unfolding (F-x) trajectories from protein sequence and structure. The core challenge in this task is bridging the modality gap: mapping a static, equilibrium structure to a stochastic, non-equilibrium unfolding process. To address this, GenUnfold employs a co-conditioning architecture of DiT with cross-attention to integrate two complementary streams of information:

1. *Coevolutionary context (y_t) by adaLN:* A high-level semantic representation of protein sequence derived from coevolutionary language models, capturing the protein's global identity.

2. *Mechanical Representation (c_t) by cross-attention:* A fine-grained, mechanically informed structure embedding. This is generated by a novel Physics-biased Attention that guides the attention mechanism using physical topology priors.

The following subsections detail the extraction of these conditions and their integration into the generative backbone.

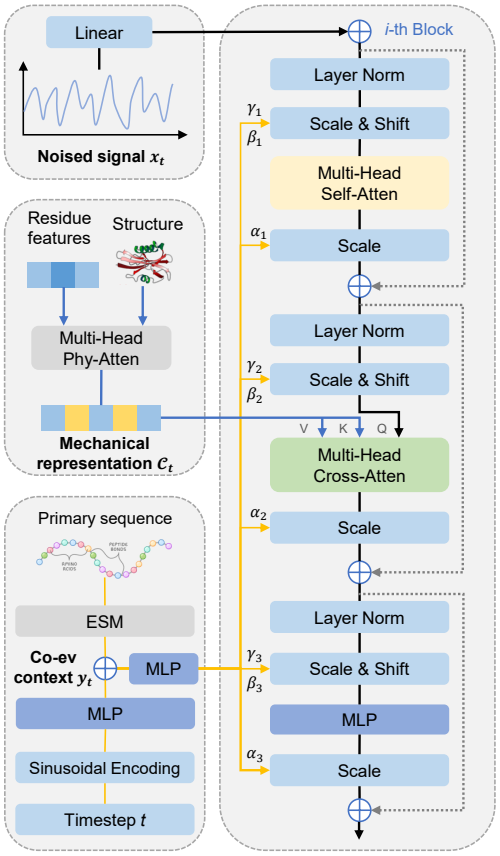

Figure 2. The Diffusion Backbone of GenUnfold. Coevolutionary context (y_t) modulates DiT blocks via adaptive layer norm (adaLN), and mechanical representations (c_t) are injected directly through Cross-Attention layers.

4.1. Coevolutionary Context

To capture coevolutionary information crucial for protein mechanics, we employ the ESM-2, a large-scale PLM trained on millions of diverse protein sequences (Lin et al., 2023). Our motivation is that the coevolutionary patterns learned by ESM-2 serve as a powerful proxy for the structural and functional relationships that govern mechanical stability.

Given a protein sequence, we extract residue-level embeddings from the last hidden layer of ESM-2 and apply mean-pooling to derive a global semantic vector $y_{\text{seq}} \in \mathbb{R}^H$, where H is the hidden size of the neural network. Concurrently, the diffusion timestep t is embedded via sinusoidal encoding and an MLP to form $t_{\text{emb}} \in \mathbb{R}^H$. Then, they are pointwise connected to form the final *coevolutionary context vector* $y_t = y_{\text{seq}} + t_{\text{emb}}$. In the diffusion backbone, y_t modulates the generation process via adaLN-Zero, effectively steering the global distribution of the generated curves based on protein identity and noise level.

4.2. Mechanical Representation via Multi-head Physics-biased Attention

While standard Transformer architectures excel at processing sequence data (Vaswani et al., 2017), they treat residues purely as tokens, may lack the explicit inductive biases required to model the long-range residue independencies and complex, heterogeneous force propagation pathways in protein unfolding. To bridge this gap, we introduce the Multi-head Physics-biased Attention (Fig. 3), which injects multi-scale physical topology priors directly into the attention mechanism, ensuring the learned local representation c respects the protein's native topology and elasticity.

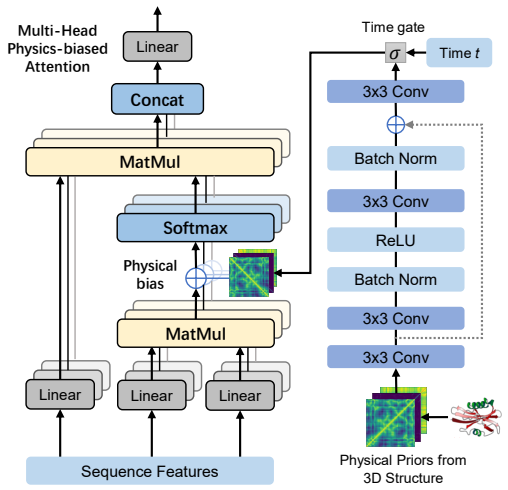

Figure 3. Architecture of the Multi-head Physics-Biased Attention. A CNN extracts features from the structural graph ($\mathcal{G}$), which are temporally gated and added to attention logits to constrain the learned mechanical representation (c_t).

Physics-Guided Multi-Channel Priors. To derive mechanically meaningful features from the 3D structure $\mathbf{X}$, we model the protein structure as a harmonic spring network (Atilgan et al., 2001). Under this setting, residues act as nodes, which are fully interconnected by simple harmonic springs that approximate the interaction stiffness and topological elasticity.

From static structure $\mathbf{X}$, We construct a multi-channel graph $\mathcal{G} = [\mathbf{D}, \mathbf{K}, \mathbf{R}] \in \mathbb{R}^{3\times L\times L}$ comprising: (1) *Distance Map* ($\mathbf{D}$): The pairwise Euclidean distance d_{ij} between residues i and j. (2) *Stiffness Map* ($\mathbf{K}$): Modeling local interaction strength. The spring constant is defined by $k_{ij} = d_{ij}^{-2}$ (for $i \neq j$) or 0 (for $i = j$) (Yang et al., 2009) to approximate contact stiffness. (3) *Resistance Map* ($\mathbf{R}$): Quantifying global mechanical coupling and elasticity. We compute the Graph Laplacian $\mathbf{L} = \mathbf{D_{deg}} - \mathbf{K}$, where the degree $(\mathbf{D_{deg}})_{ii} = \sum_j k_{ij}$, and its pseudoinverse $\mathbf{G} = \mathbf{L}^+$. The effective resistance distance is given by $R_{ij} = G_{ii} + G_{jj} - 2G_{ij}$. Intuitively, low R_{ij} indicates a "stiff corridor" efficient for force transmission, providing a global topological prior that complements local neighborhood information.

Learnable Physical Bias and Temporal gating Rather than using the priors $\mathcal{G}$ as fixed features, we employ a CNN-based Bias Module to learnable attention bias (see Fig. 3). The graph $\mathcal{G}$ is processed by a 3×3 convolutional layer followed by a residual block consisting of two stacked 3×3 convolutions with interleaved Batch Normalization and ReLU activations. A final projection maps the fused features to h distinct bias maps, $B_{\text{cnn}} \in \mathbb{R}^{h\times L\times L}$, matching the number of attention heads. This design promotes attentional specialization that one head may learn to focus on local rigidity (dominated by $\mathbf{K}$), while another attends to long-range coupling pathways (informed by $\mathbf{R}$).

Crucially, the relevance of extracted features B_{cnn} changes as the diffusion process reconstructs the trajectory. For instance, while D and K provide structural priors for the initial denoising step, R offers essential guidance for capturing the unfolding events as force propagates through topological bottlenecks in the final fine-tuning denoising stage. To model this, we introduce a Time-Dependent Gating mechanism. The diffusion time embedding t_{emb} is projected via an MLP to a scaling vector $\mathbf{g}(t) \in \mathbb{R}^H$, which allows the model to dynamically upweight or downweight physical constraints as the generative denoising progresses. The final attention bias for the i-th head is a dynamically modulated version of the physical map:

$$B^{(i)}_{\text{final}}(t) = g^{(i)}(t) \cdot B^{(i)}_{\text{cnn}} \tag{2}$$

Protein Encoder Block The encoder processes residue features $\mathbf{R}_f$ via a stack of tranformer blocks with physics-biased attention. We modify the standard self-attention mechanism by injecting the learned physical bias into the softmax logits:

$$\text{Attn}^{(i)}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{Q}^{(i)}{\mathbf{K}^{(i)}}^\top}{\sqrt{d_k}} + B^{(i)}_{\text{final}}(t)\right)\mathbf{V}^{(i)} \tag{3}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are linear projections of the layer-normalized input. This biased attention output is then processed by a standard FFN with pre-normalization to produce the final physical representation c. This embedding is subsequently fed into the cross-attention layers of the DiT backbone, enabling the generator to attend to mechanically critical residues during trajectory generation. More details of the model architecture are provided in the Appendix B.

5. Experiments

5.1. Dataset

We train and validate GenUnfold primarily using high-quality MD data sourced from the Bio-molecule Stretching Database (BSDB) (Sikora et al., 2010). We cross-referenced BSDB with the Protein Data Bank (PDB) to filter for proteins with available high-resolution experimental structures. The final curated dataset comprises 16,495 unique proteins and 32,990 F-x curves (each protein has two curves). The orginal F-x curves were preprocessed by two standardization steps to facilitate neural network training. First, to handle variable protein lengths, extension values are normalized by the theoretical contour length ($x' = x/L_{con}$) of corresponding protein and subsequently resampled onto a fixed-resolution grid of N points via linear interpolation. Second, to standardize the force range while preserving the relative mechanical strength between different proteins, force magnitudes are normalized using a global Min-Max scaling. Specifically, all force values are scaled to $[0, 1]$ based on the maximum orce observed across the training set of the dataset $\mathcal{D}$, rather than on a per-protein basis.

Conditioning signals are derived directly from PDB structures. Specifically, We precompute: (1) **Coevolutionary Context:** Coevolutionary embeddings of amino acid sequence via ESM-2. (2) **Residue-level Features:** A residue-level matrix $\mathbf{R_f}$ containing physicochemical properties (charge, polarity), geometric descriptors (secondary structure, SASA), and B-factors via DSSP (Kabsch & Sander, 1983). 3) **Physical Topology Priors:** The precomputed geometric and mechanical maps ($\mathcal{G}$) described in Sec. 4.2. Moreover, We employ a strict protein-level split (80:10:10) to create the training, validation, and test sets. Crucially, we ensure that all trajectories associated with a specific protein sequence are confined to a single split. More data and preprocessing details are provided in Appendix A.

5.2. Evaluation metrics

We evaluate performance across three complementary dimensions: signal reconstruction, distributional fidelity, and biophysical consistency. We adopt the following evaluation metrics (see Appendix C. for the detailed descriptions): (1) **Relative l_2 Error (Rel l_2)** (Ni et al., 2024) to quantify trajectory accuracy; (2) **Fréchet Inception Distance (FID)** (Jeha et al., 2022) to assess whether the generated trajectories capture the realistic, high-dimensional statistics of real unfolding curves; (3) **Jensen–Shannon distance (JSD)** to evaluate whether the model reproduces the distributions of two physically salient properties (Ni et al., 2024): unfolding force $F_{\text{unfold}} = \max_x\{F(x) \cdot \mathbf{1}(\text{unfolding event at } x)\}$ (Force-JSD) and unfolding energy $W = \int^{L_{con}} F dL_{ac}$ (Energy-JSD).

All metrics are computed on the test set for three independent runs, with means and variance reported.

5.3. Benchamark

Since no prior methods aim to generate full F-x trajectories, we design a comprehensive benchmark to evaluate the contributions of our key components: the evolution-topology co-conditioning architecture and the physics-biased attention. We evaluate two classes of models:

1. Global-Conditioned Baselines (Standard DiT). These models use a standard DiT backbone conditioned solely via adaLN with context $y \in \mathbb{R}^H$, testing if a single global vector is sufficient to capture unfolding dynamics: (1) **DiT-SelfAtten (baseline):** Uses a standard transformer to process residue-level features $\mathbf{R_f}$ and pools the output globally. (2) **DiT-ESM:** Uses pooled ESM-2 embeddings, testing the sufficiency of pure sequence/coevolutionary data. (3) **DiT-PhyAtten:** Uses our physics-biased attention but pools the output globally, testing the value of structural topology priors without coevolutionary guidance.

2. Co-Conditioning Models (CrossDiT). These models utilize the full diffusion backbone, injecting global context (y) via adaLN and fine-grained local context ($c \in \mathbb{R}^{L\times H}$) via Cross-Attention: (1) **CrossDiT-ESM-SelfAtten:** Combines global ESM context with a standard transformer. (2) **CrossDiT-ESM-ESM:** A pure sequence variant using ESM-2 for both full embeddings (global) and pooled embeddings (local), ignoring explicit physics. (3) **CrossDiT-ESM-PhyAtten (GenUnfold):** The proposed model combines global coevolutionary guidance with local physical mechanics. (4) **CrossDiT-PhyAtten-PhyAtten:** A physics-only variant excluding ESM, testing the necessity of evolutionary signals.

All variants share identical hyperparameters for the diffusion schedule, token budget, training epochs, and optimization.

6. Results

6.1. Performance Evaluation

We evaluate the performance of our model components using the metrics summarized in Table 1. Relative to the base-

Table 1. Main results on the BSDB test set. Lower is better.

Model	Rel $l_2 \downarrow$	FID $\downarrow$	Force_JSD $\downarrow$	Energy_JSD $\downarrow$
DiT-SelfAtten	0.4384 ± 0.0078	0.0256 ± 0.0106	0.0832 ± 0.0076	0.0638 ± 0.0055
DiT-ESM	0.2165 ± 0.0080	0.0220 ± 0.0083	0.0828 ± 0.0046	0.0575 ± 0.0053
DiT-PhyAtten	0.2494 ± 0.0045	0.0219 ± 0.0079	0.0584 ± 0.0075	0.0494 ± 0.0032
CrossDiT-ESM-SelfAtten	0.2117 ± 0.0199	0.0215 ± 0.0133	0.0668 ± 0.0136	0.0517 ± 0.0192
CrossDiT-ESM-ESM	$\mathbf{0.2017 \pm 0.0007}$	0.0168 ± 0.0090	0.0590 ± 0.0031	0.0533 ± 0.0097
CrossDiT-ESM-PhyAtten	$\underline{0.2070 \pm 0.0050}$	$\mathbf{0.0117 \pm 0.0011}$	$\mathbf{0.0553 \pm 0.0032}$	$\mathbf{0.0338 \pm 0.0038}$
CrossDiT-PhyAtten-PhyAtten	0.2274 ± 0.0050	$\underline{0.0165 \pm 0.0055}$	$\underline{0.0564 \pm 0.0026}$	$\underline{0.0370 \pm 0.0100}$

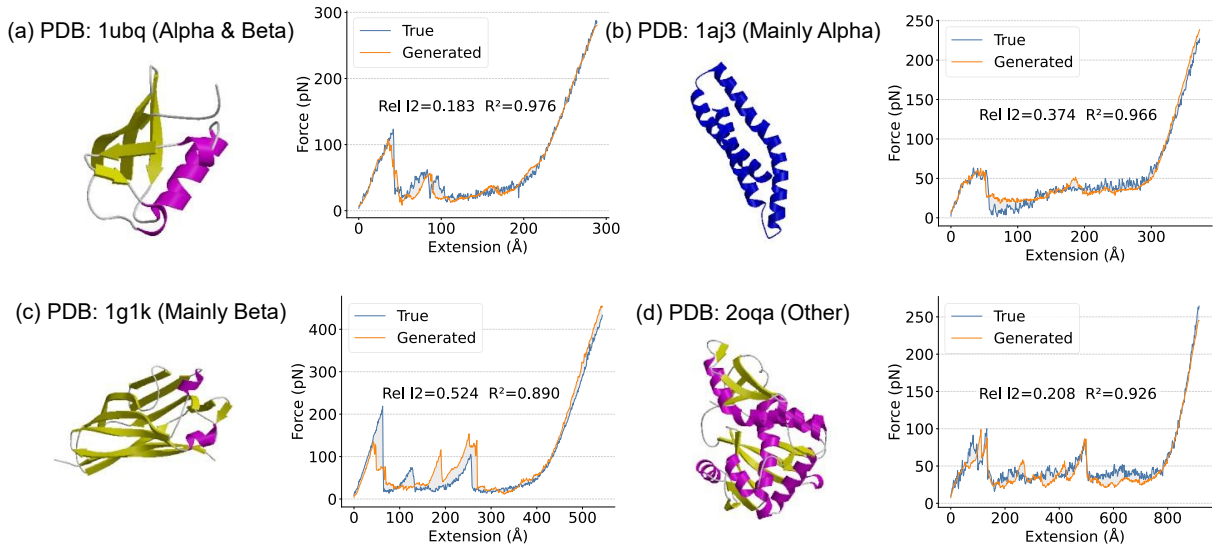

Figure 4. Representative examples from the test set showing the correspondence between 3D structure (left) and F-x curve (right).

line (DiT-SelfAtten), injecting coevolutionary information from ESM-2 (DiT-ESM) significantly improves reconstruction, reducing Rel l_2 by 50.6% and FID by 14.1%. This confirms that the global sequence context effectively acts as a high-level semantic guide for the trajectory's overall shape. On the other hand, while DiT-PhyAtten shows a more moderate improvement in Rel l_2 (43.1%) compared to the DiT-ESM, it achieves a competitive FID of 0.0219, representing a 14.5% reduction from the baseline. Crucially, DiT-PhyAtten reduces the Force-JSD by 29.8% and Energy-JSD by 22.6% compared to the baseline, outperforming DiT-ESM on these physics-sensitive metrics. The result indicates that explicit mechanical constraints are essential for steering the generation toward physically valid unfolding pathways.

Furthermore, the co-conditioning architecture yields the strongest performance. While the pure-sequence variant (CrossDiT-ESM-ESM) achieves the lowest point-wise error (Rel l_2: 0.2017), it lags in distributional fidelity. The proposed model, **GenUnfold** (CrossDiT-ESM-PhyAtten), achieves the optimal balance by fusing global coevolutionary guidance with local mechanical representation turning. It sets a new state-of-the-art on the critical generative metrics, achieving the lowest FID (0.0117), a 54.3% reduction from baseline. CrossDiT-ESM-PhyAtten also yields the most physically accurate ensembles, lowering Force-JSD by 33.5% and Energy-JSD by 47.0% (from 0.0638 to 0.0338). The dramatic improvement in Energy-JSD underscores the model's ability to maintain long-range mechanical consistency.

6.2. Structural Generalizability and Diversity

We further assess whether GenUnfold captures the structure-function relationship governing mechanostability. We classify the test set by CATH topology (Waman et al., 2025), including Mainly alpha (composed primarily of alpha-helices), Mainly beta (primarily beta-sheets), Mixed alpha and beta, and Few Secondary Structures (irregular or loop-dominated folds). As illustrated in Fig. 4, the model successfully treats the 3D structure as a fingerprint for mechanical phenotype. For beta-sheet rich proteins, which resist unfolding via cooperative hydrogen bond shearing, the model generates characteristic sawtooth patterns with high-force peaks

Table 2. Performance comparison across model variants of the proposed physics-biased attention. Lower is better.

Variant	Rel l_2	FID	Fmax_JSD	Energy_JSD
Remove K and R	**0.2053**±0.0036	0.0178±0.0010	0.0656±0.0101	0.0520±0.0206
Remover R	0.2104±0.0039	0.0167±0.0027	0.0596±0.0086	0.0396±0.0008
Remove time gate	0.2063±0.0015	**0.0128**±0.0005	**0.0557**±0.0033	0.0390±0.0025
Remove CNN learning	0.2079±0.0048	0.0132±0.0025	0.0569±0.0037	**0.0378**±0.0067

($>$ 100 pN). Conversely, for alpha-helical domains, it synthesizes compliant, noisy profiles typical of "unzipping" mechanisms. This capability confirms that GenUnfold conditions its generation on specific geometric constraints rather than memorizing a mean trajectory.

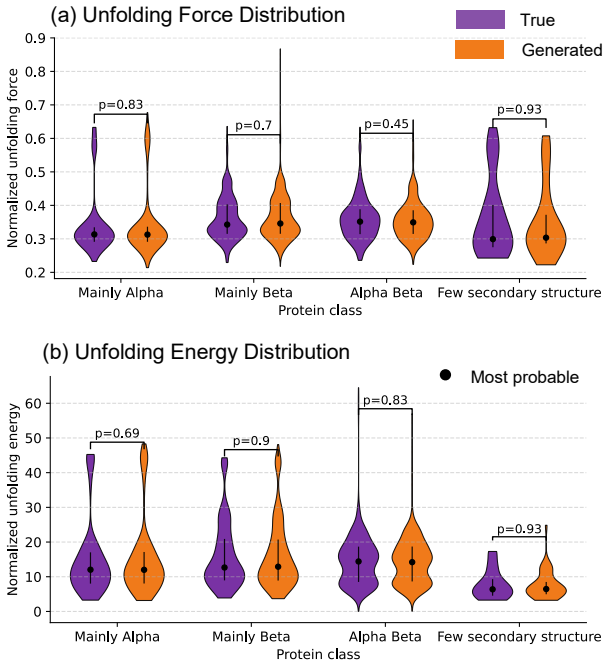

Figure 5. Statistical validation of mechanical properties across structural classes. Violin plots show normalized force and energy distributions for ground truth (blue) and model (orange). High p-values across CATH classes demonstrate topological generalizability and physical consistency of predicted unfolding trajectories.

As shown in Fig. 5, we further compared the distributions of downstream mechanical properties (unfolding Force F_{unfold} and Energy W) between predicted and ground-truth ensembles across all structural classes. It can be observed that the generated distributions (orange) closely overlay the ground truth (blue). This visual alignment is confirmed by Mann-Whitney U tests (McKnight & Najab, 2010), which yield high p-values across all categories (e.g., $p = 0.54$ for Mainly alpha, $p = 0.83$ for Mainly beta). These high p-values ($p \gg 0.05$) indicate that the generated and simulated ensembles are statistically indistinguishable.

6.3. Ablation Study of Physics-biased Attention

To isolate the source of our performance gains, we conducted a systematic ablation of the proposed physics-biased attention, with results presented in Table 2. The results show that removing the graph components ($\mathbf{K}, \mathbf{R}$) from physical topological priors $\mathcal{G}$ leads to a severe degradation in generative quality, increasing FID by 52.1% and Energy-JSD by 53.8%. Interestingly, removing $\mathbf{K}$ and $\mathbf{R}$ marginally improves point-wise Rel l_2 (0.2070 $\rightarrow$ 0.2053) but destroys the physical realism of the curves. This highlights a critical trade-off: without mechanistic priors, the model collapses toward a mean trajectory that minimizes error but fails to capture the complex, stochastic energy landscape of unfolding. Removing the CNN-based feature learning ("Remove CNN") or the time-dependent gating ("Remove time gate") results in consistent performance drops. This confirms that allowing the model to learn the dynamic optimal representation of the physical priors is beneficial.

7. Conclusion

While protein mechanical responses are fundamental to biological function, capturing their dynamic unfolding trajectories has historically relied on labor-intensive SMFS experiments or computationally expensive simulations that cannot scale to the whole proteome. We addressed this limitation by introducing the first scalable generative diffusion framework for unfolding trajectory prediction. We formalize this domain by introducing a comprehensive benchmarking suite, adapting current state-of-the-art protein language models and transformers. Moreover, we propose GenUnfold, a physics-guided diffusion model that fuses global coevolutionary context and local structural stiffness to accurately reconstruct complex force-extension profiles. The empirical results on our newly established benchmark confirm that GenUnfold achieves a new state-of-the-art in distributional accuracy, outperforming standard transformer baselines and pretrained protein language models by significant margins. Crucially, our model does not merely mimic the mean shape of the data; it exhibits high physical fidelity in predicting force and energy distributions, proving that deep generative models can internalize the fundamental physics guided by appropriate structural priors. GenUnfold provides a powerful tool for discovering force-targeted therapeutics and exploring the mechanical proteome.

Impact Statement

GenUnfold addresses a long-standing bottleneck in biophysics: the inability to rapidly and accurately characterize how proteins respond to mechanical force. By establishing the first scalable generative framework for predicting non-equilibrium unfolding trajectories, this work transitions protein mechanics from a labor-intensive, single-molecule experimental endeavor to a proteome-wide computational science.

References

Ainavarapu, S. R. K., Brujić, J., Huang, H. H., Wiita, A. P., Lu, H., Li, L., Walther, K. A., Carrion-Vazquez, M., Li, H., and Fernandez, J. M. Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophysical journal*, 92(1):225–233, 2007.

Atilgan, A. R., Durell, S., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

Beedle, A. E. and Garcia-Manyes, S. The role of single-protein elasticity in mechanobiology. *Nature Reviews Materials*, 8(1):10–24, 2023.

Bustamante, C., Chemla, Y. R., Forde, N. R., and Izhaky, D. Mechanical processes in biochemistry. *Annual review of biochemistry*, 73(1):705–748, 2004.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al. Pixart-: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Ervasti, J. M. Dystrophin, its interactions with other proteins, and implications for muscular dystrophy. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1772(2):108–117, 2007.

Fernandez, J. M. and Li, H. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science*, 303(5664):1674–1678, 2004.

Garcia Satorras, V., Hoogeboom, E., Fuchs, F., Posner, I., and Welling, M. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34: 4181–4192, 2021.

Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

Gruver, N., Stanton, S., Frey, N., Rudner, T. G., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., and Wilson, A. G. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36: 12489–12517, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hua, C., Rajaganapathy, S., Slick, R. A., Vavra, J., Muretta, J. M., Ervasti, J. M., and Salapaka, M. A physics-augmented deep learning framework for classifying single molecule force spectroscopy data. In *Forty-second International Conference on Machine Learning*, 2025.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Jeha, P., Bohlke-Schneider, M., Mercado, P., Kapoor, S., Nirwan, R. S., Flunkert, V., Gasthaus, J., and Januschowski, T. Psa-gan: Progressive self attention gans for synthetic time series. In *The tenth international conference on learning representations*, 2022.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. Torsional diffusion for molecular conformer generation. *Advances in neural information processing systems*, 35: 24240–24253, 2022.

Joshi, S. Y. and Deshmukh, S. A. A review of advancements in coarse-grained molecular dynamics simulations. *Molecular Simulation*, 47(10-11):786–803, 2021.

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

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

Lewis, S., Hempel, T., Jiménez-Luna, J., Gastegger, M., Xie, Y., Foong, A. Y., Satorras, V. G., Abdin, O., Veeling, B. S., Zaporozhets, I., et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025.

Li, P.-C. and Makarov, D. E. Theoretical studies of the mechanical unfolding of the muscle protein titin: bridging the time-scale gap between simulation and experiment. *The Journal of chemical physics*, 119(17):9260–9268, 2003.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Marszalek, P. E., Lu, H., Li, H., Carrion-Vazquez, M., Oberhauser, A. F., Schulten, K., and Fernandez, J. M. Mechanical unfolding intermediates in titin modules. *Nature*, 402 (6757):100–103, 1999.

McKnight, P. E. and Najab, J. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

Mendell, J. R., Shilling, C., Leslie, N. D., Flanigan, K. M., al Dahhak, R., Gastier-Foster, J., Kneile, K., Dunn, D. M., Duval, B., Aoyagi, A., et al. Evidence-based path to newborn screening for duchenne muscular dystrophy. *Annals of neurology*, 71(3):304–313, 2012.

Neuman, K. C. and Nagy, A. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature methods*, 5(6):491–505, 2008.

Ni, B., Kaplan, D. L., and Buehler, M. J. Forcegen: End-to-end de novo protein generation based on nonlinear mechanical unfolding responses using a language diffusion model. *Science Advances*, 10(6):eadl4000, 2024.

Noé, F., Olsson, S., Köhler, J., and Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Schönfelder, J., Perez-Jimenez, R., and Munoz, V. A simple two-state protein unfolds mechanically via multiple heterogeneous pathways at single-molecule resolution. *Nature communications*, 7(1):11777, 2016.

Sengupta, A. and Rief, M. Energy landscapes of fast-folding proteins pushing the limits of atomic force microscope (afm) pulling. *Proceedings of the National Academy of Sciences*, 118(19):e2102946118, 2021.

Sikora, M., Sułkowska, J. I., Witkowski, B. S., and Cieplak, M. Bsdb: the biomolecule stretching database. *Nucleic acids research*, 39(suppl_1):D443–D450, 2010.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Stacklies, W., Vega, M. C., Wilmanns, M., and Gräter, F. Mechanical network in titin immunoglobulin from force distribution analysis. *PLoS computational biology*, 5(3): e1000306, 2009.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Viljoen, A., Mathelié-Guinlet, M., Ray, A., Strohmeyer, N., Oh, Y. J., Hinterdorfer, P., Müller, D. J., Alsteens, D., and Dufrêne, Y. F. Force spectroscopy of single cells using atomic force microscopy. *Nature Reviews Methods Primers*, 1(1):63, 2021.

Waman, V. P., Bordin, N., Lau, A., Kandathil, S., Wells, J., Miller, D., Velankar, S., Jones, D. T., Sillitoe, I., and Orengo, C. Cath v4. 4: major expansion of cath by experimental and predicted structural data. *Nucleic Acids Research*, 53(D1):D348–D355, 2025.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Yang, L., Song, G., and Jernigan, R. L. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences*, 106(30): 12347–12352, 2009.

A. Data and Preprocessing

This section provides detailed descriptions of the datasets used for training and validating *GenUnfold*, including the simulation-derived training set, the experimental benchmark set, and the feature extraction pipeline.

A.1. BSDB Data Source and Preprocessing

The primary training data are sourced from the Bio-molecule Stretching Database (BSDB), which contains Steered Molecular Dynamics (SMD) simulations of protein unfolding under constant-velocity stretching.

Data Composition. We cross-referenced BSDB entries with the Protein Data Bank (PDB) to ensure that high-resolution experimental structures were available for every entry. The final curated dataset comprises 16,495 unique proteins.

Trajectory Details. Since the database typically provides two pulling traces per protein, the total dataset consists of 32,990 force–extension (F–x) trajectories.

Standardization Pipeline. To ensure that the model learns mechanical principles independent of protein size, each trajectory undergoes the following preprocessing steps:

- **Extension Normalization.** The extension axis is rescaled by the theoretical contour length,
$$L_{\mathrm{con}} = N \times 3.6\ \text{Å},$$
where N is the number of residues.
- **Interpolation.** Each trajectory is interpolated to a fixed resolution of $T = 512$ points on the normalized interval $[0, 1]$.
- **Force Normalization.** Force magnitudes are normalized to the unit interval $[0, 1]$ using min–max scaling to stabilize model training based on the maximum unfolding force observed across the training set.

Event Selection. Successful unfolding events are identified by characteristic saw-tooth patterns in the force–extension curves.

A.2. PDB Processing for Local Chemical Features

GenUnfold incorporates a comprehensive residue-level feature matrix $R_f \in \mathbb{R}^{L \times d_f}$ extracted directly from experimental PDB structures to provide the chemical and structural context necessary for predicting mechanical unfolding trajectories.

Extraction Method. The structural coordinates of the protein backbone and side chains are processed using the Dictionary of Protein Secondary Structure (DSSP) algorithm. This allows for the deterministic mapping of three-dimensional atomic arrangements into a set of localized physicochemical and geometric descriptors.

Feature Components. The feature vector for each residue consists of 17 distinct descriptors, categorized as follows:

- **Identity and Physicochemical Properties.** To capture the chemical nature of the polypeptide chain, we include the amino acid identity (`res_name`), net `charge`, and `hydropathy` index. We further include binary indicators for specific chemical classes: `hydrophobic`, `aromatic`, `positive`, and `negative` residues. A specific flag, `pro_or_gly`, is used to denote residues with unique conformational constraints (Proline and Glycine) that significantly impact backbone flexibility.
- **Geometric and Conformational Descriptors.** Local fold topology is represented by the secondary structure assignment (`sec_type`), solvent-accessible surface area (`sasa`), and the backbone torsion angles (`phi`, `psi`). Additionally, the `volume` of each residue is included to account for local packing density and steric effects within the protein core.
- **Dynamics and Interaction Potential.** To account for structural stability and the likelihood of bond rupture, we incorporate the `b_factor` as a proxy for local fluctuations. The `hbond_density` is calculated to quantify the local hydrogen-bonding network. Finally, we include binary markers for a residue's propensity to participate in `salt_bridge` interactions or `pi_stack` formations, both of which serve as critical "mechanical clamps" during forced unfolding.

B. Model Architecture

B.1. Noise Prediction Pipeline: From x_t to ϵ_t

In the GenUnfold framework, the core objective of the neural network ϵ_θ is to approximate the added noise at a given diffusion timestep t. GenUnfold utilizes a dual-stream conditioning strategy to modulate the latent features.

- **Global Modulation (Evolutionary):** The continuous timestep t is transformed into a frequency-based embedding followed by a two-layer MLP to obtain the temporal feature vector $\mathbf{e}_t \in \mathbb{R}^H$ with the hidden size H:

$$\mathbf{e}_t = \text{MLP}(\text{SiLU}(\text{Linear}(\text{Sinusoidal}(t)))) \tag{4}$$

  The evolutionary context y_{ESM} (extracted from pretrained ESM-2) is concatenated with time embedding $\mathbf{e}_t$ to get gobal context y_t and processed to produce scaling and shifting parameters $\{\gamma, \beta, \alpha\}$ for the Adaptive Layer Normalization (adaLN) units:

$$[\gamma, \beta, \alpha, \dots] = \text{Linear}(y_t),\ y_t = \mathbf{e}_t + \text{MLP}(y_{ESM}). \tag{5}$$

- **Local Modulation (Physics-Biased Attention):** The details of calculating local physical representation c are described in the following subsection B.2.

Given the noisy force-extension trajectory $x_t \in \mathbb{R}^{N\times 1}$, the global coevolutionary context y_{ESM}, and physical reprensentation c, the computation of the predicted noise $\hat{\epsilon}_t$ proceeds through the following stages:

1. Input Embedding The normalized noisy trajectory x_t is first projected into a high-dimensional latent space via a linear layer, yielding the initial sequence representation $\mathbf{z}_0 \in \mathbb{R}^{N\times H}$.

2. Diffusion Transformer with adaLN-Zero and cross-attention. Integrating these adaptive parameters γ, β and α, the Transformer block updates the representation $\mathbf{z_0}$ as follows:

$$\mathbf{z}_1 = \mathbf{z}_0 + \alpha_1 \cdot \text{SelfAtten}(\text{adaLN}(\mathbf{z}_0, y_t), \text{adaLN}(\mathbf{z}_0, y_t), \text{adaLN}(\mathbf{z}_0, y_t)) \tag{6}$$

$$\mathbf{z}_2 = \mathbf{z}_1 + \alpha_2 \cdot \text{CrossAtten}(\text{adaLN}(\mathbf{z}_1, y_t), c_t, c_t) \tag{7}$$

$$\mathbf{z}_3 = \mathbf{z}_2 + \alpha_3 \cdot \text{MLP}(\text{adaLN}(\mathbf{z}_2, y_t)) \tag{8}$$

In this formulation, the gating factors α_1 and α_2 are initialized to zero. This "adaLN-Zero" strategy ensures that each residual block initially acts as an identity function, significantly stabilizing the early stages of diffusion training by preventing large gradient variances from uninitialized attention weights.

The adaLN operation is defined as a conditioned linear transformation of the standard layer-normalized features:

$$\text{adaLN}(\mathbf{z}, y_t) = \gamma(y_t) \cdot \frac{x - \mu}{\sigma} + \beta(y_t),\ [\gamma, \beta] = \text{Linear}(\text{SiLU}(\mathbf{y}_t)) \tag{9}$$

where μ and σ are the mean and variance of data z along feature dimension.

For the standard Self-Attention,

$$\text{SelfAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{10}$$

where query Q, key K and value V are from the same input $\mathbf{Z} \in \mathbb{R}^{N\times d}$

$$Q = \mathbf{Z}W^Q, \quad K = \mathbf{Z}W^K, \quad V = \mathbf{Z}W^V. \tag{11}$$

For the standard Cross-Attention,

$$\text{CrossAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{12}$$

where Q is from hidden state $\mathbf{Z} \in \mathbb{R}^{N\times d}$, but K and V are from physical representation $\mathbf{C} \in \mathbb{R}^{M\times d_c}$

$$Q = \mathbf{Z}W^Q, \quad K = \mathbf{C}W^K, \quad V = \mathbf{C}W^V \tag{13}$$

3. Final Projection After the final Transformer block, the refined latent features $\mathbf{z}_L$ undergo a linear projection to map the d-dimensional features back to the trajectory space:

$$\hat{\epsilon}_t = \text{Linear}(z_3) \tag{14}$$

The predicted noise $\hat{\epsilon}_t \in \mathbb{R}^{N\times 1}$ is then used in the reverse diffusion step to update $x_t \to x_{t-1}$.

B.2. Mechanical Representation Extraction Pipeline

Physics-Guided Multi-Channel Priors Inspired by the Elastic Network Model (ENM) (Atilgan et al., 2001), we model the protein as a residue-level spring network. This allows us to derive a set of physical priors that capture distinct aspects of the mechanical landscape from the 3D structure $\mathbf{X}$.

The spring network consists of a set of nodes fully connected to each other by simple harmonic springs to model the interaction among residues. The weighted adjacency matrix $\mathbf{K} \in \mathbb{R}^{L\times L}$ defines the spring constant k_{ij} between residues i and j, which inversely proportional to their squared Euclidean distance, d_{ij} (Yang et al., 2009):

$$k_{ij} = \begin{cases} d_{ij}^{-2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{15}$$

This matrix $\mathbf{K}$ represents the local stiffness of intramolecular interactions (both covalent and non-covalent) between residue pairs. From $\mathbf{K}$, we can compute the Graph Laplacian $\mathbf{L}$ (also known as the Kirchhoff matrix) from the stiffness matrix:

$$\mathbf{L} \;=\; \mathbf{D_{deg}} - \mathbf{K}, \quad (\mathbf{D_{deg}})_{ii} = \sum_j k_{ij}. \tag{16}$$

To quantify the global mechanical coupling across the entire network, we compute the effective resistance R_{ij}. This is derived from the pseudoinverse of the Laplacian, $\mathbf{G} = \mathbf{L}^+$:

$$R_{ij} \;=\; G_{ii} + G_{jj} - 2G_{ij}, \tag{17}$$

Intuitively, R_{ij} quantifies how strongly residues i and j are mechanically coupled within the protein's elastic network. A small R_{ij} value identifies "stiff corridors" that serve as efficient pathways for force transmission.

Learnable Physical Bias and Temporal gating We aggregate the above priors into a multi-channel graph $\mathcal{G} = [\mathbf{G}, \mathbf{K}, \mathbf{R}] \in \mathbb{R}^{3\times L\times L}$, comprising the Euclidean distance map $\mathbf{D}$, stiffness map $\mathbf{K}$, and resistance map $\mathbf{R}$. Rather than directly injecting the graph $\mathcal{G}$, we employ a CNN-based Bias Module to learn optimal mechanical features (see Fig. 3). The graph $\mathcal{G}$ is processed by a 3×3 convolutional layer followed by a ResNet block consisting of two stacked 3×3 convolutions. Specifically, each convolutional layer is followed by Batch Normalization (BN) to stabilize training, with a Rectified Linear Unit (ReLU) activation applied after the first BN. A shortcut connection performs an element-wise addition between the block's input and the output of the second BN, followed by a final ReLU non-linearity to facilitate gradient flow. A final projection maps the fused features to h distinct bias maps, $B_{\text{cnn}} \in \mathbb{R}^{h\times L\times L}$, matching the number of attention heads. This design promotes attentional specialization that one head may learn to focus on local rigidity (dominated by $\mathbf{K}$), while another attends to long-range allosteric pathways (informed by $\mathbf{R}$).

Crucially, the relevance of extracted static features B_{cnn} changes as the diffusion process reconstructs the trajectory (e.g., native contacts are more relevant at low force/early time steps). To model this, we introduce a Time-Dependent Gating mechanism. The diffusion time embedding t_{emb} is projected via an MLP to a scaling vector $\mathbf{g}(t) \in \mathbb{R}^H$. The final attention bias for the i-th head is a dynamically modulated version of the static map:

$$B^{(i)}_{\text{final}}(t) = g^{(i)}(t) \cdot B^{(i)}_{\text{cnn}} \tag{18}$$

Protein Encoder Block The encoder processes the sequence of residue features $\mathbf{R}_f \in \mathbb{R}^{L\times d_f}$. The core operation is Multi-Head Physical Attention as shown in Fig. 3. First, we apply Layer Normalization of the projection of residue tokens $\mathbf{R_f}$, and the normalized tokens $\mathbf{R}_{norm} \in \mathbb{R}^{L\times d_k}$ are used to compute the attention output of each head h:

$$\mathbf{R}_{norm} = \text{LayerNorm}(\text{Linear}(\mathbf{R}_f)) \tag{19}$$

$$\mathrm{Attn}^{(i)} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(i)}(\mathbf{K}^{(i)})^\top}{\sqrt{d_k}} + B^{(i)}_{final}(t)\right)\mathbf{V}^{(i)} \tag{20}$$

where $\mathbf{Q}^{(i)} = R_{norm}\mathbf{W}^{(i)}_Q$, $\mathbf{Q}^{(i)} = R_{norm}\mathbf{W}^{(i)}_Q$, and $\mathbf{Q}^{(i)} = R_{norm}\mathbf{W}^{(i)}_Q$. $\mathbf{W}^{(i)}_Q; \mathbf{W}^{(i)}_K, \mathbf{W}^{(i)}_V \in \mathbb{R}^{d_k \times d_k}$, are query, key, and value weights, respectively; d_k is the head dimension. The outputs of all heads are concatenated and passed through a linear projection $\mathbf{W}_O$, followed by the first residual connection:

$$\mathbf{R}' = \mathrm{Linear}(\mathbf{R}_f) + \mathrm{Concat}(\mathrm{Attn}^{(1)}, \ldots, \mathrm{Attn}^{(H)})\mathbf{W}_O \tag{21}$$

Second, this intermediate representation is passed through a position-wise Feed-Forward Network (FFN), again using pre-normalization, to compute the local representaion c that is fed into the cross-attention layers of the diffusion backbone:

$$c = \mathbf{R}' + \mathrm{FFN}(\mathrm{LayerNorm}(\mathbf{R}')) \tag{22}$$

where $\mathrm{FFN}(\mathbf{x}) = (\mathrm{SiLU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2)$.

Technical Configurations. The denoising backbone consists of 2 processing layers with a hidden dimension of 256 and 4 attention heads, utilizing a patch size of 2 to process the force-extension trajectories. The Physical Encoder employs a Bias-based architecture that fuses multi-channel structural maps (including the distance map D, Laplacian stiffness Γ, and effective resistance R) via a CNN-based Bias Module with a width of 32. Global sequence context is provided by a pre-trained ESM-2 model with an embedding dimension of 1280, while local residue-level context is captured through a 24-dimensional feature matrix. The diffusion process is defined by 1000 training steps using a linear noise schedule, with inference accelerated via DDIM respacing to 100 steps. Training is conducted for 100 epochs with a batch size of 64 using the AdamW optimizer (learning rate 5×10^{-4}, weight decay 1×10^{-5}) and a cosine learning rate scheduler.

C. Evaluation metrics

We evaluate models at three complementary levels: (i) *point-wise reconstruction*, (ii) *distributional fidelity*, and (iii) *biophysical consistency* of derived mechanical properties.

Point-wise reconstruction (Rel l_2) To quantify the precise reconstruction of the unfolding pathway, we compute the Root Mean Squared Error (Rel l_2) between the ground-truth ($\mathbf{F}_{\text{true}}$) and predicted ($\mathbf{F}_{\text{pred}}$) trajectories. Both are sampled on the same normalized extension grid $\{x_m\}_{m=1}^{T}$ (where T=512):

$$Rel\ l_2 \ = \ \sqrt{\frac{1}{T}\sum_{m=1}^{T}\frac{\big(F_{\text{pred}}(x_m) - F_{\text{true}}(x_m)\big)^2}{\big(F_{\text{true}}(x_m)\big)^2}}. \tag{23}$$

We report the mean of per-curve RMSE over the test set.

Distributional fidelity (FID). To assess whether the generated trajectories capture the realistic, high-dimensional statistics of real unfolding curves, we calculate the Fréchet Inception Distance (FID). Denote means and covariances of test curves by $(\mu_{\text{real}}, \Sigma_{\text{real}})$ and $(\mu_{\text{gen}}, \Sigma_{\text{gen}})$. The FID is

$$\mathrm{FID} = \|\mu_{\text{real}} - \mu_{\text{gen}}\|_2^2 + \mathrm{tr}\big(\Sigma_{\text{real}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{real}}\Sigma_{\text{gen}})^{1/2}\big). \tag{24}$$

A lower FID indicates that the generative distribution is statistically closer to the real data manifold.

Biophysics consistency (JSD). Beyond raw signal matching, it is critical that the model reproduces the distributions of physically salient properties. We extract two key properties: unfolding force $F_{\text{unfold}} = \max_x\{F(x)\cdot\mathbf{1}(\text{unfolding event at } x)\}$ (Force-JSD) and unfolding energy $W = \int^{L_{con}} F dL_{ac}$ (Energy-JSD). We fit one-dimensional Gaussian Kernel density estimations (KDEs) of these scalars for both real (P) and generated (Q) samples of $F_{\max}$ and W. The divergence is measured via the Jensen–Shannon distance (JSD):

$$\mathrm{JSD}(P \parallel Q) = \tfrac{1}{2}\mathrm{KL}(P \parallel M) + \tfrac{1}{2}\mathrm{KL}(Q \parallel M)\,, M = \tfrac{1}{2}(P+Q). \tag{25}$$

Low JSD values indicate that the model correctly captures the stochastic ensemble properties of the protein's energy landscape.

D. Benchamark

Given the novelty of generating full force–extension trajectories directly from static structure, no established external benchmarks exist. Therefore, we design a comprehensive suite of internal baselines to isolate the contributions of our architectural components: the coevolutionary priors (ESM-2), the physics-guided bias (PhyAtten), and the co-conditioning strategy (CrossDiT).

We evaluate two classes of diffusion backbones:

1. Global-Conditioned Baselines (Standard DiT). These models utilize a standard DiT conditioned solely via adaLN using a global context vector $y \in \mathbb{R}^H$.

- **DiT-SelfAtten:** A baseline using a lightweight self-attention encoder to aggregate residue features into a global vector.
- **DiT-ESM:** Replaces the encoder with the pretrained ESM-2 model. The global vector y is the mean-pooled ESM-2 embedding, testing the value of pure coevolutionary information.
- **DiT-PhyAtten:** Uses our proposed Physical Encoder (with bias maps $B^{(h)}$) to process the structure, but pools the output into a global vector y. This tests the physics priors in a global-only context.

2. Co-Conditioning Models (CrossDiT). These models employ the full "GenUnfold" architecture, receiving both global guidance (y) via adaLN and fine-grained local guidance ($c \in \mathbb{R}^{L \times H}$) via Cross-Attention.

- **CrossDiT-ESM-SelfAtten:** Combines global ESM-2 context with a standard (unbiased) self-attention local encoder.
- **CrossDiT-ESM-ESM:** A "pure PLM" variant where both global and local contexts are derived from ESM-2 embeddings, ignoring explicit physics priors.
- **CrossDiT-ESM-PhyAtten:** The proposed full model. It combines global ESM-2 context with the local, physics-biased representation c derived from structural priors.
- **CrossDiT-PhyAtten-PhyAtten:** A physics-only variant that uses the Physical Encoder for both global (pooled) and local conditioning, excluding ESM-2 entirely.

All variants share identical hyperparameters for the diffusion schedule, token budget, training epochs, and optimization (details in Appendix). Evaluation metrics (RMSE, FID, JSD) follow the protocols defined in §C.

Table 3. Summary of model variants. The **Global Context** modulates the noise prediction via adaLN-Zero, while the **Local Context** is injected via Cross-Attention layers. "—" indicates the module is absent.

Model Name	**Backbone**	**Global Context** (y)	**Local Context** (c)
DiT-SelfAtten	DiT	SelfAtten	—
DiT-ESM	DiT	ESM	—
DiT-PhyAtten	DiT	PhyAtten	—
CrossDiT-ESM-SelfAtten	CrossDiT	ESM	SelfAtten
CrossDiT-ESM-ESM-2	CrossDiT	ESM	ESM
CrossDiT-ESM-PhyAtten	CrossDiT	ESM	PhyAtten
CrossDiT-PhyAtten-PhyAtten	CrossDiT	PhyAtten	PhyAtten