Spectral Diffusion for Protein Dynamics

Hew Phipps [1] **Matteo Cagiada** [1 2] **Santiago D. Villalba** [3] **Charlotte M. Deane** [1]

Abstract

Generative models present a promising alternative to expensive molecular dynamics for computationally querying protein dynamics, yet many existing approaches treat ensembles as unordered snapshots rather than temporally coherent trajectories, or scale poorly with protein size. We present a new physics-informed representation using Fourier transforms as an inductive bias for the multiscale temporal nature of protein dynamics. Diffusion in the spectral domain allows for disentangling of dynamics into slow conformational modes and fast atomic jitter, enabling rapid and improved prediction of dynamics across a range of temperatures. This is facilitated by denoising of structure and temperature conditioned spectral volumes where the low frequencies directly encode per-residue flexibility. Trained on the mdCATH dataset, we evaluate our model, DynaMode, on a held-out test set achieving strong performance across a set of ensemble-based metrics including a Root Mean Squared Fluctuation (RMSF) pearson r of 0.844. Code is available at https://github.com/HPuntu/DynaMode.

1. Introduction

Proteins can dynamically adopt diverse conformational states that are often difficult to capture at high resolution with experimental methods. Where experimental approaches are limited and difficult to interpret, computationally modeling protein dynamics can offer insights into key biological processes such as folding, binding and allostery (Huynh et al., 2025; Noé et al., 2019). Traditionally, Molecular Dynamics (MD) numerically solves the Newtonian mechanics of a protein structure over time but is computationally expensive. Enhanced sampling methods have been used to accelerate conformational space exploration but their reliance on system specific collective variables makes them less general (Hénin et al., 2022; Zhu et al., 2026). Other modeling approaches, such as Gõ-based Ising-like models, use state-based discretisations based on binary native-like contact formation, but these are often poor approximations of the underlying N-body chain mechanics (Takada, 2019; Jiang & Hansmann, 2012).

With the success of generative models in structure prediction from sequence (Jumper et al., 2021), methods like diffusion and flow-matching offer a cheap alternative to solving dynamics numerically by learning to instead predict MD trajectories directly from structure using deep neural networks (Lewis et al., 2025; Janson et al., 2025). Such "MD emulators" hold potential for a range of possible applications across temporal upsampling, interpolation between conformations, and inpainting (Jing et al., 2024b). However, dynamics generation requires sampling long, temporally consistent trajectories where motion is highly autocorrelated and influenced by environmental conditions such as temperature, making it a distinct challenge from structure prediction.

We sought to address this problem through a change of representation from the time to the spectral domain which has shown success in generative image dynamics (Li et al., 2024). The spectral domain provides a powerful inductive bias for temporal dynamics by disentangling multiscale temporal correlations and coupled motions into an orthogonal frequency basis, yielding a better conditioned learning problem (Ngueabou & Oloniiju, 2025). Additionally, extending this approach to proteins offers unique opportunities for direct per-residue flexibility prediction.

In this work, we present DynaMode, a diffusion model trained on mdCATH (Mirarchi et al., 2024) to generate Discrete Cosine Transform (DCT) spectral volumes for monomers under 576 residues at temperatures between 300K to 450K. DynaMode is a general dynamics generator that captures protein motion across temperatures over a set of ensemble metrics whilst maintaining temporal coherence including on out-of-distribution temperature regimes.

[1]Department of Statistics, University of Oxford, Oxford, United Kingdom [2]Department of Biology, Copenhagen University [3]Bayer, Berlin, Germany. Correspondence to: Charlotte Deane <charlotte.deane@stats.ox.ac.uk>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

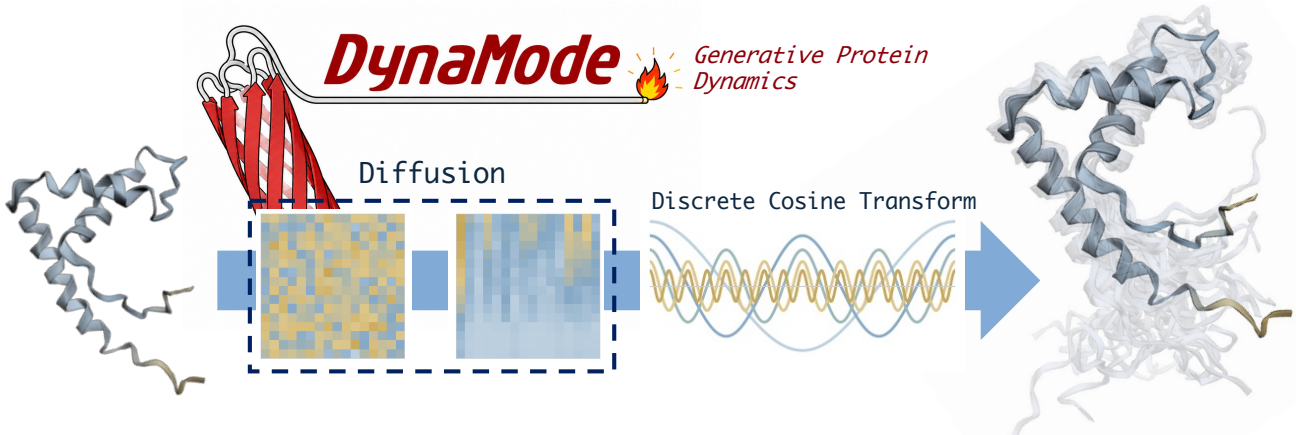

Figure 1. DynaMode is a diffusion model that iteratively denoises a spectral volume representation of protein dynamics given an input structure and temperature. The predicted spectral volume is inverse DCT-II transformed into a trajectory of structures over time.

1.1. Contributions

1. **Fast and Accurate MD Emulation** We achieve superior performance on key metrics for protein dynamics and ensemble properties including an RMSF Pearson of $r = 0.844$ on the mdCATH test set and $r = 0.734$ on the out-of-distribution ATLAS dataset with sampling times of ~ 1 seconds per 250 frame trajectory on a GH200 GPU.
2. **Spectral Convolution Architecture** We develop a custom spectral convolution architecture inspired by Fourier Neural Operators (FNOs) that enables this rapid sampling speed through block-wise spectral mixing.
3. **Per-Residue Flexibility Prediction** We show through Parseval's theorem how the low frequencies analytically approximate RMSF whilst being more expressive measures of residue motion. Through x_0 prediction diffusion denoising with an MSE loss the model also functions as a zero-shot RMSF-like per-residue flexibility predictor for a given structure.
4. **Spectral Protein Dynamics** We show that the DCT transformation is more robust to discontinuity boundaries in high temperature non-equilibrium protein dynamics than the Discrete Fourier Transform (DFT).

Although our representation provides a powerful inductive bias for dynamics, we note that without a post-inference energy minimisation structural validity suffers.

2. Background

Diffusion for Proteins Diffusion models provide a general framework for generative models by learning to iteratively reverse a Gaussian corruption process. Let $\mathbf{x}_0 \in \mathbb{R}^d$ denote a generic protein object, such as cartesian C_α coordinates, and let $\mathbf{c}$ denote conditioning information such as sequence, structural context, or simulation conditions. In the DDPM formulation (Ho et al., 2020) noisy samples are defined as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$, inducing

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. We adopt the cosine schedule of Improved DDPM (Nichol & Dhariwal, 2021). Rather than predicting the added noise, a denoiser directly predicts the clean object from the noised input,

$$\hat{\mathbf{x}}_0 = D_\theta(\mathbf{x}_t, t, \mathbf{c}), \tag{2}$$

This clean prediction defines the score estimate

$$s_\theta(\mathbf{x}_t, t, \mathbf{c}) = \frac{\sqrt{\bar{\alpha}_t}\,D_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{x}_t}{1-\bar{\alpha}_t} \approx \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{c}). \tag{3}$$

At inference, we use the deterministic DDIM update (Song et al., 2022), permitting a reduced number of reverse steps. Training is performed with a denoising regression objective, and inference proceeds by sampling from high-variance Gaussian noise and applying the learned reverse process. Here, we use the same denoising framework for generative modeling of protein dynamics, but jointly over T structures at once.

Ensemble Samplers Prior work has shown that finetuning existing structure prediction models on MD data improves sampling of multiple conformational states across the MD-derived Boltzmann distribution (Jing et al., 2024a; Lewis et al., 2025; Janson et al., 2025). However, these typically rely on large pretrained structure modules or sequence-based embeddings like ESM that can slow inference.

Repeated querying of these models to sample an ensemble results in a set of structures without explicit temporal ordering. Such approaches focus on conformational

state exploration as opposed to dynamics (Janson et al., 2025; Jing et al., 2023; Kapuśniak et al., 2026; Sengar et al., 2025a). The recent MarS-FM explicitly learns metastable states through defining Markov State Models (MSMs) which guide parallel conformational sampling, resulting in drastically improved sampling speed and conformational space traversal (Kapuśniak et al., 2026). Our approach seeks to explicitly model state sampling over time, which we do so efficiently by sampling whole trajectory windows at once.

MD Emulators There has been a recent emergence of so-called MD-emulators which one-shot generate complete trajectories given an input structure, having been effectively demonstrated in small systems over a range of functionalities including upsampling, interpolation and inpainting (Jing et al., 2024b). In our case we are explicitly interested in trajectory generation/extension, which has also been approached by iterative structure sampling as a function of time (Feng et al., 2025; Xu et al., 2025). These methods tackle long-scale dynamics generation for larger systems with curriculum based sampling, where a "forecaster" or "planner" samples sparsely separated conformations across long timescales, followed by an interpolator.

Rather than representing the multiple scales of protein dynamics by training on different timestep partitions of the data (Xu et al., 2025), here we explicitly expose the slow and fast modes through the DCT transform. Recent success in latent space diffusion models evidences the field is moving in the direction of lower dimensional representations for protein dynamics (Sengar et al., 2025a;b). The current state of the art, ATMOS, uses latent states that scale linearly with the number of structures sampled whilst driving temporal sampling for all-atom generative dynamics (Shi et al., 2026).

Low Dimensional Representations of Protein Dynamics MD trajectories are by nature high-dimensional, but proteins exhibit strong spatiotemporal autocorrelations, making it well-established in the biophysics literature to project dynamics onto a small number of collective modes. Methods such as Principal Component Analysis (PCA) and Time-lagged Independent Component Analysis (tICA) identify these slow-mode subspaces directly from simulation data (Schultze & Grubmüller, 2021). Whilst standard tools for MD analysis, their utility for generative modeling is fundamentally limited by their dependence on protein-specific trajectory data to define the basis, precluding generalisation to unseen sequences and motions.

Separately, Normal Mode Analysis (NMA) recovers collective motions analytically by treating the protein as a system of harmonic oscillators around an energy minimum, providing a simulation-free but equilibrium-specific basis (Bauer et al., 2019). Analogously, EigenFold formulates protein structure generation around eigenvectors of the residue graph Laplacian, defining a coarse-to-fine diffusion schedule in which low eigenmodes establish global topology before high eigenmodes resolve local geometry (Jing et al., 2023). Latent diffusion on spatial graph Laplacian modes specifically leverages the multiscale nature of protein geometry (Sengar et al., 2025a).

In contrast, Fourier transformation over the time axis offers a universal basis across different proteins and temperatures that is generative, whilst yielding similar multiscale dynamics separation and slow-mode collective motion encoding into the low frequencies.

Fourier Basis Spectral representations are standard in mechanics and signal processing, where Fourier transformations offer a natural representation for temporal processes by decomposing trajectories into orthogonal modes ordered by frequency (Canuto et al., 2010; Ngueabou & Oloniiju, 2025). Applied along the time axis of protein MD trajectories, the DCT is a real-valued Fourier transform which separates slow, collective conformational changes from faster local fluctuations by assigning them to low- and high-frequency coefficients, respectively (Ahmed et al., 1974). This is consistent with established frequency-domain analyses of MD, including vibrational spectra from Fourier transforms of time-correlation functions and multivariate frequency-domain analysis of protein dynamics (Matsunaga et al., 2009), as well as trajectory-compression and acceleration using essential-dynamics and PCA or DCT representations (Meyer et al., 2006; Kumar et al., 2013; Liang et al., 2026).

DCT is particularly suited to finite trajectory windows where it avoids the strong periodicity assumption of DFT which manifests as undesirable spectral leakage of coefficient correlations across the high frequencies (Wallace, 1992; Yaroslavsky & Wang). Inspired by recent spectral-volume diffusion models for image dynamics (Li et al., 2024), we use DCT spectral volumes as the predictive target for diffusion.

3. Methods

3.1. Data

We trained on the mdCATH dataset of 5,398 monomer domain MD trajectories simulated at 5 temperatures (320K, 348K, 379K, 413K, 450K) with 5 repeats each. Each trajectory consists of up to 450 structures over 1 ns timesteps. We use the standard 80 / 10 / 10 train / val /test split set out in related work (Jing et al., 2024a; Kapuśniak et al., 2026). The test set was curated with mmseqs2 (Steinegger & Söding, 2017) so that no sequence holds $> 20\%$ sequence similarity within the test set, yielding splits with 4304 / 538 / 495 do-

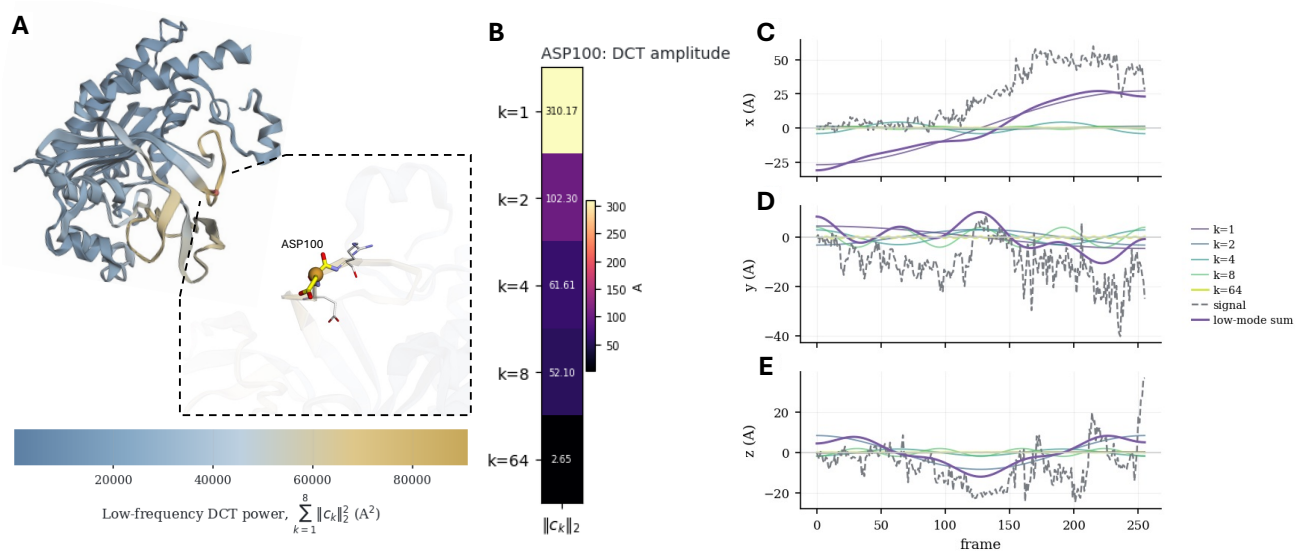

Figure 2. The DCT transform gives τ frequency coefficients for each C_α coordinate channel (x,y,z) of each residue. **A** Illustrative cartoon depiction 12asA00, a 327 residue mdCATH domain coloured by the per-residue spectral power of the lowest $k = 8$ frequencies over a 450K trajectory. Dashed box: Close up of ASP100, the residue with the highest spectral power. **B** DCT spectral amplitudes for a number of different frequencies for ASP100. Each frequency is represented by its cosine wave over time for the x, y and z channels (**C**, **D**, **E** respectively). The purple line is the sum of the shown cosine waves, depicting the nature of the Fourier transform.

mains respectively. We denote the dataset $\mathcal{D}$ of N training trajectories

$$\mathcal{D} = \left\{ \mathbf{X}^{(n)} \in \mathbb{R}^{T_n \times L_n \times 3} \right\}_{n=1}^{N}, \tag{4}$$

where $\mathbf{X}_{t,i}^{(n)} \in \mathbb{R}^3$ denotes the Cartesian coordinates of the i-th residue's C_α atom at time t. For each $\mathbf{X}^{(n)}$ we sample random, contiguous temporal windows and crop random, contiguous residue subsets

$$\mathbf{X}_{s:s+\tau-1,I} \in \mathbb{R}^{\tau \times \ell \times 3}, \quad s \sim U\{0, ..., T_n - \tau\}, \quad I \subset \{1, ..., L_n\}, \ |I| = \ell, \tag{5}$$

with $\tau = 256$ and $\ell = 576$. Sequence-based cropping is done after sampled trajectory windows are aligned using a two-stage iterative rigid-core alignment strategy to the native structure detailed in Algorithm 1 to reduce global rotations and translations whilst still retaining large relative subunit motion (Appendix A.1).

3.2. Spectral Transformations

We denote the input reference structure for any given trajectory $\mathbf{X}_{\text{ref}}^{(n)} \in \mathbb{R}^{L_n \times 3}$ and compute the displacement trajectory for each frame in the sampled window:

$$\Delta\mathbf{X}_{s+t,i,c}^{(n)} = \mathbf{X}_{s+t,i,c}^{(n)} - \mathbf{X}_{\text{ref},i,c}^{(n)} \tag{6}$$

Given a displacement trajectory $\Delta\mathbf{X}_{s:s+\tau-1,I}$, we apply the orthonormal DCT-II transformation along the time dimension. For each residue $i \in I$ and coordinate $c \in \{x, y, z\}$, the transform is given by

$$\mathbf{Z}_{k,i,c}^{(n)} = \alpha_k \sum_{t=0}^{\tau-1} \Delta\mathbf{X}_{s+t,i,c}^{(n)} \cos\left[\frac{\pi}{\tau}\left(t + \frac{1}{2}\right)k\right], \tag{7}$$

$$\alpha_k = \begin{cases} \sqrt{1/\tau}, & k = 0, \\ \sqrt{2/\tau}, & k > 0, \end{cases} \tag{8}$$

where α_k make the transform orthonormal, yielding $k \in \{0, ..., \tau - 1\}$ strictly real-valued frequency coefficients

$$\mathbf{Z}^{(n)} \in \mathbb{R}^{\tau \times \ell \times 3}. \tag{9}$$

The DCT avoids DFT-induced boundary discontinuities by implicitly extending the signal with mirrored symmetry. Under this extension, the endpoint $\mathbf{\Delta}X_{s+\tau-1,i,c}^{(n)}$ connects continuously to its reflection, eliminating the wrap-around discontinuity, making it better suited to non-equilibrium dynamics such as high-temperature unfolding (Yaroslavsky & Wang).

It is standard in the field of electrical engineering to refer to the first (lowest) frequency of the spectral volume $k = 0$ as the Direct Current (DC) component. The DC component represents the per-residue per-channel mean over the trajectory, hence spectral transformation of displacements rather than absolute coordinates means the DC component encodes the mean displacement over the trajectory as opposed to the mean structure.

Going forward, truncation of the spectral volume will refer specifically to setting all frequency coefficients $> k$ to 0.

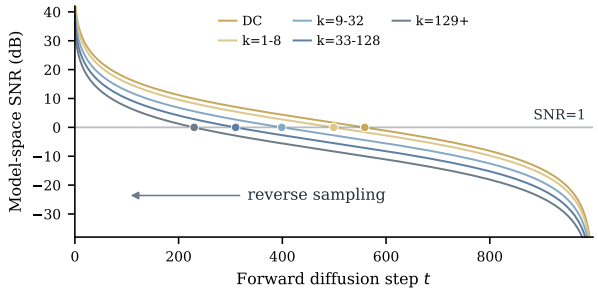

Figure 3. The log-SNR shifted cosine noise schedule enforces hierarchical low to high frequency denoising. Model-space SNR (dB) is the SNR after applying frequency normalisation.

3.3. Diffusion on Spectral Volumes

Schedule and noise. For a clean spectral target $\mathbf{Z}_0$ the forward diffusion process is defined

$$\mathbf{Z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{Z}_0 + \sqrt{1-\bar{\alpha}_t}\,(\mathbf{w}\odot\boldsymbol{\epsilon}), \qquad \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}). \quad (10)$$

Here $\bar{\alpha}_t$ is given by a log-SNR-shifted cosine schedule (Nichol & Dhariwal, 2021). The vector $\mathbf{w}$ contains unit-RMS anisotropic noise multipliers computed from the train-set frequency-scale vectors used for spectral volume normalisation (Appendix A.2), with anisotropy strength $\gamma = 0.5$ (Appendix A.7). This preserves the total noise power of isotropic diffusion while separating the effective log-SNR trajectories across frequency groups (Figure 3). In effect it enforces a hierarchical low to high frequency denoising, low frequencies denoise first setting the global shape and dynamics whilst high frequencies encode atomistic detail akin to how recent diffusion approaches proceed (Chu et al., 2024; Jing et al., 2023).

Objective. All models use $\mathbf{Z}_0$-prediction. As the frequency coefficients are the prediction target the model directly reasons on dynamics. Our training loss is a masked $\mathbf{Z}_0$-MSE plus a curriculum scheduled auxiliary detailed in Appendix A.7:

$$\mathcal{L}_{\text{MSE}} = \frac{\sum_l m_l \,\|\hat{\mathbf{Z}}_{0,l} - \mathbf{Z}_{0,l}\|_2^2}{\sum_l m_l}, \quad (11)$$

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{aux}}\,\mathcal{L}_{\text{aux}}. \quad (12)$$

Model. Our model, DynaMode, is a spectral convolution architecture that combines a frequency-band mixing full-spectrum trunk with a dedicated low-frequency amplitude-calibration branch. The trunk is inspired by Fourier Neural Operators (FNOs) (Li et al., 2021) and the low-frequency specialist is designed to boost the accuracy of the dominant DC / low-k amplitudes that encode shape and flexibility. DynaMode achieves rapid sampling time compared to full self-attention transformers by splitting a full linear frequency mixing operation across the spectral volume into blocks that respect approximate timescale groupings (Table 5). Appendix A.6 gives the full architectural specification.

4. Experiments

4.1. Spectral Truncation Destroys Structural Validity

Spectral transformations are commonly used in video compression where the highest $> k$ frequencies are discarded or quantized, generally encumbering negligible losses in video fidelity (Wallace, 1992). Considering how MD trajectories could be similarly compressed, we first explored how truncating the spectral volume down to the lowest k modes can effect reconstruction quality through a number of structural validity and dynamics metrics. Specifically, for the set of $k = b/K, \quad b \in \{0.0625, 0.125, 0.25, 0.5, 0.75, 1.0\}$ lowest frequency fractions of the full volume[1], we perform both the DFT and DCT transforms on the trajectory for comparison, zero the frequencies $> k$, before inverse transforming to give the reconstructed trajectories. We aggregated the following metrics over the training set:

1. **RMSF Spearman**: Measures per-residue flexibility consistency as a proxy for consistency of dynamics.

2. **Backbone RMSD**: Measures deviation from the original trajectory giving a direct error measurement.

3. **Neighbouring $\mathbf{C}_\alpha$-$\mathbf{C}_\alpha$ Distances**: Physical validity.

We compared DFT and DCT across temperatures and specifically considered trajectory boundaries (first and last 5 frames) at each temperature to explore how DCT improves upon DFT in the non-equilibrium setting (Figure 4). Interestingly, truncation of the spectral volume had negligible effect on RMSF, (mean 0.96 RMSF spearman at $k = 16$ for DCT and $k = 8$ for DFT) evidencing the collection of dynamics in the lowest frequencies (Appendix Table 4).

We expected DCT to be more robust to error, especially at the boundaries as it does not suffer the same Gibbs ringing phenomenem of DFT on non-equilibrium dynamics after truncation of the high frequencies (Gottlieb & Shu, 1997). Indeed, we found that DCT was marginally more accurate than DFT at trajectory boundaries (Figure 4B), motivating its use throughout this work.

4.2. Spectral Amplitude Captures Residue Flexibility

Given the assumption that the low frequencies represent slow collective motion, we next asked how well these low frequency modes represent per-residue flexibility by comparing directly to RMSF. We compute the spectral power for each residue from the lowest k Fourier frequencies f as

[1] We use fractions of the total number of frequencies as the DFT transform has half the number of coefficients as DCT.

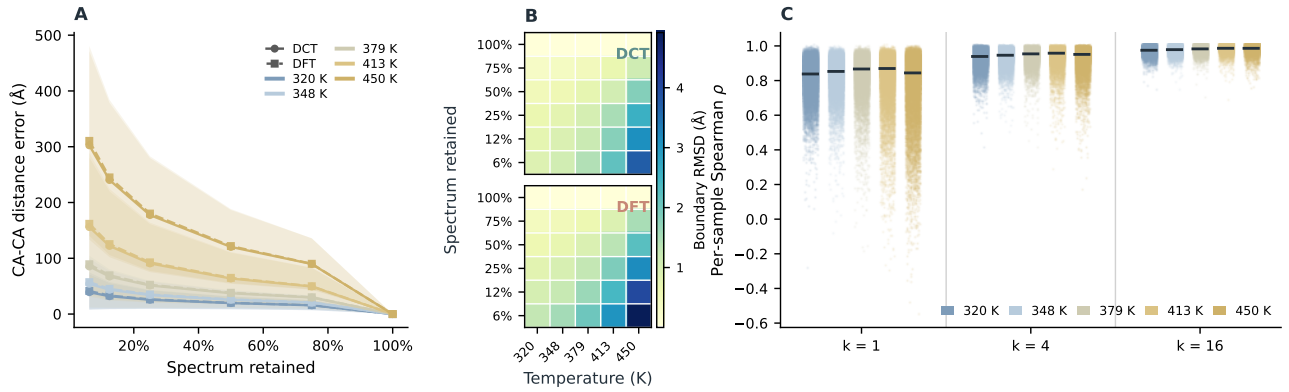

Figure 4. Spectral transformation with DCT followed by truncation leads to significant reconstruction errors. **A** $C_\alpha - C_\alpha$ distances break down with minimal spectral truncation and this scales with simulation temperature. **B** While spectral truncation destroys structural validity, DCT (top) is more robust to this than DFT (bottom) at the trajectory boundaries (first and last 5 frames). **C** A strong positive relationship between RMSF as a measure of flexibility and DCT spectral amplitude of the lowest k frequencies exists when reported by per-sample Spearman correlation (ρ).

the squared ℓ_2-norm of those frequencies

$$\|f_Z(k)\|_2^2 = \sum_{m=1}^{k} \sum_{c\in\{x,y,z\}} |Z_{c,m}|^2. \tag{13}$$

With displacements from native $d_t = x_t - x_{\text{nat}}$, Parseval's identity for the orthonormal DCT states that the per-residue trajectory variance about its mean position $\text{RMSF}_i^2(\mathbf{X})$, equals the non-DC frequencies' summed spectral power,

$$\text{RMSF}_i^2(\mathbf{X}) \;=\; \tfrac{1}{\tau}\sum_{m=1}^{\tau-1}\sum_{c} |Z_{i,c,m}|^2 \;=\; \tfrac{1}{\tau}\,\|f_{Z_i}(\tau{-}1)\|_2^2, \tag{14}$$

so $\frac{1}{\tau}\|f_{Z_i}(k)\|_2^2$ is a truncated RMSF^2 that converges to the full quantity as $k \to \tau{-}1$. Because slow modes carry the bulk of the variance, This is already a good proxy given low k. The DC term is orthogonal to flexibility as $\sum_c |Z_{i,c,0}|^2/\tau = \|\bar{d}_i\|^2$ is the squared mean displacement from native. Thus, the per-residue mean-square distance from native decomposes as $\|\bar{d}_i\|^2 + \text{RMSF}_i^2$.

As the non-DC frequencies encode the temporal scale of motion (slow hinge vs. fast inter-frame jitter) and its direction in $\mathbb{R}^3$, the spectral amplitude is more expressive than RMSF. This decomposition is the primary inductive bias of the spectral target - a model trained to reproduce low-k spectral volumes is by construction reasoning over RMSF.

Analysis of the training set shows a strong positive correlation (per-sample spearman ρ) between per-residue RMSF and DCT amplitude profiles even down to the lowest $k = 4$ non-DC frequencies (Figure 4B). Thus, accurate prediction of the lowest non-DC frequencies is the most important prediction task for learning protein flexibility in this regime. This is our primary motivation for the use of a dedicated

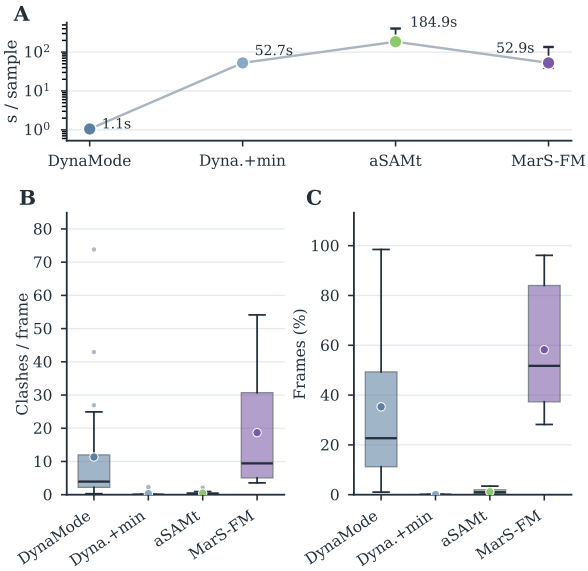

Figure 5. **A** Median inference time per generated 500-frame trajectory on the same GH200 hardware, with error bars showing the interquartile range. Bottom: Structural-validity distributions across targets, reporting nonbonded C_α–C_α clashes below 3.5 Å per frame (**B**) and the percentage of frames containing a nonlocal backbone trace distance below 1.0 Å (**C**).

low-frequency residual correction head in the model (Section A.6) and x_0 prediction over noise or v prediction.

4.3. Spectral Diffusion Effectively Learns Protein Dynamics

We trained our diffusion model, DynaMode, to generate full DCT spectral volumes for $\tau = 256$ frame windows, given an input monomer structure and temperature, which are inverse transformed into 256 temporally ordered structures.

Table 1. Trajectory benchmark on mdCATH (all five temperatures, 320–450 K). MDGen, AlphaFlow-MD and Tempo competitor numbers from published evaluations. ↑: higher is better. ↓: lower is better. **Bold**: best per column among non-oracle methods. Oracle uses the held-out MD trajectory as prediction. Values are medians over 2475 test trajectories.

METHOD	PAIR. RMSD r ↑	GLOBAL RMSF r ↑	RMWD ↓	PCA $\mathcal{W}_2$ ↓	PC-SIM ↑	WEAK J ↑	TRANS. J ↑
ORACLE	0.992	0.885	3.08	2.21	–	0.822	0.482
MDGEN	0.710	0.670	**3.36**	2.62	17.19%	0.410	0.200
ALPHAFLOW-MD	0.410	0.410	5.62	2.38	**21.88%**	0.420	**0.270**
TEMPO	0.770	0.670	4.21	**2.33**	7.81%	0.430	0.200
DYNAMODE	**0.854**	**0.844**	4.12	2.78	17.13%	**0.620**	0.246

Table 2. Trajectory benchmark on ATLAS (300 K). MDGen, AlphaFlow-MD and Tempo competitor numbers from published evaluations. Note that MDGen, Alphaflow and TEMPO each train on a subset of the ATLAS dataset whilst ours is a more difficult out-of-distribution test using the same test set. ↑: higher is better. ↓: lower is better. **Bold**: best per column among non-oracle methods. Oracle uses the held-out MD trajectory as prediction. Values are medians over test targets.

METHOD	PAIR. RMSD r ↑	GLOBAL RMSF r ↑	RMWD ↓	PCA $\mathcal{W}_2$ ↓	PC-SIM ↑	WEAK J ↑	TRANS. J ↑
ORACLE	0.835	0.910	1.85	1.25	–	0.720	0.52
MDGEN	0.480	0.500	2.69	1.89	10%	0.510	**0.410**
ALPHAFLOW-MD	0.480	0.600	2.61	1.52	44%	0.620	0.290
TEMPO	**0.910**	**0.890**	**1.49**	**0.60**	**76%**	**0.740**	0.380
DYNAMODE	0.665	0.734	2.65	1.72	4%	0.491	0.225

Evaluation protocol. We evaluated DynaMode with a comprehensive set of trajectory evaluation benchmarks used in recent generative protein dynamics work to assess both ensemble properties and structural validity. Specifically, we follow the method described by (Jing et al., 2024a) and used in AlphaFlow (Jing et al., 2024a), MDGen (Janson et al., 2025), and TEMPO (Xu et al., 2025) with the same test splits for comparison with their reported results:

1. The held-out mdCATH test set (495 domains across 5 temperatures, 320–450 K).
2. An 82 domain subset of the ATLAS dataset (Vander Meersche et al., 2024) of equilibrium MD simulations at 300 K (100 ns trajectories sliced to 1 ns resolution), serving as an out-of-distribution test at a temperature below the training range.

For a detailed description of the protocol and metrics used we refer the reader to Appendix A.11.

DynaMode is Fast but Suffers from Steric Clashes Given reported inference times are not fairly comparable across systems, and other models do not report structural validity metrics, we ran a small direct comparison against the recent ensemble samplers aSAMt and MarS-FM comparing inference times and structural validity before the full test set benchmark. As aSAMt uses a different train/test split, we used the 6 domain (24 trajectories) union of our mdCATH test sets. For each model we adhered to their respective inference protocols and evaluated structural validity according to the distributions of nonbonded $C_\alpha - C_\alpha$ distances in the generated trajectories (Appendix A.14).

DynaMode is 2 orders of magnitude faster than the next best performing model aSAMt (Figure 5A) on the same GH200 GPU. However, this result highlighted poor structural validity across generated structures in our base approach which, like aSAMt, motivated the use of a brief post-inference energy minimsation (Janson et al., 2025) (detailed in Section A.9) which we found resolved most steric clashes (Figure 5B&C) but dampened our superior inference times (Figure 5A). Although our comparison shows validity superior to MarS-FM and aSAMt after energy minimisation, the size of the dataset used limits the power of the comparison.

Benchmark Results on the mdCATH held-out test set We next report performance across the entire mdCATH test set of 495 domains (Table 1). Due to computational limits and TEMPO not providing code, the reported results of competitor models are their own published values. We also do not perform post-inference energy minimsation but note that the relaxation step has negligible effect on RMSF. aSAMt is excluded as it used a different train/test split to the other models.

DynaMode achieves the strongest performance on the mdCATH test set in pairwise RMSD r and global RMSF r, evidencing the effectiveness of our DCT representation in capturing dynamics (Table 1) even in non-equilibrium settings like unfolding (Figure 6A). Although MarS-FM (Kapuśniak et al., 2026) uses the same splits and benchmark, they report 320K and 450K stratified results on mdCATH only, which we match in Section A.13, showing that they narrowly out-

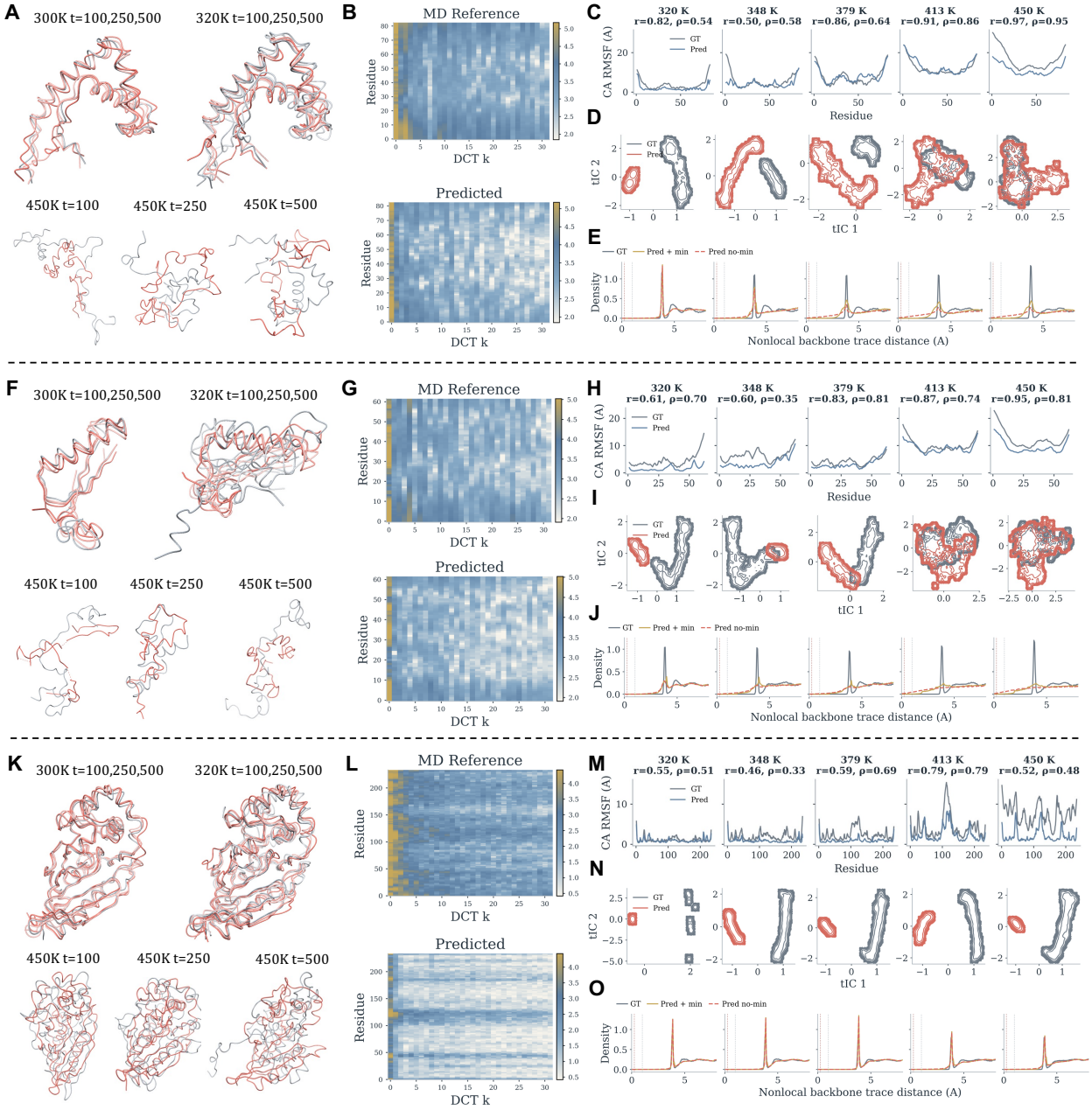

Figure 6. We show three case study domains for DynaMode: 1. (top) mdCATH domain ID 1aabA00 is an 83 residue chain from the train set. 2. (middle) mdCATH domain ID 2fm7A00 is a small 62 residue unfolder from the validation set. 3. (bottom) mdCATH domain ID 4c23B01 from the test set is a 234 residue globular protein. Predicted trajectories were post-inference energy minimised. For each we show structures representing frames 100, 250 and 500 from the predicted (red) and reference MD trajectories (grey) at 300K and 320K overlayed and aligned on each other with decreasing transparency with frame index, and for 450K separately (**A,F,K**). For 300K we show only the native input PDB structure in grey as a reference MD trajectory is not available at 300K for these mdCATH domains. **B,G,L** Heatmaps of the first $k = 32$ lowest per-residue frequency amplitudes (ℓ_2-norm over x,y,z channels) of the MD reference and predicted spectral volumes. For each of the mdCATH set temperatures 320K, 348K, 379K, 413K and 450K RMSF plots with Pearson r and Spearman ρ show the predicted trajectories against the MD reference (**C,H,M**). Similarly tICA free energy overlays at each temperature (**D,I,N**) and non-local backbone trace distance distributions for the MD reference, predicted with and without energy minimisation (**E,J,O**) are shown. tICA folding free energy, RMSF and steric clashes calculations are defined in appendix A.11.

perform us on all metrics at both temperatures.

We present three case study domains, 1aabA00, 2fm7A00 and 4c23B01 where 4c23B01 is an example of a particularly

poor prediction, clearly visible from the difference in spectral volumes (Figure 6L). The significantly lower predicted amplitudes correspond to reduced flexibility and conformational space misalignment (Figure 6N), which aligns with the remarkable consistency in nonlocal backbone trace distance distributions with the the reference MD, particularly compared to 1aabA00 (Figure 6E & J). This highlights the trade off between dynamics and validity, where conservative spectral coefficient predictions (near zero DCT amplitudes) represent model collapse into the input structure, and confident predictions with high amplitudes reproduce dynamics but unless accurately composed fail to maintain validity.

Benchmark Results on the ATLAS dataset Table 2 reports trajectory benchmark results on the ATLAS test set. We emphasize that, whilst the performance is significantly worse on the ATLAS test set, this is an out-of-distribution test set at 300K. MDGen, Alphaflow-MD and TEMPO each train on a subset of the ATLAS dataset.

5. Discussion

DynaMode is capable of producing temporally coherent C_α-only trajectories for single chain monomers under 576 residues at temperatures from 300-450K with a given structure as input. 256 temporally ordered structures are generated in a single diffusion pass in ~ 1 seconds per domain on a GH200 GPU beating existing models substantially (Figure 5), due in part to our spectral convolution architecture (Appendix A.6) which is significantly more lightweight than a full self-attention transformer with heavy sequence embeddings or pretrained structure modules. Further, diffusion in the spectral domain offers a fundamental inductive bias for learning dynamics due to the relationship between DCT frequencies and RMSF.

By projecting the highly autocorrelated time-domain coordinates onto an orthogonal Fourier basis, we globally decorrelate the temporal signal, eliminating the need for deep autoregressive or expansive convolutions to capture long-range macroscopic state transitions. However, this representation circumvents spatial reasoning which, with imperfect spectral volume predictions, collapses structural geometry (Section A.14). In this regard, ensemble-based methods may outperform (Jing et al., 2024a; Lewis et al., 2025; Janson et al., 2025; Kapuśniak et al., 2026). We expect this is mostly due to inadequate prediction of the low-mid frequencies (Figure 6) which define shape and geometry. This posits the potential for a more elegant architectural combination of spatial and spectral reasoning, which we save for future work. Although we resolve steric clashes with post-inference energy minimisation in a small subset test set, it almost negates our inference speed advantage (Figure 5).

DynaMode is limited to C_α and proteins under 576 residues; other existing methods achieve similar accuracies in all atom regimes (Xu et al., 2025; Kapuśniak et al., 2026; Feng et al., 2025) at the cost of increased inference times and more complex curriculum training (Xu et al., 2025; Feng et al., 2025). Although the recent MarS-FM (Kapuśniak et al., 2026) achieves higher accuracy compared to DynaMode on ensemble metrics at 320K and 450K on the mdCATH test set (Appendix A.13), it does not generate temporally contiguous trajectories and requires construction of trajectory-specific MSMs, making it less universal than our DCT approach.

DCT, although more robust to spectral volume truncation than DFT (Figure 4), is still ill-posed for non-equilibrium dynamics due to compensation of non-stationary modes via spectral leakage. Alternative spectral parameterisations such as Chebyshev polynomials (Kondov, 2024; Fain et al., 2002) could be explored for such non-periodic directional data. Although we originally hoped for representation compression by truncating the spectral volume to the low frequency modes only, this was shown to destroy physical validity (Appendix 4.1). However, as is common in video compression, quantisation and/or wavelet transforms of the high frequencies could be explored as a less destructive alternative to truncation (Lee et al., 1997; Bagheri Zadeh et al., 2008).

Alongside general generative dynamics, DynaMode's low-frequency correction head could be engineered into a self-contained queryable prediction module for per-residue flexibility and dynamics. Indeed, we showed that the lowest frequency coefficients of DCT transformed displacements from native coordinates are a more expressive measure of residue motion than RMSF which provides only a single scalar average movement value (Section 4.2). We imagine potential applications in rapid structural flexibility prediction of input structures.

It would also be worth exploiting the ameanability of the spectral domain for temporal upsampling by padding the predicted spectral volumes with 0s which, when inverse transformed, increases the resolved frame counts of the trajectories by a smoothing-like interpolation. We emphasise that given the 1ns frame gaps in the training data, the high frequencies do not represent true atomistic and bond vibration timescales as they would using the transform on the full resolution 10ps per-frame ATLAS dataset. Instead, the high frequencies capture inter-frame differences which are artifactual of MD sampling. Either way, the results herein demonstrate the effectiveness of explicit temporal modelling of protein dynamics through spectral transformations.

Impact Statement

This work advances machine learning for generative biology by introducing a spectral representation for modelling

protein dynamics, together with an architecture designed to accelerate the sampling of molecular dynamics trajectories. By reducing the computational cost of generating dynamical protein ensembles, this work may support downstream applications in biomedical research, protein engineering, and therapeutic development.

The model is trained on publicly available molecular simulation datasets and does not use human-subject data or personally identifiable information. Potential risks include misuse of generative modelling tools for unsupported biological claims, biased conclusions arising from limited simulation coverage, and downstream use in protein design workflows with dual-use implications. These risks are mitigated in part by the coarse-grained Cα trajectory setting, the reliance on established scientific software, and the need for substantial additional validation before any experimental or biomedical application.

6. Code availability

Model training and inference code, including scripts used for evaluation results, is provided through GitHub (https://github.com/HPuntu/DynaMode).

7. Acknowledgements

The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023] (McIntosh-Smith et al., 2024).

The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (Richards, 2015).

This work was supported through research funding from the UK Biotechnology and Biological Sciences Research Council (BBSRC) [Grant number EP/S024093/1] and Bayer AG, in collaboration with Bayer AG.

We are grateful to valuable advice from colleagues in the Oxford Protein Informatics Group (OPIG) including Nele Quast for her helpful comments on first drafts.

8. Competing interests statement

C.M.D. discloses membership of the Scientific Advisory Board of Fusion Antibodies and AI Proteins 632 as well as being a founder of Dalton. All other authors declare no conflict of interest.

References

Ahmed, N., Natarajan, T., and Rao, K. R. Discrete Cosine Transfom. *IEEE Trans. Comput.*, 23(1):90–93, January 1974. ISSN 0018-9340. doi: 10.1109/T-C.1974.223784.

Bagheri Zadeh, P., Buggy, T., and Sheikh Akbari, A. Statistical, DCT and vector quantisation-based video codec. *IET Image Processing*, 2(3):107–115, June 2008. doi: 10.1049/iet-ipr:20070181.

Bauer, J. A., Pavlović, J., and Bauerová-Hlinková, V. Normal Mode Analysis as a Routine Part of a Structural Investigation. *Molecules*, 24(18):3293, September 2019. ISSN 1420-3049. doi: 10.3390/molecules24183293.

Canuto, C., Hussaini, M., Quarteroni, A., and Zang, T. Spectral Methods: Fundamentals in Single Domains. November 2010.

Chu, A. E., Kim, J., Cheng, L., El Nesr, G., Xu, M., Shuai, R. W., and Huang, P.-S. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, July 2024. doi: 10.1073/pnas.2311500121.

Fain, B., Xia, Y., and Levitt, M. Design of an optimal Chebyshev-expanded discrimination function for globular proteins. *Protein Science : A Publication of the Protein Society*, 11(8):2010–2021, August 2002. ISSN 0961-8368. doi: 10.1110/ps.0200702.

Feng, B., Zhang, J., Zhang, X., Liu, Z., and Li, Y. BioMD: All-atom Generative Model for Biomolecular Dynamics Simulation, 2025.

Gottlieb, D. and Shu, C.-W. On the Gibbs Phenomenon and Its Resolution. *SIAM Review*, 39(4):644–668, January 1997. ISSN 0036-1445. doi: 10.1137/S0036144596301390.

Hekkelman, M. L., Salmoral, D. Á., Perrakis, A., and Joosten, R. P. DSSP 4: FAIR annotation of protein secondary structure. *Protein Science : A Publication of the Protein Society*, 34(8):e70208, July 2025. ISSN 0961-8368. doi: 10.1002/pro.70208.

Hénin, J., Lelièvre, T., Shirts, M. R., Valsson, O., and Delemotte, L. Enhanced sampling methods for molecular dynamics simulations. *Living Journal of Computational Molecular Science*, 4(1):1583, December 2022. ISSN 2575-6524. doi: 10.33011/livecoms.4.1.1583.

Ho, J. and Salimans, T. Classifier-Free Diffusion Guidance, July 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International*

Conference on Neural Information Processing Systems, NIPS '20, pp. 6840–6851, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Huynh, N., Kazan, I. C., Lu, J., Kolbaba-Kartchner, B., Mills, J. H., and Ozkan, S. B. A protein dynamics–based deep learning model enhances predictions of fitness and epistasis. *Proceedings of the National Academy of Sciences*, 122(42):e2502444122, October 2025. doi: 10.1073/pnas.2502444122.

Janson, G., Jussupow, A., and Feig, M. Deep generative modeling of temperature-dependent structural ensembles of proteins. *Communications Chemistry*, 8(1): 354, November 2025. ISSN 2399-3669. doi: 10.1038/s42004-025-01737-2.

Jiang, P. and Hansmann, U. H. E. Modeling Structural Flexibility of Proteins with Go-Models. *Journal of Chemical Theory and Computation*, 8(6):2127–2133, June 2012. ISSN 1549-9618. doi: 10.1021/ct3000469.

Jing, B., Erives, E., Pao-Huang, P., Corso, G., Berger, B., and Jaakkola, T. EigenFold: Generative Protein Structure Prediction with Diffusion Models. 2023. doi: 10.48550/ARXIV.2304.02198.

Jing, B., Berger, B., and Jaakkola, T. AlphaFold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 22277–22303, Vienna, Austria, July 2024a. JMLR.org.

Jing, B., Stärk, H., Jaakkola, T., and Berger, B. Generative Modeling of Molecular Dynamics Trajectories, September 2024b.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

Kapuśniak, K., Gabellini, C., Bronstein, M., Tossou, P., and Giovanni, F. D. MarS-FM: Generative Modeling of Molecular Dynamics via Markov State Models, March 2026.

Kondov, I. Polynomial propagators for classical molecular dynamics, July 2024.

Kräutler, V., van Gunsteren, W. F., and Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, 22(5):501–508, 2001. ISSN 1096-987X. doi: 10.1002/1096-987X(20010415)22:5⟨501::AID-JCC1021⟩3.0.CO;2-V.

Kumar, A., Zhu, X., Tu, Y.-C., and Pandit, S. Compression in Molecular Simulation Datasets. In Sun, C., Fang, F., Zhou, Z.-H., Yang, W., and Liu, Z.-Y. (eds.), *Intelligence Science and Big Data Engineering*, pp. 22–29, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-42057-3. doi: 10.1007/978-3-642-42057-3_4.

Lawrence, J., Bernal, J., and Witzgall, C. A Purely Algebraic Justification of the Kabsch-Umeyama Algorithm. *Journal of Research of the National Institute of Standards and Technology*, 124:124028, October 2019. ISSN 2165-7254. doi: 10.6028/jres.124.028.

Lee, M. C., Chan, R. K. W., and Adjeroh, D. A. Quantization of 3D-DCT Coefficients and Scan Order for Video Compression. *Journal of Visual Communication and Image Representation*, 8(4):405–422, December 1997. ISSN 1047-3203. doi: 10.1006/jvci.1997.0365.

Lewis, S., Hempel, T., Jiménez-Luna, J., Gastegger, M., Xie, Y., Foong, A. Y. K., Satorras, V. G., Abdin, O., Veeling, B. S., Zaporozhets, I., Chen, Y., Yang, S., Foster, A. E., Schneuing, A., Nigam, J., Barbero, F., Stimper, V., Campbell, A., Yim, J., Lienen, M., Shi, Y., Zheng, S., Schulz, H., Munir, U., Sordillo, R., Tomioka, R., Clementi, C., and Noé, F. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, July 2025. doi: 10.1126/science.adv9817.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier Neural Operator for Parametric Partial Differential Equations, May 2021.

Li, Z., Tucker, R., Snavely, N., and Holynski, A. Generative Image Dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24142–24153, 2024.

Liang, J., Lu, L., Barnett, A., Greengard, L., and Jiang, S. Accelerating molecular dynamics simulations using fast Ewald summation with prolates. *Nature Communications*, May 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-73232-8.

Matsunaga, Y., Fuchigami, S., and Kidera, A. Multivariate frequency domain analysis of protein dynamics. *The Journal of Chemical Physics*, 130(12):124104, March 2009. ISSN 0021-9606. doi: 10.1063/1.3090812.

McIntosh-Smith, S., Alam, S. R., and Woods, C. Isambard-AI: A leadership class supercomputer optimised specifically for Artificial Intelligence, November 2024.

Meyer, T., Ferrer-Costa, C., Pérez, A., Rueda, M., Bidon-Chanal, A., Luque, F. J., Laughton, C. A., and Orozco, M. Essential Dynamics: A Tool for Efficient Trajectory Compression and Management. *Journal of Chemical Theory and Computation*, 2(2):251–258, March 2006. ISSN 1549-9618. doi: 10.1021/ct050285b.

Mirarchi, A., Giorgino, T., and De Fabritiis, G. mdCATH: A Large-Scale MD Dataset for Data-Driven Computational Biophysics. *Scientific Data*, 11(1):1299, November 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-04140-z.

Ngueabou, Y. V. and Oloniiju, S. D. Integrating Spectral Methods with Neural Network Architectures: A Review of Hybrid Approaches to Solving Differential Equation. *Archives of Computational Methods in Engineering*, December 2025. ISSN 1886-1784. doi: 10.1007/s11831-025-10475-6.

Nichol, A. and Dhariwal, P. Improved Denoising Diffusion Probabilistic Models, February 2021.

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, September 2019. doi: 10.1126/science.aaw1147.

Peebles, W. and Xie, S. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00387.

Richards, A. University of Oxford Advanced Research Computing. August 2015. doi: 10.5281/zenodo.22558.

Schultze, S. and Grubmüller, H. Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. *Journal of Chemical Theory and Computation*, 17(9):5766–5776, September 2021. ISSN 1549-9618. doi: 10.1021/acs.jctc.1c00273.

Sengar, A., Hariri, A., Probst, D., Barth, P., and Vandergheynst, P. Generative Modeling of Full-Atom Protein Conformations using Latent Diffusion on Graph Embeddings, August 2025a.

Sengar, A., Zhang, J., Vandergheynst, P., and Barth, P. Beyond Ensembles: Simulating All-Atom Protein Dynamics in a Learned Latent Space, 2025b.

Shi, L., Lu, J., Liu, J., Shi, C., Yang, Z., and Tang, J. Atomic Trajectory Modeling with State Space Models for Biomolecular Dynamics, March 2026.

Song, J., Meng, C., and Ermon, S. Denoising Diffusion Implicit Models, October 2022.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063.

Takada, S. Gō model revisited. *Biophysics and Physicobiology*, 16:248–255, November 2019. ISSN 2189-4779. doi: 10.2142/biophysico.16.0_248.

Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.-C., and Galochkina, T. ATLAS: Protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1084.

Wallace, G. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, February 1992. ISSN 1558-4127. doi: 10.1109/30.125072.

Xu, Y., Wang, D., Zhou, Z., Yu, T., and Chen, M. TEMPO: Temporal Multi-scale Autoregressive Generation of Protein Conformational Ensembles, 2025.

Yaroslavsky, L. and Wang, Y. DFT, DCT, MDCT, DST AND SIGNAL FOURIER SPECTRUM ANALYSIS.

Zhu, K., Trizio, E., Zhang, J., Hu, R., Jiang, L., Hou, T., and Bonati, L. Enhanced Sampling in the Age of Machine Learning: Algorithms and Applications. *Chemical Reviews*, 126(1):671–713, January 2026. ISSN 0009-2665. doi: 10.1021/acs.chemrev.5c00700.

A. Appendix

A.1. Alignment

All MD trajectories are aligned to the native reference structure using the Kabsch algorithm (Lawrence et al., 2019) which solves for

$$\mathbf{R} = \mathbf{U}\,\mathrm{diag}(1, 1, \det(\mathbf{U}\mathbf{V}^\top))\,\mathbf{V}^\top, \quad \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathrm{SVD}(\mathbf{X}^\top\mathbf{Y}). \tag{15}$$

We use a two a 2-stage iterative rigid-body alignment where we first perform global alignment, before selecting the lowest 50% RMSF residues as the rigid core and re-aligning by computing the rotation matrix on the rigid core and applying it to the globally aligned trajectory as detailed in Algorithm 1.

Algorithm 1 Rigid-Core Trajectory Alignment

Input: Trajectory $\mathcal{T} = \{\mathbf{X}_t\}_{t=1}^T$ where $\mathbf{X}_t \in \mathbb{R}^{N\times 3}$, Native Structure $\mathbf{X}_{nat}$
Output: Aligned Trajectory $\mathcal{T}'$
{Step 1: Global Alignment}
$\mathbf{X}_{nat} \leftarrow \mathbf{X}_{nat} - \mathrm{mean}(\mathbf{X}_{nat})$ {Mean centering}
for $t = 1$ **to** T **do**
  $\mathbf{X}_t \leftarrow \mathbf{X}_t - \mathrm{mean}(\mathbf{X}_t)$
  $\mathbf{R}_t \leftarrow \mathrm{Kabsch}(\mathbf{X}_t^{C_\alpha}, \mathbf{X}_{nat}^{C_\alpha})$ {Compute rotation via SVD}
  $\mathbf{X}_t \leftarrow \mathbf{X}_t\mathbf{R}_t$
end for
{Step 2: Rigid Core Identification}
Compute RMSF for all C_α atoms: $\rho_i = \sqrt{\frac{1}{T}\sum_{t=1}^T \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_i\|^2}$
Let $\mathcal{S}_{core}$ be the indices of C_α atoms where $\rho_i \leq \mathrm{median}(\rho)$
{Step 3: Refined Alignment}
for $t = 1$ **to** T **do**
  $\mathbf{R}'_t \leftarrow \mathrm{Kabsch}(\mathbf{X}_t^{\mathcal{S}_{core}}, \mathbf{X}_{nat}^{\mathcal{S}_{core}})$
  $\mathbf{X}'_t \leftarrow \mathbf{X}_t\mathbf{R}'_t$
end for
return $\mathcal{T}' = \{\mathbf{X}'_t\}_{t=1}^T$

A.2. Conditioned Frequency Normalisation

We compute bucketed frequency normalisation factors for each Cartesian channel c of each frequency coefficient k from the full train set of DCT transformed native-displacement C_α spectral volume z_d. For each train set trajectory we sample enough windows of size $\tau = 256$ to cover the number of frames T typically giving overlapped trajectory windows for each trajectory. For each spectral feature $d = (k, c)$, the reference statistic is the bucket-specific robust amplitude

$$a_d^{(b)} = Q_{0.75}(|z_d| \mid b)\,, \tag{16}$$

where b is a temperature, size (number of residues), or temperature–size bucket. The deployed scale table uses a shrunk temperature–size estimate

$$\sigma_d^{(b)} = 0.25\,\sigma_d^{(\mathrm{global})} + 0.75\,a_d^{(b)}. \tag{17}$$

This shrinkage retains the dominant dependence on temperature and protein size while avoiding a fully independent scale for every bucket and feature.

Figure 7 shows the statistical landscape of the train-set DCT spectral volumes. The global DCT amplitude falls steeply from the DC mode and then flattens, while both temperature and size mainly perturb the DC and low-frequency scales. Temperature partitioning has the most pronounced effect on frequency statistics. High-temperature trajectories require much larger low-k scales, consistent with unfolding drift and slow collective motion.

To quantify the normalisation effect, we compare the observed bucket statistic $a_d^{(b)}$ against the scale selected by each scheme

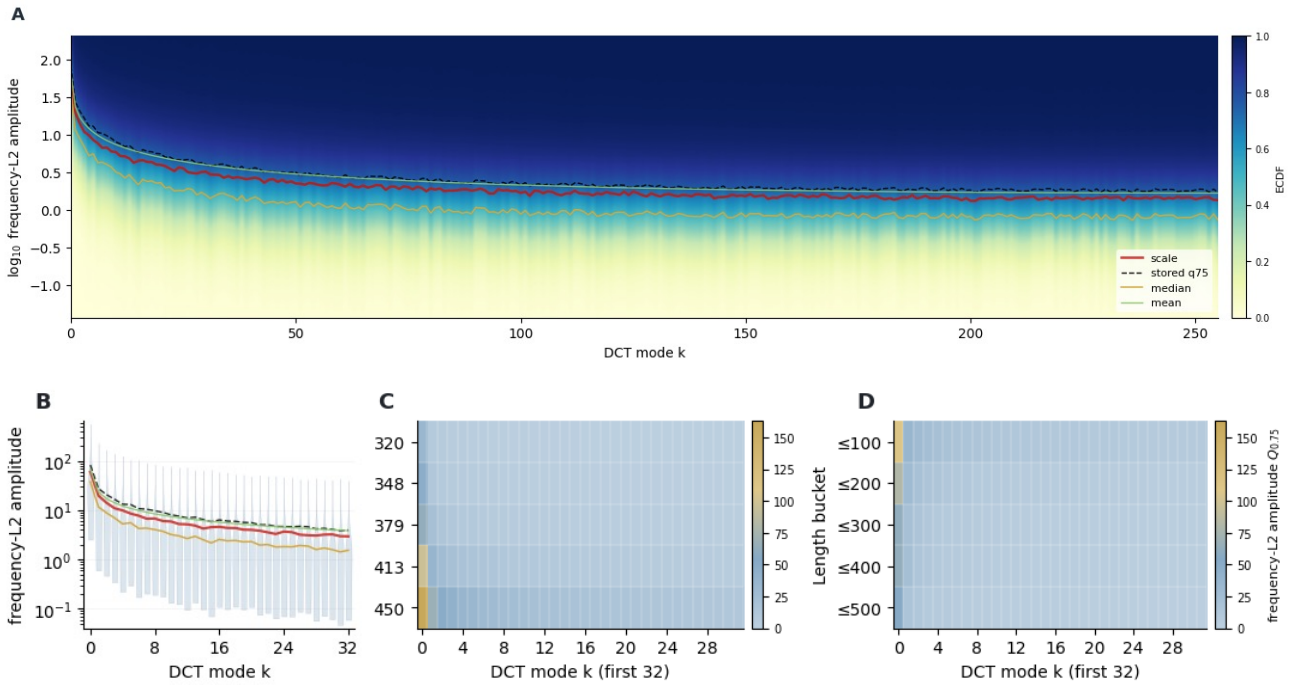

Figure 7. Conditioned DCT frequency-scale statistics used for spectral normalisation. **A** All-frequency ECDF heatmap over all $K = 256$ DCT modes, with the selected normalisation scale (red line) and stored $Q_{0.75}$ amplitude (dashed) overlaid. **B** Low-frequency log-amplitude distribution summary shown as one vertical violin per mode, again with same scale overlaid. **C** Temperature-conditioned $Q_{0.75}$ amplitudes for the first 32 modes. **D** Size-conditioned $Q_{0.75}$ amplitudes for the first 32 modes.

Table 3. Normalisation-scale mismatch across train-set temperature/size buckets. Values are mean absolute log-ratios between observed bucket $Q_{0.75}(|z|)$ and the scale used by each scheme - lower is better. Bands follow the $K = 256$ DCT frequency grouping used for the normalisation analysis, with the DC mode separated from the remaining drift modes.

Scheme	All-k	DC	$0 < k < 8$	$8 \leq k < 32$	$32 \leq k < 128$	$128 \leq k < 256$
Global	0.531	0.524	0.751	0.715	0.562	0.461
Temperature only	0.263	0.224	0.335	0.343	0.279	0.233
Size only	0.490	0.489	0.701	0.652	0.516	0.429
Temperature×size shrunk	0.137	0.147	0.220	0.201	0.145	0.114

using mean absolute log-ratio,

$$\Delta = \frac{1}{|\mathcal{B}|\,|\mathcal{D}|} \sum_{b \in \mathcal{B}} \sum_{d \in \mathcal{D}} \left| \log \frac{a_d^{(b)}}{\sigma_d^{(b)}} \right|, \tag{18}$$

averaged over the 25 observed temperature/size training buckets. Lower values mean the normalised coefficient amplitudes are closer to a common robust scale across regimes. Table 3 shows that temperature conditioning explains most of the improvement over a single global table, while the final shrunk temperature–size table gives the best match, especially for the DC and $0 < k < 8$ drift modes.

Since these scales remove some physically meaningful amplitude variation, the selected low-frequency scales are also passed back to the model through the global scale-conditioning vector described in Section A.5. This enables us to preserve some awareness of the magnitude of differences across frequencies which is likely a strong conditioning factor to predict spectral volumes across temperatures and sizes.

A.3. Residualised DC target.

We additionally stabilise the most dominant spectral component, the DC term, by subtracting a train-set baseline before diffusion. The DC coefficient corresponds to the mean displacement over the trajectory window, so it carries large systematic

offsets associated with temperature, size, and overall structural drift. Asking the model to predict this raw baseline directly wastes capacity on a largely predictable offset. From the train-set we compute a per-protein per-temperature per-residue baseline mean DC component over repeats,

$$\boldsymbol{\mu}_{\mathrm{DC}}^{(n,\mathrm{temp})} \in \mathbb{R}^{L\times C}, \tag{19}$$

Before the forward diffusion process we residualise the DC coefficient of the clean spectral volume so that

$$\mathbf{Z}_{l,0,:}^{\mathrm{res}} = \mathbf{Z}_{l,0,:} - \boldsymbol{\mu}_{\mathrm{DC},l,:}^{(n,\mathrm{temp})}. \tag{20}$$

The model is therefore trained to predict a residualised low-frequency target rather than the absolute DC coefficient. After denoising, the stored DC baseline is added back before inverse transformation to the time domain. This allows the model to focus on learning residual fluctuation around the expected low-frequency drift, rather than the mean drift itself.

A.4. Tokenisation

For a window represented by K retained DCT coefficients and $C = 3$ Cartesian displacement channels, each residue l is encoded as a flattened spectral token

$$\tilde{\mathbf{Z}}_l^{(n)} \in \mathbb{R}^{K\times C}, \tag{21}$$

where $\tilde{\mathbf{Z}}$ denotes the normalised spectral volume. In practice we use $K = \tau$ but this formulation permits spectral volume truncation by choosing $K < \tau$. A per-frequency linear map is applied independently to each (k, c) slice, projecting into the hidden dimension H. The resulting tensor is flattened to $D = KH$. We also add a learned frequency embedding before flattening so the model can distinguish DCT modes explicitly. Rotary positional embeddings (RoPE) (Su et al., 2024) are then applied along the residue axis.

A.5. Conditioning

Native-structure conditioning enters locally at the token input. For each residue and frequency mode we concatenate the normalised spectral coefficient, the native C_α coordinates (scaled by a constant factor), and residue-level features including amino-acid identity and secondary structure before the per-frequency input projection.

Global Conditioning. Global conditioning is injected through AdaLN-Zero (Peebles & Xie, 2023). The conditioning vector

$$\mathbf{c} = f_{t_d}(t_d) + f_{\text{temp}}(\text{temp}) + f_s(s) + f_{\text{size}}(L_{\text{eff}}) + f_{\text{seq}}(\mathbf{a}) + f_{\text{ss}}(\mathbf{q}) + f_{\text{scale}}(\boldsymbol{\sigma}) \tag{22}$$

combines diffusion time t_d, temperature (temp), normalised window position s, effective sequence length L_{eff}, pooled sequence features $\mathbf{a}$, pooled secondary-structure features $\mathbf{q}$, and the selected normalisation-scale features $\boldsymbol{\sigma}$. We use $d_{\text{cond}} \ll D$ to decouple conditioning capacity from token width. All AdaLN modulation layers and the final per-frequency output projection are zero-initialised.

Sequence Embedding. Amino acid sequence is one-hot encoded and embedded in the model through a single linear layer. We also precompute secondary structure using dssp (Hekkelman et al., 2025) and embed that through another linear layer. Both feature types are used in two ways. First, they are injected locally by concatenating the per-residue embeddings to the spectral token input before the per-frequency projection. Second, they are pooled across valid residues with a masked mean and added to the global conditioning vector $\mathbf{c}$.

Size conditioning. Protein size is discretised into coarse bins matching the those used for the normalisation factors.

$$L_{\text{eff}} \in \{\leq 100, \leq 200, \leq 300, \leq 400, \leq 500, \leq 600, > 600\}, \tag{23}$$

computed from the masked residue count in the current crop. A learned embedding of this size bin is projected into the global conditioning vector.

A.6. Architecture

Spectral Convolution. The main body of DynaMode is a spectral-convolution diffusion trunk that predicts the full K-mode output for every residue. Each block contains (i) masked self-attention across residues and (ii) an FNO-style

spectral operator acting along the frequency axis. For (ii), tokens are reshaped to $(B \times L, H, K)$, treating frequency as a one-dimensional domain, and three operations are applied in parallel:

1. **SpectralConv1d**: a learned convolution (H, H, K_{modes}) that mixes the lowest K_{modes} frequencies.
2. **Pointwise Conv1d**: a 1×1 pointwise convolution over the coordinate channels.
3. **Cross-frequency mixing**: a frequency mixer operating independently within physically motivated frequency groups.

The outputs are summed, passed through a GELU nonlinearity, and reshaped back to (B, L, D). The cross-frequency mixer is *block diagonal* rather than fully dense. For $K = 256$ we partition the frequency bands as follows

$$[0, 8],\ [8, 32],\ [32, 128],\ [128, 256], \tag{24}$$

corresponding roughly to drift/slow collective motion, intermediate backbone-scale modes, and high-frequency thermal jitter. This keeps mixing within physically similar spectral regimes and reduces destructive leakage between slow and fast modes.

Low-frequency amplitude calibration head. Given the importance of the low frequencies for dynamics representation we boosted the model's capacity to predict them with an amplitude-calibration branch acting as a residual correction on top of the stronger main trunk for a small set of the lowest frequencies. In practice performance was better with a narrower band of low frequencies (< 8).

Let M denote the number of target low-frequency modes to recalibrate and $M_{\text{ctx}} \geq M$ the number of low-k context modes exposed to the head. The trunk first predicts the full spectral volume $\hat{\mathbf{Z}}^{\text{trunk}}_{l,k,:}$. We then extract the context tensor

$$\mathbf{u}_l = \text{vec}\Big(\hat{\mathbf{Z}}^{\text{trunk}}_{l,\,0:M_{\text{ctx}}-1,:}\Big), \tag{25}$$

concatenate it with residue-local conditioning features, and process it with a small residue-level transformer to produce per-mode log-gains

$$\mathbf{g}_l = g_{\text{amp}}\big(\mathbf{u}_l, \mathbf{c}, \mathbf{e}^{\text{seq}}_l, \mathbf{e}^{\text{ss}}_l, \mathbf{x}^{\text{ref}}_l\big) \in \mathbb{R}^M. \tag{26}$$

The gains are applied multiplicatively to the amplitudes of the first M trunk-predicted vectors while preserving their direction:

$$\hat{\mathbf{z}}_{l,k} = \hat{\mathbf{Z}}^{\text{trunk}}_{l,k,:}, \qquad a_{l,k} = \|\hat{\mathbf{z}}_{l,k}\|_2, \qquad \mathbf{d}_{l,k} = \hat{\mathbf{z}}_{l,k}/\max(a_{l,k}, \epsilon). \tag{27}$$

The calibrated low-frequency prediction becomes

$$\hat{\mathbf{Z}}^{\text{low}}_{l,k,:} = \exp(g_{l,k})\, a_{l,k}\, \mathbf{d}_{l,k}, \qquad k < M, \tag{28}$$

while $\hat{\mathbf{Z}}^{\text{low}}_{l,k,:} = \hat{\mathbf{Z}}^{\text{trunk}}_{l,k,:}$ for $k \geq M$.

Post-DCT geometry refiner and differentiable SHAKE. To account for observed structural collapse and steric clashes we augment the spectral predictor with an explicit coordinate-space geometry module. After denoising, de-normalisation, DC reconstruction, and inverse DCT, the predicted representation is mapped back to absolute C_α coordinates $\hat{\mathbf{X}} \in \mathbb{R}^{T\times L\times 3}$. A lightweight residual refiner r_θ then acts directly on these reconstructed coordinates,

$$\mathbf{X}^{\text{ref}} = \hat{\mathbf{X}} + \Delta_\theta(\hat{\mathbf{X}}), \qquad \|\Delta_{\theta,t,i,:}\|_\infty \leq \Delta_{\max}, \tag{29}$$

where Δ_θ is a small 1-D convolutional stack over the residue axis applied independently to each frame. The final projection layer is zero-initialised, so the auxiliary learns only a bounded post-iDCT residual correction. We used a refiner with hidden width 32, depth 2, kernel size 5, and $\Delta_{\max} = 0.5\,\text{Å}$.

Because the DCT inverse is coordinate-wise and does not enforce chain geometry, we follow the refiner with a differentiable SHAKE-style C_α–C_α projection (Kräutler et al., 2001). For every adjacent residue pair at every frame, define

$$\mathbf{b}_{t,i} = \mathbf{X}^{\text{ref}}_{t,i+1,:} - \mathbf{X}^{\text{ref}}_{t,i,:}, \qquad r_{t,i} = \max(\|\mathbf{b}_{t,i}\|_2, \epsilon), \tag{30}$$

and apply the symmetric update

$$\boldsymbol{\delta}_{t,i} = \frac{1}{2}\,(d^* - r_{t,i})\,\frac{\mathbf{b}_{t,i}}{r_{t,i}}, \qquad \mathbf{X}_{t,i,:} \leftarrow \mathbf{X}_{t,i,:} - \boldsymbol{\delta}_{t,i}, \quad \mathbf{X}_{t,i+1,:} \leftarrow \mathbf{X}_{t,i+1,:} + \boldsymbol{\delta}_{t,i}, \tag{31}$$

with target distance $d^* = 3.8$ Å. We iterate this projection for at most 2 passes over valid adjacent pairs. Since the projection remains inside the forward graph, training can expose the spectral trunk and coordinate refiner to the geometry-corrected trajectory used for downstream losses. We also penalise the refiner's SHAKE residual during the geometry-loss phase so that the learned refiner, rather than SHAKE alone, absorbs systematic bond-length corrections.

A.7. Training

We train on randomly sampled 256-frame windows translating to a 256-mode DCT. Sequences are cropped to 576 residues. Distributed training across 4 GH200 GPUs allows for a batch size of 200. We trained for 300 epochs with AdamW, using a peak learning rate of 3×10^{-5} taking 24 hours in total. Weight decay is applied in the standard decoupled form with coefficient 0.05 on matrix weights and no decay on bias or normalisation parameters. Optimisation used a OneCycle learning-rate schedule with cosine annealing. As the schedule sets $\texttt{pct_start} = \min(0.1, 5/\texttt{epochs})$, this corresponds to a five-epoch warmup in the final 300-epoch run. We clip the global gradient norm to 1.0, and use bfloat16 mixed precision.

Noise Schedule We sample from 1000 diffusion timesteps each training step using a log-SNR-shifted cosine schedule for the scalar $\bar{\alpha}_t$. Spectral anisotropy is applied in the flattened (frequency,coordinate) feature dimension $d = (k, c)$, using the same train-set frequency-scale vector $\sigma_d^{(b)}$ used for spectral normalisation in bucket b. The unnormalised and unit-RMS noise multipliers are

$$\tilde{w}_d = \left(\frac{\sigma_d^{(b)}}{\min_j \sigma_j^{(b)}} \right)^{\gamma}, \qquad w_d = \frac{\tilde{w}_d}{\sqrt{|\mathcal{F}|^{-1} \sum_{j \in \mathcal{F}} \tilde{w}_j^2}}, \qquad \gamma = 0.5, \tag{32}$$

where $\mathcal{F}$ is the set of flattened spectral features. Since the largest scales occur in the DC and low-k modes, this gives low-frequency trajectory components proportionally stronger corruption in both forward diffusion and the initial sampler noise while preserving unit-RMS total noise power.

Algorithm 2 summarizes the training loop. The implementation supports $\mathbf{Z}_0$-, noise-, and v-prediction targets, but all reported DynaMode models use $\mathbf{Z}_0$-prediction. The clean target is the normalised, optionally DC-residualised DCT spectral volume of the native-frame displacement trajectory. We employ Classifier Free Guidance (CFG) (Ho & Salimans, 2022) dropout on temperature and structural conditioning with 15% probability.

The training objective combines a masked $\mathbf{Z}_0$-MSE loss with auxiliary low frequency spectral amplitude and geometry terms:

$$\mathcal{L}_{\text{MSE}} = \frac{\sum_l m_l \, \|\hat{\tilde{\mathbf{Z}}}_{0,l} - \tilde{\mathbf{Z}}_{0,l}\|_2^2}{\sum_l m_l + \epsilon}, \tag{33}$$

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{amp}} \, \mathcal{L}_{\text{amp}} + \lambda_{\text{geo}} \, \mathcal{L}_{\text{geo}}. \tag{34}$$

$$\mathcal{L}_{\text{amp}} = \frac{\sum_{l,k} m_l \, \text{Huber}_\delta\big(A_{l,k}(\hat{\tilde{\mathbf{Z}}}_0) - A_{l,k}(\tilde{\mathbf{Z}}_0)\big)}{K \sum_l m_l}. \tag{35}$$

The spectral amplitude loss $\mathcal{L}_{\text{amp}}$ matches per-residue, per-frequency magnitudes

$$A_{l,k}(\tilde{\mathbf{Z}}) = \big\|\tilde{\mathbf{Z}}_{l,\,3k:3k+3}\big\|_2, \tag{36}$$

between the predicted clean representation $\hat{\tilde{\mathbf{Z}}}_0$ and ground truth $\tilde{\mathbf{Z}}_0$ using a Huber loss ($\delta = 0.5$). This preserves the full K-dimensional amplitude profile per residue, encouraging the model to capture frequency-dependent motion and helping prevent suppression of low-frequency amplitudes by the $\mathbf{Z}_0$-MSE objective.

Finally we apply exponential decay frequency weights inside the spectral losses so that the low-frequency modes, which dominate the physical motion and are hardest to calibrate, receive stronger supervision.

A.8. Inference

The reference structure $\mathbf{X}_{\text{ref}}^{(n)}$ in the form of C_α coordinates, and temperature, provide the primary input structural conditioning. Generated trajectories are implicitly aligned to this input structure as training trajectories were aligned to their

Algorithm 2 Spectral Diffusion Training

Input: Training set $\mathcal{D}$, model f_θ, transform pipeline Φ, diffusion process q, epochs E, retained modes K, window length $T = 256$, optimizer O, scheduler H
Output: Best checkpoint $\theta^\star$
Initialise AdamW with decoupled decay on matrix weights only, OneCycle scheduler, and best validation score $b \leftarrow -\infty$
for epoch $e = 1$ **to** E **do**
  Enable random temporal-window sampling in the training dataset
  for minibatch $\mathcal{B}$ from the training loader **do**
    Load aligned trajectory windows $\mathbf{X}$, native structures $\mathbf{X}_{\mathrm{nat}}$, temperatures T_{temp}, residue masks $\mathbf{m}$, residue features $\mathbf{a}$, DSSP features $\mathbf{q}$, and optional torsion features
    Form coordinate representation $\mathbf{Y} = R(\mathbf{X}, \mathbf{X}_{\mathrm{nat}})$ and concatenate torsion channels when present
    Jitter temperatures during training and compute $\bar{T}_{\text{temp}} = \mathrm{clip}((T_{\text{temp}} - 250)/200, 0, 1)$
    Sample classifier-free conditioning dropout mask $\mathbf{d}$
    $\tilde{\mathbf{Z}}_0 \leftarrow \Phi(\mathbf{Y} \odot \mathbf{m}; K)$ (DCT, truncation, frequency normalisation)
    Residualise the DC coefficient of $\tilde{\mathbf{Z}}_0$ using per-residue or bucket-level baselines when available
    Sample diffusion steps t and noise $\boldsymbol{\epsilon}$; compute $\tilde{\mathbf{Z}}_t = q(\tilde{\mathbf{Z}}_0, t, \boldsymbol{\epsilon})$
    Build residue/channel mask $\mathbf{M}$ and, for hierarchical runs, sample visible and target frequency groups
    $\hat{\mathbf{u}} \leftarrow f_\theta(\tilde{\mathbf{Z}}_t, t, \bar{T}_{\text{temp}}, \mathbf{X}_{\mathrm{nat}}, \mathbf{a}, \mathbf{q}, \mathbf{m}, \mathbf{d})$
    Choose target $\mathbf{u}$: $\tilde{\mathbf{Z}}_0$ for $\mathbf{Z}_0$-prediction, $\boldsymbol{\epsilon}$ for noise-prediction, or v_t for v-prediction
    $\mathcal{L}_{\mathrm{denoise}} \leftarrow \dfrac{\sum \mathbf{M}\, w_k\, w_t\, \|\hat{\mathbf{u}} - \mathbf{u}\|_2^2}{\sum \mathbf{M}\, w_k\, w_t + \epsilon}$
    Recover $\hat{\tilde{\mathbf{Z}}}_0$ from $\hat{\mathbf{u}}$ when auxiliary clean-target losses are active
    Add scheduled auxiliary losses: spectral amplitude, signed low-k, DC, and IDCT-space geometry/refiner/SHAKE losses
    $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{denoise}} + \mathcal{L}_{\mathrm{aux}}$
    if $\mathcal{L}$ or gradients are non-finite **then**
      Skip the optimizer update on all distributed ranks
    else
      Backpropagate $\mathcal{L}$, clip global gradient norm to 1.0, step O, and step H
    end if
  end for
  Disable random validation windows and evaluate one-step loss plus full inference metrics
  Save the latest checkpoint
  $s \leftarrow \mathrm{RMSF\ Spearman} + \mathrm{LDDT}$
  if $s > b$ **then**
    $b \leftarrow s$; save $\theta^\star$ as the best inference checkpoint
  end if
end for
return $\theta^\star$

respective input native structures. Classifier-free guidance (CFG) dropout ($p = 0.15$) is applied to temperature and reference conditioning only. Standard sampling uses 50 denoising steps; the predicted spectral representation $\hat{\tilde{\mathbf{Z}}}_0$ is inverse-DCT transformed and decoded to absolute coordinates over time. For structural refinement these coordinates are then passed through the coordinate refiner and differentiable SHAKE projection described above.

Inference is implemented as a windowed trajectory sampler, summarised in Algorithm 3. Given one or more input PDB structures, we extract C_α coordinates, residue identity features, DSSP features, optional torsion features, and a residue mask. Each requested trajectory is generated as one or more $T = 256$ frame windows. For multi-window trajectories, the sampler can either condition every window on the original native structure or chain windows by conditioning the next window on the final frame of the previous window. The generated windows are concatenated, trimmed to the requested frame count, and optionally passed through a C_α minimisation pass before export.

Given one or more input PDB structures, we extract C_α coordinates, residue identity features, DSSP features, optional torsion features, and a residue batch padding mask. $K = 256$ DCT spectral volumes $\hat{\tilde{\mathbf{Z}}}_0$ are denoised over 50 steps before inverse transforming into a $T = 256$ frame window which is implicitly aligned to the input structure. For structural refinement these coordinates are then passed through the coordinate refiner and differentiable SHAKE projection described above. Inference can be chained by providing say, the final frame of one prediction, as input, and output trajectories stacked for trajectory sampling beyond 256 frames. Optional C_α energy minimsation can be performed.

A.9. C_α Energy Minimisation

As an optional post-processing step, generated trajectories can be relaxed with a lightweight C_α-only energy minimiser. This is not intended to be a full physical force field. Instead, it is a geometry-cleanup objective that reduces nonlocal C_α–C_α clashes while preserving the local shape of the generated trajectory.

For each generated trajectory, coordinates are flattened over frames and optimised in batches of 50 structures. We construct a fixed C_α topology containing adjacent bonds $(i, i+1)$, bend angles $(i, i+1, i+2)$, pseudo-dihedrals $(i, i+1, i+2, i+3)$, and nonbonded C_α–C_α pairs with sequence separation at least two. Nonbonded pairs are cached every 10 optimisation steps by retaining pairs that are within 10 Å in any structure in the current batch.

The minimised energy is

$$
\begin{aligned}
E(\mathbf{X}) = k_b \sum_{(i,j)\in\mathcal{B}} m_{ij} \left(d_{ij} - d^0_{ij}\right)^2 + k_\theta \sum_{(i,j,k)\in\mathcal{A}} m_{ijk} \Delta(\theta_{ijk}, \theta^0_{ijk})^2 \\
+ k_\phi \sum_{(i,j,k,l)\in\mathcal{D}} m_{ijkl} \Delta(\phi_{ijkl}, \phi^0_{ijkl})^2 + k_{\mathrm{nb}} \sum_{(i,j)\in\mathcal{N}} m_{ij} \left[r_{\mathrm{nb}} - d_{ij}\right]_+^2 ,
\end{aligned} \tag{37}
$$

where d_{ij} is a C_α–C_α distance, θ and ϕ are the C_α bend-angle and pseudo-dihedral angle, $\Delta(a, b) = \mathrm{atan2}(\sin(a - b), \cos(a - b))$ handles angle periodicity, and m masks invalid residues. Angle and dihedral targets are set from the generated input coordinates, so minimisation is biased toward preserving the sampled conformation. For adjacent bonds, the target d^0_{ij} is the initial generated bond length if it already lies in $[3.57, 4.11]$ Å, and otherwise the ideal 3.8 Å. In the mdCATH protocol we use $k_b = 10000$, $k_\theta = 100$, $k_\phi = 10$, $k_{\mathrm{nb}} = 250$, and $r_{\mathrm{nb}} = 3.5$ Å. An optional segment–segment self-intersection penalty is implemented but disabled in the reported default settings.

The default protocol first runs a short Adam warm-start stage for up to 50 steps with $k_{\mathrm{nb}} = 100$, then an L-BFGS stage for 30 outer steps with 10 inner iterations per step. Optimisation stops early when the average number of cached nonbonded clashes below 3.5 Å is at most 0.7 per structure. During inference the minimiser can be applied either to each generated window independently or to the concatenated trajectory after all windows have been stacked.

A.10. Reconstruction Error

Table 4 details the DCT and DFT reconstruction errors for different spectral volume truncation across a number of metrics over the whole train set.

Algorithm 3 Windowed Spectral Diffusion Inference

Input: Checkpoint/configuration $\mathcal{C}$, PDB structures $\{\mathbf{P}^{(b)}\}_{b=1}^{B}$, temperatures $\{T_{\text{temp},b}\}_{b=1}^{B}$, requested frames F, window length $T = 256$, retained modes K, ODE steps N, CFG scale w
Output: Generated trajectories $\{\hat{\mathbf{X}}_{1:F}^{(b)}\}_{b=1}^{B}$
Build runtime from $\mathcal{C}$: model f_θ, diffusion sampler S, spectral transform pipeline Φ, representation map R
Load checkpoint weights and set f_θ to evaluation mode
for each input structure $\mathbf{P}^{(b)}$ **do**
  Extract native coordinates $\mathbf{X}_{\text{nat}}^{(b)}$, topology, residue features $\mathbf{a}^{(b)}$, DSSP features $\mathbf{q}^{(b)}$, and mask $\mathbf{m}^{(b)}$
  Optionally compute native torsion features and an NMA RMSF prior from $\mathbf{X}_{\text{nat}}^{(b)}$
end for
Pad all residue-wise tensors in the batch to $L_{\max}$
$J \leftarrow \lceil F/T \rceil$; initialise conditioning structures $\mathbf{C}_1^{(b)} \leftarrow \mathbf{X}_{\text{nat}}^{(b)}$
for $j = 1$ **to** J **do**
  $s_j \leftarrow ((j-1)T)/\max(F-1, 1)$ {Normalised window-position conditioning}
  $D \leftarrow K \cdot C_{\text{model}}$ for DCT spectral models
  Sample initial latent $\mathbf{z}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{aniso}})$ with invalid residues and channels masked out
  Normalise temperatures $\bar{T}_{\text{temp},b} \leftarrow \text{clip}((\tau_b - 250)/200, 0, 1)$
  Wrap f_θ with classifier-free guidance scale w
  $\hat{\tilde{\mathbf{Z}}}_{0,j} \leftarrow S_{\text{ODE}}(f_\theta, \mathbf{z}_N, \mathbf{C}_j, \bar{T}_{\text{temp}}, s_j, \mathbf{a}, \mathbf{q}, \mathbf{m}; N)$
  Restore residualised DC coefficients and denormalise with Φ
  $\hat{\mathbf{Y}}_j \leftarrow \Phi^{-1}(\hat{\tilde{\mathbf{Z}}}_{0,j})$ {Inverse DCT to time-domain representation}
  $\hat{\mathbf{X}}_j \leftarrow R^{-1}(\hat{\mathbf{Y}}_j, \mathbf{C}_j)$ {For displacement models, add the conditioning structure}
  if coordinate refiner or SHAKE projection is enabled **then**
    Refine $\hat{\mathbf{X}}_j$ and project adjacent C_α–C_α distances toward $3.8\,\text{Å}$
  end if
  Append valid residues of $\hat{\mathbf{X}}_j$ to each output trajectory
  if chained generation is enabled **then**
    $\mathbf{C}_{j+1}^{(b)} \leftarrow \hat{\mathbf{X}}_{j,T}^{(b)}$ for all b
  else
    $\mathbf{C}_{j+1}^{(b)} \leftarrow \mathbf{X}_{\text{nat}}^{(b)}$ for all b
  end if
end for
Concatenate windows, trim to F frames, optionally minimise the full trajectory, and export PDB/XTC files
return $\{\hat{\mathbf{X}}_{1:F}^{(b)}\}_{b=1}^{B}$

A.11. Evaluation

Benchmark datasets and trajectory construction. We evaluate on two test sets, the mdCATH held-out test set which uses the same train/val/test split as MarS-FM (Jing et al., 2024a; Kapuśniak et al., 2026), and the 82 domain ATLAS test set. For mdCATH, each domain is simulated with five replicas at each of five temperatures $T_{\text{temp}} \in \{320, 348, 379, 413, 450\}$ K, sampled at $1\,\text{ns}$ per frame. For ATLAS, each domain has three trajectories at $300\,\text{K}$; we subsample each trajectory with stride 100 so that the evaluation resolution matches the $1\,\text{ns}$ timestep used by our model. Following the ensemble evaluation protocol used by MarS-FM (Kapuśniak et al., 2026), we generate multiple trajectories and compare to pooled MD ensembles for each target, followed by averaging over the results.

For mdCATH, one generated sample is formed by predicting two 256-frame windows conditioned at window positions $s = 0$ and $s = 256$, concatenating them, and trimming down to a 500-frame trajectory. For ATLAS, one generated sample is a single 256-frame trimmed to the first 100 frames. We generate 5 independent samples per target on mdCATH and 3 independent samples per target on ATLAS, evaluate each sample against the pooled MD ensemble, and average the resulting per-repeat metrics.

Alignment and oracle baseline. All generated and reference trajectories are rigid-body aligned to the first frame of the first MD replica for that target. We also report an *oracle* MD upper bound via leave-one-out pooling: for mdCATH, one replica is held out and compared against the pool of the remaining four, repeated over all five choices; for ATLAS, one replica is held out and compared against the remaining two. The oracle therefore measures intrinsic replicate-to-replicate agreement under the same metrics, providing a realistic ceiling for what any stochastic emulator can achieve.

RMSF and correlation metrics. For a trajectory $\mathbf{X} \in \mathbb{R}^{T\times L\times 3}$, the per-residue root-mean-square fluctuation (RMSF) is

$$\text{RMSF}_i(\mathbf{X}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{X}_{t,i,:} - \bar{\mathbf{X}}_{i,:}\right\|_2^2}, \qquad \bar{\mathbf{X}}_{i,:} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{t,i,:}. \tag{38}$$

RMSF is a direct measure of residue-wise flexibility, so it is especially important for testing how well our model captures conformational dynamics especially at high temperatures. For each target n, we compute a per-target Pearson correlation

$$r^{(n)}_{\text{RMSF}} = \text{corr}\Big(\text{RMSF}(\hat{\mathbf{X}}^{(n)}), \text{RMSF}(\mathbf{X}^{(n)}_{\text{MD}})\Big), \tag{39}$$

and report the median across targets. We also compute the related per-target Spearman correlation $\rho^{(n)}_{\text{RMSF}}$, which is less sensitive to amplitude calibration and instead emphasizes correct flexibility ranking - we prioritise Pearson as we observed the model struggles more in matching magnitude than residue profiles and this is the metric most other models report (Kapuśniak et al., 2026). Global RMSF correlation pools all residues from all targets into one concatenated vector before computing a single Pearson r or Spearman ρ. Thus, global RMSF tests overall flexibility calibration across the entire benchmark, whereas per-target RMSF tests whether the model gets the residue-level fluctuation profile right for each individual protein.

Table 4. DCT and DFT truncation reconstruction errors for native-frame C_α displacement at matched fractions of the usable spectrum. Values are mean ± standard deviation over the same sampled trajectories used for the reconstruction-error figure. ↑: higher is better; ↓: lower is better.

	SPECTRUM	CA RMSD (Å) ↓	BOUNDARY RMSD (Å) ↓	C_α–C_α W_1 (Å) ↓	RMSF ρ ↑	ENERGY ↑
DCT	6.2%	1.934 ± 1.620	1.841 ± 1.546	0.480	0.976	0.837
DCT	12.5%	1.650 ± 1.371	1.541 ± 1.292	0.374	0.987	0.877
DCT	25.0%	1.353 ± 1.102	1.288 ± 1.064	0.270	0.994	0.914
DCT	50.0%	0.997 ± 0.791	0.918 ± 0.746	0.173	0.998	0.951
DCT	75.0%	0.679 ± 0.534	0.573 ± 0.468	0.123	0.999	0.977
DCT	100.0%	0.000 ± 0.000	0.000 ± 0.000	0.000	1.000	1.000
DFT	6.2%	1.986 ± 1.650	2.543 ± 1.956	0.496	0.975	0.889
DFT	12.5%	1.680 ± 1.389	2.043 ± 1.594	0.383	0.987	0.917
DFT	25.0%	1.371 ± 1.113	1.599 ± 1.250	0.275	0.993	0.943
DFT	50.0%	1.009 ± 0.799	1.148 ± 0.890	0.176	0.998	0.968
DFT	75.0%	0.680 ± 0.533	0.770 ± 0.599	0.124	0.999	0.985
DFT	100.0%	0.000 ± 0.000	0.000 ± 0.000	0.000	1.000	1.000

Pairwise RMSD correlation and distribution. For each target, we estimate the mean pairwise C_α-RMSD within an ensemble,

$$\overline{\mathrm{RMSD}}(\mathbf{X}) = \frac{2}{T(T-1)} \sum_{1 \le t < t' \le T} \mathrm{RMSD}(\mathbf{X}_t, \mathbf{X}_{t'}), \tag{40}$$

and report the Pearson correlation across targets between the generated and MD values of $\overline{\mathrm{RMSD}}$ - evaluating how the model reproduces structural diversity for each protein. In addition, we compare the full distribution of sampled pairwise RMSD values between predicted and MD ensembles using the Jensen–Shannon divergence (JSD). Whereas the correlation metric tests calibration of the target-level summary statistic, the JSD tests whether the overall diversity distribution is reproduced.

Binned distributional metrics. Several scalar observables are compared via histogram-based divergences. Given scalar samples u, we estimate discrete distributions P and Q using 100-bin histograms over a shared range, add a floor $\epsilon = 10^{-5}$, and renormalise. The forward Kullback–Leibler (KL) divergence is

$$D_{\mathrm{KL}}(P \| Q) = \sum_b P_b \log \frac{P_b}{Q_b}, \tag{41}$$

and the Jensen–Shannon divergence is

$$\mathrm{JSD}(P, Q) = \frac{1}{2} D_{\mathrm{KL}}(P \| M) + \frac{1}{2} D_{\mathrm{KL}}(Q \| M), \qquad M = \frac{1}{2}(P + Q). \tag{42}$$

These distributional comparisons are used for pairwise RMSD, RMSF, radius of gyration, fraction of native contacts, Global Distance Test – Total Score (GDT-TS), and spectral amplitudes.

RMWD and PCA-based ensemble distances. To compare full positional ensembles, we approximate the generated and reference positional distributions for atom i by Gaussians $\mathcal{N}(\boldsymbol{\mu}_i^{\mathrm{pred}}, \Sigma_i^{\mathrm{pred}})$ and $\mathcal{N}(\boldsymbol{\mu}_i^{\mathrm{gt}}, \Sigma_i^{\mathrm{gt}})$. The squared 2-Wasserstein distance is

$$\mathcal{W}_2^2(i) = \|\boldsymbol{\mu}_i^{\mathrm{pred}} - \boldsymbol{\mu}_i^{\mathrm{gt}}\|_2^2 + \mathrm{Tr}\left(\Sigma_i^{\mathrm{pred}} + \Sigma_i^{\mathrm{gt}} - 2\left[(\Sigma_i^{\mathrm{gt}})^{1/2} \Sigma_i^{\mathrm{pred}} (\Sigma_i^{\mathrm{gt}})^{1/2} \right]^{1/2} \right). \tag{43}$$

The root mean Wasserstein distance (RMWD) is then

$$\mathrm{RMWD} = \sqrt{\frac{1}{L} \sum_{i=1}^{L} \mathcal{W}_2^2(i)}. \tag{44}$$

RMWD penalises both mean positional shifts and covariance mismatch, so it is more sensitive than RMSF to errors in the full spatial distribution of each residue.

We also compute PCA-based ensemble distances. PCA is fit on the reference MD ensemble, both predicted and reference structures are projected onto the first two reference PCs, projected distances are normalised by $\sqrt{L}$, and an assignment-based empirical 2-Wasserstein distance is computed between equal-sized samples. This asks whether the generated ensemble covers the same dominant collective modes as MD.

tICA free-energy visualisations. For the case-study visualisations in Figure 6, we additionally project trajectories onto slow collective coordinates using tICA. For each domain and temperature, we construct native-displacement C_α features by flattening $\mathbf{X}_t - \mathbf{X}_{\mathrm{nat}}$ over residues and coordinates, and fit a two-component tICA basis jointly on the aligned MD and generated trajectories using a lag time of $\tau = 10$ frames. Both trajectories are projected onto the first two tICs. We then estimate a shared two-dimensional histogram density $p(y_1, y_2)$ and plot the dimensionless relative free-energy surface

$$F(y_1, y_2) = -\log p(y_1, y_2) + C, \tag{45}$$

where C shifts the minimum finite value to zero. These overlays give a visual gauge of how generated trajectories occupy the same slow-mode basins as the MD reference.

It is worth noting, these plots a largely qualitative reporter of ensemble overlap. Smaller proteins tend to show more overlap in our tICA space overlays (results not shown) likely due to the increased dimensionality of the conformational landscape in larger proteins and the use of just the top two components.

Principal-component similarity and contact metrics. Let $\mathbf{v}_1^{\mathrm{pred}}$ and $\mathbf{v}_1^{\mathrm{gt}}$ denote the leading principal components of the generated and reference ensembles. We compute the unsigned cosine similarity

$$s_{\mathrm{PC1}} = \left| \frac{\langle \mathbf{v}_1^{\mathrm{pred}}, \mathbf{v}_1^{\mathrm{gt}} \rangle}{\|\mathbf{v}_1^{\mathrm{pred}}\|_2 \, \|\mathbf{v}_1^{\mathrm{gt}}\|_2} \right|, \tag{46}$$

and summarise the percentage of targets with $s_{\mathrm{PC1}} > 0.5$. This metric tests whether the model recovers the direction of the dominant collective motion.

For contact metrics, we form C_α contact probabilities using an 8 Å threshold. *Weak contacts* are native contacts whose contact probability drops below 0.9, and *transient contacts* are non-native contacts whose contact probability rises above 0.1. We compare the predicted and MD weak/transient contact sets using the Jaccard similarity

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{47}$$

These metrics test whether the model captures the same topological diversity as MD, often more mechanistically meaningful than raw coordinate error.

Native-contact, folding, and structural-validity metrics. From the native structure we extract the set of non-local native C_α contacts and compute the fraction of native contacts retained in each frame using the smooth logistic definition from prior folding-emulation work. This yields a per-frame Fraction of Native Contacts trajectory $\mathrm{FNC}(t)$. We compare predicted and MD FNC distributions using JSD, and also report the mean predicted and reference FNC.

Using a folding threshold $\mathrm{FNC} > 0.5$, we estimate the folded-state probability p_{fold} and corresponding folding free energy

$$\Delta G_{\mathrm{fold}} = -kT \log \frac{p_{\mathrm{fold}}}{1 - p_{\mathrm{fold}}}, \tag{48}$$

where kT uses the evaluation temperature. We report predicted and reference ΔG_{fold}, the absolute folding free-energy error, and the corresponding folded fractions. These quantities test whether the generated ensemble preserves the correct balance between folded and unfolded states.

We additionally compute per-frame GDT-TS to the native structure, yielding both mean GDT-TS and a GDT-TS JSD. This measures native-state structural similarity at multiple distance cutoffs and is useful for checking whether the model preserves native-like geometry while still producing non-trivial motion. Finally, we report mean consecutive C_α distances and their JSD, serving as a lightweight geometric-validity diagnostic analogous to a bond-length sanity check.

Radius of gyration and compactness. For each frame we compute the C_α-based radius of gyration

$$R_g(\mathbf{X}_t) = \sqrt{\frac{1}{L} \sum_{i=1}^{L} \left\| \mathbf{X}_{t,i,:} - \bar{\mathbf{X}}_{t,:} \right\|_2^2}. \tag{49}$$

We compare predicted and MD R_g distributions using both JSD and forward KL, and also report the mean predicted and reference values as a measure of global compactness consistency.

Spectral metrics. For our spectral model, we evaluate the predicted spectral volume directly. Let $\hat{\mathbf{Z}}$ and $\mathbf{Z}$ denote predicted and ground-truth normalised spectral tensors. We compute the spectral mean-squared error

$$\mathrm{SpecMSE} = \frac{1}{LKC} \sum_{l=1}^{L} \sum_{k=0}^{K-1} \sum_{c=1}^{C} \left(\hat{Z}_{l,k,c} - Z_{l,k,c} \right)^2. \tag{50}$$

We also compute the per-residue amplitude at frequency k,

$$A_{l,k}(\mathbf{Z}) = \|\mathbf{Z}_{l,k,:}\|_2, \tag{51}$$

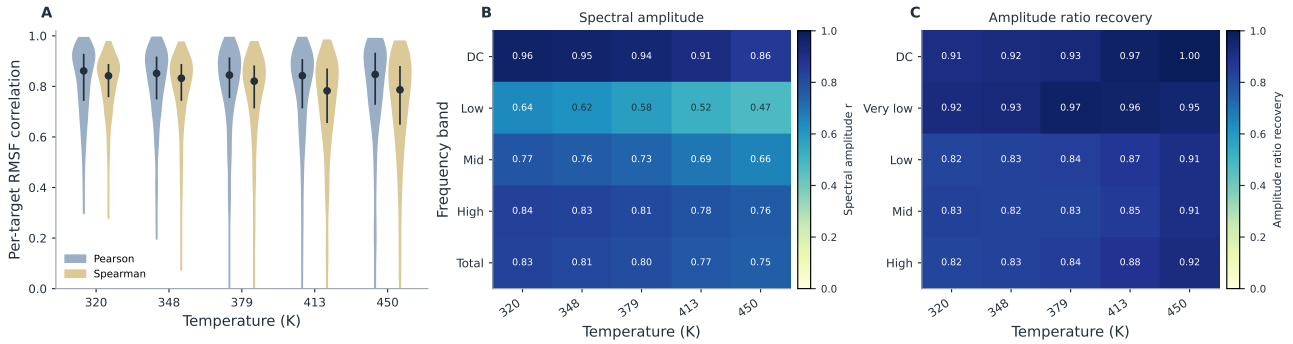

Figure 8. Residue flexibility (RMSF) and spectral volumes prediction accuracy across temperatures on the mdCATH test set. **A** Violin plots show the distribution of RMSF Pearson and Spearman correlations (predicted vs reference MD trajectory following the protocol in Appendix A.11 across temperatures. We also show spectral volume prediction accuracy across different regions of the spectral volume defined by grouping frequencies according to Table 5 at different temperatures. Accuracy is assessed by comparing predicted 256 frame window spectral volumes against same starting timepoint 256 window for MD trajectory references averaged across 5 repeats. **B** Pearson r correlation. **C** Ratio of summed spectral amplitudes of predicted vs reference.

Table 5. Frequency bands used for the frequency-resolved spectral-volume accuracy analysis in Figure 6. Bands are defined over DCT mode index k for 256-frame trajectory windows. The grouping separates the trajectory mean, slow collective drift, low-frequency conformational changes, intermediate backbone-scale motion, and fast local fluctuations. Timescales are approximate because DCT basis functions are finite-window cosine modes rather than periodic Fourier modes.

BAND	DCT MODES k	APPROX. TIMESCALE
DC	0	TRAJECTORY MEAN
VERY LOW	1–4	$\gtrsim 128$ ns
LOW	5–16	32–102 ns
MID	17–64	8–30 ns
HIGH	65–128	$\lesssim 8$ ns

and aggregate amplitudes into DC ($k = 0$), low-, mid-, high-, and total-frequency bands. From these amplitudes we report bandwise Pearson correlations, JSDs, and amplitude-recovery scores. For a reference amplitude vector a^{ref} and predicted amplitude vector a^{pred}, the amplitude-recovery score is

$$\text{Rec}(a^{\text{ref}}, a^{\text{pred}}) = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{|a_i^{\text{pred}} - a_i^{\text{ref}}|}{a_i^{\text{pred}} + a_i^{\text{ref}} + \varepsilon} \right), \tag{52}$$

with $\varepsilon = 10^{-8}$. These diagnostics are especially important because RMSF magnitude is dominated by the DC and neighboring low-frequency modes. For example, the model can achieve reasonable coordinate MSE while still underestimating the low-k amplitudes that control large-scale flexibility. We therefore report DC-specific metrics - DC Pearson correlation, DC JSD, and DC amplitude recovery - alongside broader low/mid/high/total band metrics.

A.12. Spectral Volume Prediction Accuracy

Figure 8B & C shows spectral prediction accuracy measurements across temperatures and frequency groupings (Table 5). The high frequencies are clearly the easiest to predict (Figure 6), which is consistent with the effectiveness of diffusion models in the white-noise regime. In contrast, we observed that the low frequencies, characterised by a highly structured signal, remained the most difficult to predict.

This frequency-resolved error attribution motivated the dedicated low-k amplitude-calibration pathway (Section A.6) which we found improved DC component prediction substantially, as shown by the high correlation in Figure 8B. Directly improving mean displacement prediction like so is not available to time-domain methods, where reconstruction errors are entangled across all temporal scales.

Table 6. Trajectory benchmark on mdCATH at 320 K ($n = 478$ test targets). BioEMU and MarS-FM results from the MarS-FM paper (mean ± SEM). Oracle uses the held-out MD trajectory as prediction. For DynaMode, ± denotes bootstrap SEM across targets. ↑: higher is better; ↓: lower is better. **Bold**: best per row among non-oracle methods.

METRIC	ORACLE (320 K)	BIOEMU	MARS-FM	DYNAMODE (320 K)
PAIR. RMSD r ↑	0.984	$0.580 \pm .001$	$\mathbf{0.900 \pm .001}$	$0.833 \pm .028$
GLOBAL RMSF r ↑	0.894	$0.630 \pm .004$	$\mathbf{0.870 \pm .001}$	$0.823 \pm .017$
PER-TGT RMSF r ↑	0.894	$0.840 \pm .002$	$\mathbf{0.900 \pm .003}$	$0.861 \pm .007$
RMWD ↓	1.85	–	–	$2.73 \pm .102$
PCA $\mathcal{W}_2$ ↓	1.33	–	–	$1.63 \pm .070$
WEAK J ↑	0.780	–	–	$0.538 \pm .004$
TRANS. J ↑	0.508	–	–	$0.258 \pm .003$
R_g JSD ↓	0.09	$0.36 \pm .001$	$\mathbf{0.14 \pm .001}$	$0.15 \pm .004$
ΔG MAE ↓	0.60	$0.83 \pm .003$	$\mathbf{0.58 \pm .001}$	$1.23 \pm .119$

Table 7. Trajectory benchmark on mdCATH at 450 K ($n = 487$ test targets). BioEMU and MarS-FM results from the MarS-FM paper (mean ± SEM). Oracle uses the held-out MD trajectory as prediction. For DynaMode, ± denotes bootstrap SEM across targets. ↑: higher is better; ↓: lower is better. **Bold**: best per row among non-oracle methods.

METRIC	ORACLE (450 K)	BIOEMU	MARS-FM	DYNAMODE (450 K)
PAIR. RMSD r ↑	0.990	$0.250 \pm .002$	$\mathbf{0.650 \pm .004}$	$0.590 \pm .034$
GLOBAL RMSF r ↑	0.862	$0.410 \pm .004$	$\mathbf{0.710 \pm .003}$	$0.691 \pm .022$
PER-TGT RMSF r ↑	0.862	$0.660 \pm .001$	$\mathbf{0.890 \pm .001}$	$0.847 \pm .009$
RMWD ↓	4.82	–	–	$6.38 \pm .100$
PCA $\mathcal{W}_2$ ↓	3.37	–	–	$4.04 \pm .080$
WEAK J ↑	0.930	–	–	$0.715 \pm .003$
TRANS. J ↑	0.511	–	–	$0.191 \pm .004$
R_g JSD ↓	0.07	$0.40 \pm .001$	$\mathbf{0.10 \pm .001}$	$0.12 \pm .003$
ΔG MAE ↓	1.58	$4.67 \pm .004$	$\mathbf{1.05 \pm .003}$	$2.50 \pm .111$

A.13. Temperature Stratified Evaluation Against MarS-FM

For direct comparison against the recent MarS-FM we detail our trajectory benchmark results on the mdCATH test set limited to 320K (Table 6) and 450K (Table 7).

A.14. Structural Validity

Nonbonded and backbone-trace distances. To diagnose steric collapse in C_α-only trajectories, we compute nonbonded C_α–C_α distances for all residue pairs separated by at least two positions in sequence,

$$d_{ij}(t) = \|\mathbf{X}_{t,i} - \mathbf{X}_{t,j}\|_2 , \qquad |i - j| \geq 2. \tag{53}$$

We report the per-frame minimum distance and the number or fraction of frames with at least one pair below 3.5, 3.0, or 2.5 Å. For the nonlocal backbone-trace distance shown in Figure 6, each consecutive C_α pair defines a line segment $S_i(t) = [\mathbf{X}_{t,i}, \mathbf{X}_{t,i+1}]$. For segment pairs with $|i - j| \geq 2$, we compute

$$d_{ij}^{\text{seg}}(t) = \min_{u,v \in [0,1]} \|(1-u)\mathbf{X}_{t,i} + u\mathbf{X}_{t,i+1} - (1-v)\mathbf{X}_{t,j} + v\mathbf{X}_{t,j+1}\|_2 . \tag{54}$$

Segment distances below 1.0 Å indicate severe trace contacts, and distances below 0.25 Å are treated as trace-intersection proxies.

Structural validity metrics across the mdCATH and ATLAS test sets are collected in Table 8 including neighbouring C_α–C_α distance statistics, FNC, GDT-TS, LDDT, and the average minimum of nonlocal (non-neighbouring) C_α–C_α distances and the fraction of structures in each trajectory with at least one C_α–C_α $< 1.0\mathring{A}$ as reporters of steric clashes.

Although C_α–C_α distances, FNC, GDT-TS and LDDT show good structural integrity compared to the MD references, there is a significant number of steric clashes reported. Whilst generally we can expect more steric clashes from any generative model than the reference MD, 20 and 6.56 of nonbonded C_α–C_α on average per frame showing steric clashes, and 91.8% and 87.6% of frames per trajectory with at least one steric clash for mdCATH and ATLAS respectively is likely a direct

Table 8. Structural validity metrics split by dataset and temperature. mdCATH overall and per-temperature columns are followed by a vertically separated ATLAS 300 K column; each condition reports the MD reference and predicted ensemble side by side.

Measure	mdCATH all		mdCATH 320 K		mdCATH 348 K		mdCATH 379 K		mdCATH 413 K		mdCATH 450 K		ATLAS 300 K	
	MD	Pred	MD	Pred	MD	Pred	MD	Pred	MD	Pred	MD	Pred	MD	Pred
C_α–C_α mean (Å)	3.831	3.804	3.833	3.808	3.832	3.805	3.831	3.803	3.830	3.802	3.829	3.804	3.836	3.813
C_α–C_α std (Å)	0.009	0.009	0.008	0.009	0.008	0.009	0.008	0.009	0.009	0.009	0.010	0.009	0.006	0.007
C_α–C_α 1st–99th pct. width (Å)	0.041	0.044	0.038	0.045	0.040	0.043	0.040	0.042	0.041	0.042	0.043	0.044	0.031	0.033
FNC mean	0.613	0.696	0.774	0.848	0.731	0.804	0.658	0.738	0.541	0.622	0.360	0.464	0.899	0.907
GDT-TS mean	0.447	0.555	0.601	0.728	0.555	0.674	0.479	0.594	0.375	0.469	0.226	0.308	0.656	0.764
LDDT mean	0.617	0.654	0.737	0.794	0.704	0.751	0.649	0.690	0.563	0.585	0.431	0.447	0.837	0.839
Mean min nonlocal C_α–C_α (Å)	3.488	1.339	3.513	2.206	3.504	1.835	3.497	1.352	3.480	0.858	3.446	0.439	3.443	2.188
Frames with any nonbonded C_α–C_α < 3.5 Å(%)	0.4	91.8	0.3	80.2	0.3	87.6	0.4	93.0	0.4	98.2	0.5	99.9	1.3	87.6
Frames with any nonbonded C_α–C_α < 3.0 Å(%)	0.0	84.2	0.0	65.6	0.0	76.1	0.0	85.1	0.0	94.8	0.0	99.3	0.0	72.1
Nonbonded C_α–C_α pair clashes/frame < 3.5 Å	0.005	20.755	0.003	5.909	0.003	8.485	0.004	13.372	0.007	25.742	0.008	49.996	0.013	6.566
Nonbonded C_α–C_α pair clashes/frame < 3.0 Å	0.001	12.761	0.000	3.339	0.000	4.911	0.000	7.958	0.002	15.798	0.002	31.624	0.000	3.394
Nonbonded C_α–C_α pair clashes/frame < 2.5 Å	0.001	7.320	0.000	1.819	0.000	2.713	0.000	4.468	0.002	9.037	0.002	18.458	0.000	1.762
Nonbonded C_α–C_α pair clashes/clashing frame < 2.5 Å	87.488	9.659	0.000	3.519	0.000	4.271	0.000	5.923	105.312	10.067	76.080	18.831	0.000	3.201
Nonbonded C_α–C_α pair clashes/trajectory < 3.5 Å	11.4	51888.6	5.9	14772.3	7.4	21213.4	8.4	33429.8	16.5	64354.0	18.9	124990.0	3.8	3283.2
Segment crossings/frame < 1.0 Å	0.001	4.146	0.000	0.791	0.000	1.301	0.000	2.571	0.002	5.268	0.002	10.854	0.000	0.621
Frames with nonlocal C_α–C_α segment < 1.0 Å(%)	0.0	51.0	0.0	21.2	0.0	31.0	0.0	47.8	0.0	68.3	0.0	87.1	0.0	18.4

result of our DCT representation. These results are the raw model inference output, not post-inference energy minimised. Particularly concerning is backbone trace nonlocal C_α–C_α segments $< 1.0\dot{A}$ occurring at least once in 51% and 18.4% of mdCATH and ATLAS test set generated structures respectively which are possible grounds for topological changes from chain crossover which would not be fixable by energy minimsation.