Beyond Ensembles: Simulating All-Atom Protein Dynamics in a Learned Latent Space

Aditya Sengar[*†‡] aditya.sengar@epfl.ch
Ali Hariri[*] ali.hariri@epfl.ch
Pierre Vandergheynst[*] pierre.vandergheynst@epfl.ch
Patrick Barth[†§] patrick.barth@epfl.ch

Abstract

Simulating the long-timescale dynamics of biomolecules is a central challenge in computational science. While enhanced sampling methods can accelerate these simulations, they rely on pre-defined collective variables that are often difficult to identify. A recent generative model, LD-FPG, demonstrated that this problem could be bypassed by learning to sample the static equilibrium ensemble as all-atom deformations from a reference structure, establishing a powerful method for all-atom ensemble generation. However, while this approach successfully captures a system's probable conformations, it does not model the temporal evolution between them. Here we extend LD-FPG with a temporal propagator that operates within the learned latent space and compare three classes: (i) score-guided Langevin dynamics, (ii) Koopman-based linear operators, and (iii) autoregressive neural networks. Within a unified encoder–propagator–decoder framework, we evaluate long-horizon stability, backbone and side-chain ensemble fidelity, and functional free-energy landscapes. Autoregressive neural networks deliver the most robust long rollouts; score-guided Langevin best recovers side-chain thermodynamics when the score is well learned; and Koopman provides an interpretable, lightweight baseline that tends to damp fluctuations. These results clarify the trade-offs among propagators and offer practical guidance for latent-space simulators of all-atom protein dynamics.

1 Introduction

Molecular simulations are indispensable for studying the complex dynamics that govern biological function, yet brute-force approaches struggle to access the slow, functionally relevant motions—such as protein folding, ligand binding, or allosteric switching—due to rugged energy landscapes and the dominance of rare events [1, 2]. To mitigate this gap—beyond what enhanced sampling can offer when suitable collective variables are hard to specify—a complementary strategy has gained traction: shifting the burden from raw coordinates to learned *latent* coordinates. In this *representation-first* view, the simulation problem is recast as a modular encoder–propagator–decoder pipeline: an encoder maps high-dimensional atomic configurations into a continuous, low-dimensional latent space; a propagator evolves the system's state within this simplified space; and a decoder maps the resulting latent trajectory back to all-atom coordinates [3, 4].

[*]Signal Processing Laboratory (LTS2), EPFL, Lausanne, Switzerland
[†]Institute of Bioengineering, EPFL, Lausanne, Switzerland
[‡]Corresponding author
[§]Ludwig Institute for Cancer Research, Lausanne, Switzerland

Preprint.

Progress within this paradigm comes from two complementary directions. The first focuses on learning the underlying physics, employing score-based diffusion, flow matching, and energy-based models to learn generative surrogates—sometimes yielding differentiable force fields—that implicitly define the system's potential of mean force [5, 6]. The second centers on learning simplified dynamical coordinates, using time-aware autoencoders or Koopman/DMD analysis to discover an intrinsic manifold where the long-term dynamics become stable, predictable, or even approximately linear [7–10].

While this framework is well-established, the choice of the latent **propagator**—the engine that drives the dynamics—remains a critical open question, as different models offer distinct trade-offs between physical rigor, long-term stability, and expressive power. Within this paradigm, we adopt LD-FPG [11] as the encoder–decoder backbone that learns an all-atom equilibrium ensemble in a pooled latent space, and we augment it with a temporal component. We then systematically compare, within the *same* latent space, three propagator classes: (i) **score-guided Langevin dynamics**, which leverages learned forces from the equilibrium distribution; (ii) **Koopman-based linear operators**, which offer long-horizon stability and interpretability; and (iii) **flexible neural networks (MLPs)**, which capture non-linear memory effects but can drift during long autoregressive rollouts [12–14].

We evaluate these propagators on alanine-dipeptide and two GPCRs (A1AR, A2AR), assessing long-horizon stability, backbone and side-chain ensemble fidelity, and functional free-energy landscapes. In brief, autoregressive neural networks provide the most reliable long rollouts; score-guided Langevin best recovers side-chain thermodynamics when the score is well learned; and Koopman serves as a lightweight, interpretable baseline that tends to damp fluctuations.

2 Related Work

The encoder–propagator–decoder blueprint. The core idea of simulating in a low-dimensional space is well established. Molecular Latent Space Simulators (LSS) [3] explicitly factorize the problem into three components: an encoder to find slow collective variables (CVs), a latent propagator to evolve them, and a decoder to generate all-atom structures. Similarly, Deep Generative MSMs (DeepGenMSM) [4] pair a latent Markovian transition model with a generative decoder to emit molecular configurations for each state. This blueprint has been realized in various forms, including autoregressive simulators with RNN/LSTM propagators [12], trajectory-level generators that cast MD as video synthesis [15], and invertible models like Boltzmann Generators that learn a direct map to the equilibrium distribution [16]. We follow this modular design but (i) *hold the LD-FPG decoder fixed* to control for reconstruction quality and (ii) focus our analysis on the choice of latent *propagator*.

Learning dynamically aware latent spaces (encoders). The quality of a latent simulation hinges on the quality of the latent space itself. While linear methods like time-lagged independent component analysis (TICA) remain strong baselines for identifying slow variables [17], deep learning has enabled far more expressive encoders. Time-lagged autoencoders such as the Variational Dynamics Encoder (VDE) learn non-linear representations predictive over a time delay Δt [7, 18]. VAMPnets use a variational principle to approximate the leading Koopman eigenfunctions, corresponding to the slowest dynamical processes [8]. Similarly, iterative methods like RAVE use VAEs to learn a latent distribution that guides the discovery of an optimal reaction coordinate for enhanced sampling [19]. To remove hand-crafted features, modern approaches leverage graph neural networks (GNNs) to learn CVs directly from coordinates in a permutation- and symmetry-aware manner [20–23], often with information-bottleneck objectives to explicitly optimize predictiveness of future states [24].

Linear dynamics in latent space (Koopman/DMD propagators). The Koopman-operator framework provides a route to linearize non-linear dynamics in a learned observable space [9]. Extended dynamic mode decomposition (EDMD) [25] and DMD [26] showed that a wide class of systems admits accurate linear predictors; subsequent work learns such observables end-to-end with deep encoders [27]. This has inspired Koopman autoencoders that enforce a linear evolution rule within the latent space [28, 29], yielding exceptional long-horizon stability by avoiding the compounding errors of iterated non-linear models, with continued advances in scalable kernels and consistency-enforcing architectures [10].

Non-linear sequence models (neural propagators). A complementary path directly learns the non-linear transition function. The Learning Effective Dynamics (LED) framework uses LSTMs to propagate latent variables and capture memory effects [12]; others pair RNNs with physical resampling such as Maximum Caliber to enforce kinetic consistency [30]. Continuous-time learners (Neural ODEs) offer flexible parametric flows [31], and graph-based simulators exploit locality for rollouts in interacting systems [32]. To curb instability, physics-preserving architectures (Hamiltonian/Lagrangian networks) enforce conservation laws [13, 14]. Our MLP propagator serves as a streamlined baseline in this family.

Physics-guided stochastic models (Langevin and diffusion). Generative models increasingly embed statistical-mechanics structure. Score-based diffusion has been used to learn effective force fields for coarse-grained MD [5], linking denoising, Langevin dynamics [5], and Fokker–Planck evolution [6]. Recent works (e.g., DiffMD, Score Dynamics) demonstrate larger stable time steps while retaining short-time kinetic signals [33, 34]. A key theme is *consistency* between the sampled equilibrium ensemble and the stationary distribution of the learned dynamics [6]; stability can also benefit from noise augmentation during training [5].

GPCRs as a proving ground for slow dynamics. G protein-coupled receptors (GPCRs) are a demanding testbed: their function hinges on slow transitions among metastable states mapped by landmark MD/MSM studies [1, 2, 35]. High-resolution structures (e.g., β_2AR–G_s) provide anchors for validating pathways [36], and reviews emphasize micro-switches and long-range allostery spanning orders of magnitude in time [37–39]. By evaluating on A_1AR/A_2AR, we benchmark whether latent simulators recover metastable states and kinetic pathways relevant to activation.

Positioning among recent generative simulators. Recent efforts span coordinate-space trajectory synthesis (e.g., diffusion-based GeoTDM) [40], SE(3)-equivariant flow matching for coarse-grained rollouts (F^3low) [41], and neural operators for full 3D dynamics beyond next-step prediction (EGNO) [42]. Latent and physics-informed approaches accelerate sampling or enforce structure—LAST for adaptive MD [43], ConfRover for joint conformation–dynamics learning [44], and NeuralMD with symmetric neural ODEs for binding dynamics [45]—while comparative studies benchmark diffusion, flow matching, and normalizing flows on MD tasks [46]. Complementing these advances, this work isolates the *propagator* choice by evaluating linear (Koopman), neural, and score-guided dynamics *within the same learned latent and fixed decoder*, enabling a controlled comparison of long-horizon stability, ensemble fidelity, and functional landscapes. We also considered MD-Gen [15], but as it was pre-trained on the Atlas dataset, which does not include alanine-dipeptide, re-training it for a fair comparison was beyond the scope of this work.

3 Methods

Our framework extends the Latent Diffusion for Full Protein Generation (LD-FPG) model by incorporating a temporal propagator that operates within its learned conformational latent space. To provide a self-contained description, we first briefly outline the relevant components of the LD-FPG architecture before detailing the three propagator models developed in this work.

3.1 Latent Space Representation from LD-FPG

The foundation of our method is the encoder-decoder architecture from LD-FPG. An encoder, implemented as a Chebyshev Graph Neural Network (ChebNet), learns a mapping from high-dimensional, all-atom protein coordinates $X(t) \in \mathbb{R}^{N\times 3}$ (where N is the number of heavy atoms) to a low-dimensional latent embedding $z(t) \in \mathbb{R}^d$. This encoder is trained on a Molecular Dynamics (MD) trajectory, producing a time-series of latent vectors $\{z_0, z_1, \ldots, z_M\}$ that captures the essential conformational dynamics of the protein. Our goal is to model the time evolution within this latent space, which can be expressed as a discrete-time update rule:

$$z_{t+1} = f(z_t) + \eta_t \tag{1}$$

where f is the propagator function we aim to learn, and η_t represents a stochastic noise term. We systematically compare three distinct classes of propagators for learning f. Key notation is summarized in Table S1, and a glossary in Table S2.

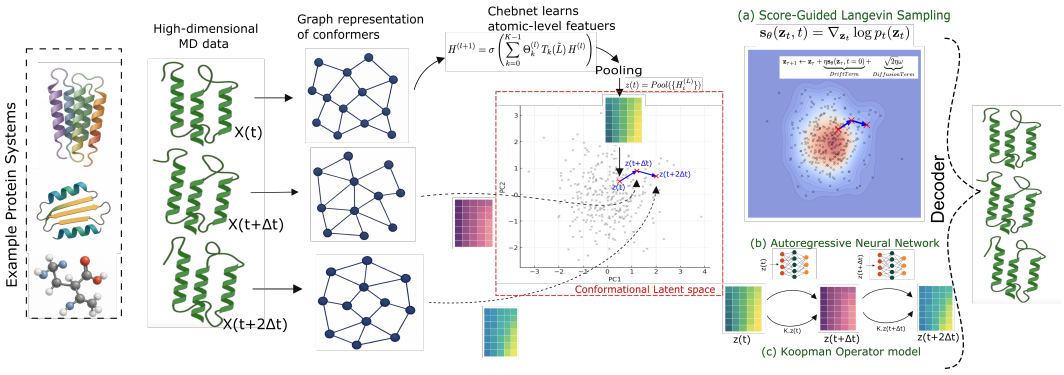

Figure 1: **Framework overview.** A pre-trained LD-FPG encoder (ChebNet; left) maps all-atom coordinates $X(t)$ to a pooled latent $z(t)$. Within this fixed latent, one of three *propagators* advances the state (red box): **(a)** score-guided Langevin using the LD-FPG denoiser to estimate $s_\theta(z,\tau) = \nabla_z \log p_\tau(z)$ at a fixed low-noise level; **(b)** an autoregressive neural network $z_{t+1} = f_\theta(z_t)$; and **(c)** a Koopman linear operator $z_{t+1} = Az_t$. The frozen LD-FPG decoder (right) maps the latent trajectory back to all-atom structures $\hat{X}(t+\Delta t)$

We use a one–frame latent stride for training pairs (z_t, z_{t+1}). Unless stated, rollout noise is $\eta_t \sim \mathcal{N}(0, I)$.

3.2 Koopman Propagator via Dynamic Mode Decomposition

We approximate the latent dynamics with a linear map

$$z_{t+1} \approx Az_t, \qquad A \in \mathbb{R}^{d\times d}. \tag{2}$$

Let the snapshot matrices collect *columns as time*:

$$\mathbf{X} = [\, z_0, \ldots, z_{M-2} \,] \in \mathbb{R}^{d\times(M-1)}, \quad \mathbf{Y} = [\, z_1, \ldots, z_{M-1} \,] \in \mathbb{R}^{d\times(M-1)}.$$

DMD solves $\min_A \|\mathbf{Y} - A\mathbf{X}\|_F^2$ with closed form

$$A = \mathbf{Y}\,\mathbf{X}^+, \tag{3}$$

where $^+$ is the Moore–Penrose pseudoinverse. For stability, we compute $\mathbf{X}^+$ via truncated SVD of $\mathbf{X}$ at rank $r < d$ (EDMD); r is chosen by retaining singular values above a fixed energy fraction (e.g., 95%). New trajectories follow $\hat{z}_{t+1} = A\hat{z}_t + \eta_t$ with optional $\eta_t \sim \mathcal{N}(0, \sigma^2 I)$.

3.3 Autoregressive Neural Network Propagator

To capture potentially complex, non-linear relationships in the dynamics, we employ a standard **Multi-Layer Perceptron (MLP)** as the propagator. The model learns a general non-linear function f_θ parameterized by weights θ:

$$z_{t+1} = f_\theta(z_t) \tag{4}$$

Our implementation of f_θ is a sequential network consisting of fully-connected layers with ReLU activation functions and Dropout for regularization. The model is trained to predict the state one step ahead by minimizing the **Mean Squared Error (MSE)** loss between the predicted state and the true next state over the training data:

$$\mathcal{L}(\theta) = \frac{1}{M-1} \sum_{t=0}^{M-2} \|f_\theta(z_t) - z_{t+1}\|^2 \tag{5}$$

The optimization is performed using the Adam optimizer. Similar to the Koopman model, trajectories are generated autoregressively from a starting point $\hat{z}_0$ by iteratively applying the learned function: $\hat{z}_{i+1} = f_\theta(\hat{z}_i) + \eta_i$.

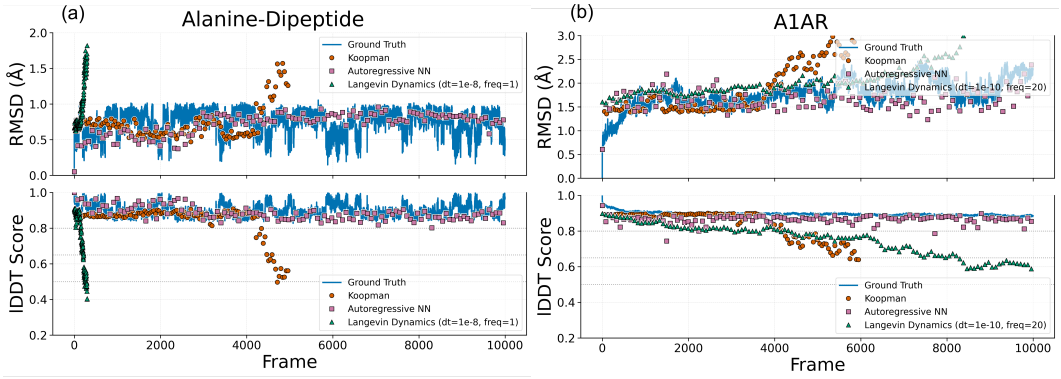

Figure 2: **Stability over long rollouts.** RMSD and lDDT versus frame index for (a) alanine-dipeptide and (b) A1AR. The *failure time* is defined as the first frame where lDDT (computed relative to the first frame of the trajectory) drops below 0.65. Autoregressive NN maintains the longest stable rollout on A1AR (no failure within 10,000 frames), while Koopman and Langevin fail earlier; on alanine-dipeptide, Koopman and NN persist for thousands of frames whereas Langevin fails early.

3.4 Score-Guided Langevin Propagator

This approach frames the dynamics from a statistical mechanics perspective, simulating the evolution of the system under the influence of a potential of mean force. The dynamics are governed by the **overdamped Langevin equation**, which in discretized form is:

$$z_{t+1} = z_t - \nabla_z U(z_t)\, h + \sqrt{2Th}\, \eta_t, \tag{6}$$

where $U(z)$ is the potential of mean force, Δt is the integration time step, T is the temperature, and $\eta_t \sim \mathcal{N}(0, I)$ is a random Gaussian vector.

The key insight is that the force term, $-\nabla_z U(z)$, can be related to the score of the equilibrium (Boltzmann) distribution, $s(z) = \nabla_z \log p(z)$, since $p(z) \propto \exp(-U(z)/T)$. The pre-trained diffusion model from LD-FPG, $\epsilon_\theta(z_\tau, \tau)$, provides a direct way to estimate this score. According to score-based generative modeling theory, the score of a data distribution perturbed with noise level σ_τ is related to the optimal denoiser:

$$\nabla_z \log p_\tau(z_\tau) \approx -\frac{\epsilon_\theta(z_\tau, \tau)}{\sigma_\tau^2}. \tag{7}$$

By evaluating the model at a low, fixed noise level (i.e., a small diffusion timestep τ_{noise}), we approximate the score of the true data distribution, $s(z) \approx -\frac{\epsilon_\theta(z, \tau_{\text{noise}})}{\sigma^2_{\tau_{\text{noise}}}}$ [47**?**].

Substituting the score for the force term, we arrive at the simulation update rule:

$$z_{i+1} = z_i + Th\, s(z_i) + \sqrt{2Th}\, \eta_i, \qquad \eta_i \sim \mathcal{N}(0, I). \tag{8}$$

This method directly leverages the learned equilibrium distribution from LD-FPG to drive a physically-motivated, stochastic simulation in the latent space. To enhance numerical stability during long rollouts, we also implement optional score clipping, where the norm of the score vector $s(z_i)$ is capped at a predefined maximum value.

4 Results and Discussion

We benchmark the three latent-space propagators—Koopman, Autoregressive Neural Network (NN), and score-guided Langevin—within the unified LD-FPG latent (Fig. 1). Datasets and code are summarized in Appendix S1.2. Metrics include stability (RMSD, lDDT), equilibrium-ensemble fidelity (backbone and side-chain dihedral JSD), and functional free-energy landscapes. Unless noted, *failure time* is the first frame where lDDT relative to the initial frame drops below 0.65.

4.1 Long-horizon stability: autoregressive NN is most robust

Figure 2 tracks RMSD and lDDT through time; *failure time* is the first sampled point where lDDT (vs. the first frame) drops below 0.65. For alanine-dipeptide (Fig. 2a), Koopman and the Autoregressive NN remain stable for several thousand sampled frames (failure at 4443 and 3176, respectively), whereas Langevin fails early (206). On A1AR (Fig. 2b), the Autoregressive NN completes the full 10,000-frame rollout without failure; Langevin remains stable to 7476; Koopman fails at 5740. A2AR (Table 1; SI) follows the same ranking: Autoregressive NN $>$ Langevin $>$ Koopman for long-rollout stability.

A note on "frames" for Langevin. For Koopman and NN, each frame corresponds to the dataset stride (the models are trained to map $z_t \to z_{t+1}$), so the horizontal axis coincides with the MD sampling stride. By contrast, the Langevin propagator integrates a latent SDE with internal step size Δt and a separate sampling stride; thus a "frame" is a *sampled* SDE state, not a single update of the integrator. To make curves visually comparable, we *calibrated* $(\Delta t, \text{sampling stride})$ per system to match the short-horizon RMSD across Langevin replicas to the base simulation. Concretely, for alanine-dipeptide we used $\Delta t = 10^{-8}$ and sampled *every* step (one internal step per plotted frame), whereas for A1AR we used $\Delta t = 10^{-10}$ and sampled *every 20 steps* (effective time per plotted frame $= 20\,\Delta t$). The reported failure indices for Langevin therefore count sampled outputs; converting to physical time would scale the A1AR axis by $20\,\Delta t$ and the alanine axis by Δt.

Why Langevin fails early on alanine. The rapid failure of Langevin dynamics on the dipeptide likely stems from a combination of challenges in the training data, the system's intrinsic properties, and the simulation hyperparameters. First, the large time step between frames in the source MD simulation can result in a sparsely sampled, fragmented latent manifold. This makes it difficult for the diffusion model to learn a smooth and continuous score function ($s(z) = \nabla_z \log p(z)$), yielding a noisy or inaccurate effective force field that is prone to instability. Second, the dipeptide's small size leads to large intrinsic fluctuations and high successive-frame displacements (RMSD ~ 0.9 Å), making the lDDT failure criterion particularly stringent. Finally, the simulation hyperparameters were reused from GPCR settings [11]. The large effective step size, chosen to model such highly diffusive systems, proves too aggressive when combined with the imperfect score function. This combination rapidly drives the simulation into unphysical regions, a deviation that is quickly detected by the sensitive lDDT metric.

4.2 Equilibrium-ensemble fidelity: NN best on backbone; Langevin best on side-chains

We evaluate how well each propagator reproduces the *equilibrium* ensemble in dihedral space using 2D free-energy maps for backbone (ϕ, ψ) and side-chain (χ_1, χ_2) angles (Fig. 3). Fidelity is quantified with the Jensen–Shannon divergence (JSD) between model and ground-truth distributions (Table 1); side-chain JSD is aggregated over all residues so that it captures global rotamer statistics.

Alanine-dipeptide. All three models recover the canonical Ramachandran basins. The *Autoregressive NN* most closely matches basin *shape* and separation (backbone JSD $= 0.0056$), slightly outperforming *Koopman* (0.0085). *Langevin* under-samples and over-smooths the landscape (backbone JSD $= 0.029$), consistent with its early rollout failure on this small, rapidly fluctuating system (Fig. 2a). For comparison, an Equivariant Graph Neural Operator (EGNO) baseline [42] yielded a significantly higher backbone JSD of 0.3875, underscoring the effectiveness of our specialized latent-space propagators.

A1AR : For the large GPCR, the *Autoregressive NN* gives the most faithful backbone ensemble (backbone JSD $= 0.0443$), while *Langevin* broadens low-energy regions (backbone JSD $= 0.1943$). In contrast, side-chains tell the opposite story: the score-guided *Langevin* dynamics sharply recovers rotameric structure in (χ_1, χ_2) with the lowest divergence (JSD $= 0.0223$), the NN is second-best (0.0436), and *Koopman* is notably diffuse (0.1144). Visually, the Langevin maps reproduce the expected χ_1 bands (g^+, anti, g^-) present in the ground truth, whereas the NN blurs band edges and Koopman largely washes them out (Fig. 3, right).

Why side-chains favor Langevin. Side-chain rotamers are governed by local barriers and short-range couplings. Because the Langevin propagator uses the learned score $s(z) = \nabla_z \log p(z)$ from

LD-FPG, it injects a *thermodynamically consistent* drift toward high-probability neighborhoods in latent space and adds small isotropic noise. This combination encourages frequent, barrier-crossing micro-moves that preserve the stationary rotamer distribution. The NN, optimized for one-step prediction, can accumulate exposure bias and smooth sharp multimodality over long rollouts; Koopman's linear evolution further damps variance, leading to broadened or merged rotamer basins.

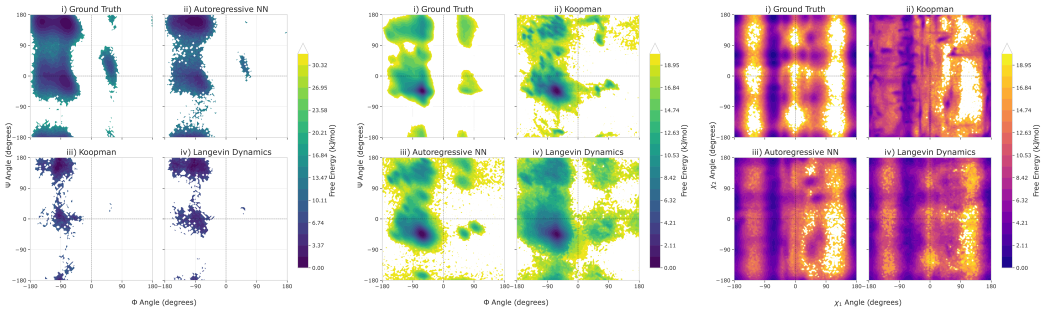

Figure 3: **Ensemble fidelity in dihedral space.** Free-energy maps for backbone (ϕ, ψ) and, for A1AR, side-chain (χ_1, χ_2) angles. Alanine-dipeptide (left) shows recovery of the canonical Ramachandran basins by all methods, with the Autoregressive NN most closely matching basin shapes given the long rollout. For A1AR (middle/right), NN best matches backbone statistics, while score-guided Langevin achieves the lowest divergence for side-chains (side-chain JSD aggregated over all residues).

Amplitude of motion. Across systems, *Koopman* systematically underestimates fluctuations (lower mean RMSF than ground truth), reflecting its variance-damping bias, whereas *Langevin* and the *NN* better match the amplitude of motion (Table 1). Residue-resolved trends follow the same pattern (see Fig. S1).

4.3 Functional GPCR surfaces: TM-distance free energies capture activation

We test whether each propagator reproduces the inactive↔active switching of GPCRs by projecting trajectories onto a two-dimensional free-energy surface $F(\mathrm{TM3\text{–}6}, \mathrm{TM3\text{–}7})$. Activation is characterized by an outward motion of TM6 that increases TM3–6, while TM3–7 helps resolve the geometry of the intracellular opening. In Fig. 4 (A1AR) the *background* heat map is the ground-truth surface; *overlaid contours* correspond to Koopman (purple), Autoregressive NN (cyan), and Langevin (orange). Right panels show the corresponding one-dimensional profiles.

A1AR. Both the NN and Langevin models recover the location and curvature of the principal low–free-energy valley. Langevin spans the valley most extensively, covering the transition corridor between inactive and active-like states. The NN tracks the same valley with a tighter footprint, under-sampling the flanks. Koopman identifies the basin center but exhibits stiffer, more isotropic contours and elevated apparent barriers, limiting coverage of the transition path. The one-dimensional slices mirror these trends: NN and Langevin reproduce the primary minimum and overall shape, whereas Koopman inflates barriers.

A2AR (reference). For A2AR (Fig. S2), switching involves coordinated changes along *both* axes: TM3–6 increases with the TM6 outward motion and TM3–7 shifts as the intracellular pocket reshapes, yielding a diagonal valley in the $(\mathrm{TM3\text{–}6}, \mathrm{TM3\text{–}7})$ plane. Langevin again achieves the broadest coverage along both coordinates and reaches the transition corridor; the NN follows the valley with narrower support; Koopman under-covers the corridor and smooths anisotropy.

Summary. Together with the dihedral analyses, these surfaces indicate that the non-linear (NN) and score-guided (Langevin) propagators better explore and *connect* metastable states relevant to GPCR activation, while the strictly linear Koopman rule provides a conservative baseline that tends to over-regularize barriers and shrink anisotropy.

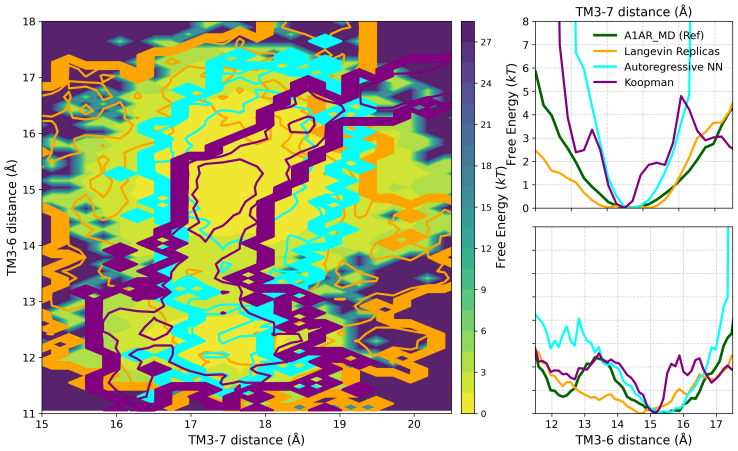

Figure 4: **Functional free-energy surface for A1AR.** Two-dimensional free-energy over TM3–6 ("TM36") and TM3–7 ("TM37") distances. The *background heat map* shows the ground-truth surface; *overlaid contours* show Koopman (purple), Autoregressive NN (cyan), and Langevin (orange). Right panels: corresponding 1D free-energy profiles along each coordinate. NN and Langevin track the main basin and curvature; Koopman identifies the basin but overestimates barriers and misses anisotropy.

Table 1: **Quantitative summary across systems.** Failure time is the first frame with lDDT (vs. the initial frame) < 0.65. JSDs measure divergence from ground-truth dihedral distributions. Average RMSF is in Å. For alanine-dipeptide, we include an EGNO baseline [42] trained for 100 frames. A2AR results are in the SI.

System / Metric	Model	Failure Time (frames)	Backbone JSD	Sidechain JSD	Average RMSF (Å)
Alanine-Dipeptide	Ground Truth	N/A	N/A	N/A	0.8222
	Koopman	4443	0.0085	N/A	0.8320
	Autoregressive NN	3176	0.0056	N/A	0.8171
	Langevin Dynamics	206	0.029	N/A	0.8233
	EGNO (baseline)	1000	0.3875	N/A	1.0481
A1AR	Ground Truth	N/A	N/A	N/A	1.5809
	Koopman	5740	0.0472	0.1144	0.7482
	Autoregressive NN	No failure (10000)	0.0443	0.0436	0.7880
	Langevin Dynamics	7476	0.1943	0.0223	1.1303
A2AR	Ground Truth	N/A	N/A	N/A	1.9454
	Koopman	2324	0.1195	0.1380	1.2531
	Autoregressive NN	5789	0.0679	0.0691	0.8200
	Langevin Dynamics	5432	0.1211	0.0065	0.7463

4.4 Outlook

Across systems, the propagators fill complementary roles. The **autoregressive NN** provides the most reliable long-horizon rollouts and the closest backbone statistics (it is the only A1AR run without failure). **Score-guided Langevin** best reproduces *side-chain* thermodynamics on GPCRs and yields realistic local fluctuations, though it is sensitive to score quality and step size. **Koopman** remains a fast, interpretable baseline that offers medium-term stability but damps variance and blurs rotamer structure. In practice, we favor the NN when long all-atom trajectories are the goal and Langevin when rotamer distributions and local thermodynamics are paramount, provided the denoiser and latent connectivity are well tuned.

References

[1] Ron O Dror, Daniel H Arlow, Paul Maragakis, Thomas J Mildorf, Albert C Pan, Huafeng Xu, David W Borhani, and David E Shaw. Activation mechanism of the β 2-adrenergic receptor. *Proceedings of the National Academy of Sciences*, 108(46):18684–18689, 2011.

[2] Naomi R Latorraca, AJ Venkatakrishnan, and Ron O Dror. Gpcr dynamics: structures in motion. *Chemical reviews*, 117(1):139–155, 2017.

[3] Hythem Sidky, Wei Chen, and Andrew L Ferguson. Molecular latent space simulators. *Chemical Science*, 11(35):9459–9467, 2020.

[4] Hao Wu, Andreas Mardt, Luca Pasquali, and Frank Noe. Deep generative markov state models. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zugner, Marco Federici, Cecilia Clementi, Frank Noé, Robert Pinsler, and Rianne van den Berg. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *Journal of Chemical Theory and Computation*, 19(18):6151–6159, 2023.

[6] Michael Plainer, Hao Wu, Leon Klein, Stephan Günnemann, and Frank Noé. Consistent sampling and simulation: Molecular dynamics with energy-based diffusion models. *arXiv preprint arXiv:2506.17139*, 2025.

[7] Carlos X Hernández, Hannah K Wayment-Steele, Mohammad M Sultan, Brooke E Husic, and Vijay S Pande. Variational encoding of complex dynamics. *Physical Review E*, 97(6):062412, 2018.

[8] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):5, 2018.

[9] Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

[10] Omri Azencot, N Benjamin Erichson, Vanessa Lin, and Michael Mahoney. Forecasting sequential data using consistent koopman autoencoders. In *International Conference on Machine Learning*, pages 475–485. PMLR, 2020.

[11] Aditya Sengar, Ali Hariri, Daniel Probst, Patrick Barth, and Pierre Vandergheynst. Generative modeling of full-atom protein conformations using latent diffusion on graph embeddings. *arXiv preprint arXiv:2506.17064*, 2025.

[12] Pantelis R Vlachas, Julija Zavadlav, Matej Praprotnik, and Petros Koumoutsakos. Accelerated simulations of molecular systems through learning of effective dynamics. *Journal of Chemical Theory and Computation*, 18(1):538–549, 2021.

[13] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

[14] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

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

[16] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[17] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1), 2013.

[18] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics*, 148(24), 2018.

[19] João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary. Reweighted autoencoded variational bayes for enhanced sampling (rave). *The Journal of chemical physics*, 149(7), 2018.

[20] Ziyue Zou, Dedi Wang, and Pratyush Tiwary. A graph neural network-state predictive information bottleneck (gnn-spib) approach for learning molecular thermodynamics and kinetics. *Digital Discovery*, 4(1):211–221, 2025.

[21] Mahdi Ghorbani, Samarjeet Prasad, Jeffery B Klauda, and Bernard R Brooks. Graphvampnet, using graph neural networks and variational approach to markov processes for dynamical modeling of biomolecules. *The Journal of Chemical Physics*, 156(18), 2022.

[22] Jintu Zhang, Luigi Bonati, Enrico Trizio, Odin Zhang, Yu Kang, TingJun Hou, and Michele Parrinello. Descriptor-free collective variables from geometric graph neural networks. *Journal of Chemical Theory and Computation*, 20(24):10787–10797, 2024.

[23] Alexandre Agm Duval, Victor Schmidt, Alex Hernández-Garcıa, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. Faenet: Frame averaging equivariant gnn for materials modeling. In *International Conference on Machine Learning*, pages 9013–9033. PMLR, 2023.

[24] Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154(13), 2021.

[25] Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

[26] Jonathan H Tu. *Dynamic mode decomposition: Theory and applications*. PhD thesis, Princeton University, 2013.

[27] Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):4950, 2018.

[28] Kshitij Tayal, Arvind Renganathan, Rahul Ghosh, Xiaowei Jia, and Vipin Kumar. Koopman invertible autoencoder: Leveraging forward and backward dynamics for temporal modeling. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 588–597. IEEE, 2023.

[29] Xingjian Wu, Xiangfei Qiu, Hongfan Gao, Jilin Hu, Bin Yang, and Chenjuan Guo. K2 vae: A koopman-kalman enhanced variational autoencoder for probabilistic time series forecasting. *arXiv preprint arXiv:2505.23017*, 2025.

[30] Sun-Ting Tsai, Eric Fields, Yijia Xu, En-Jui Kuo, and Pratyush Tiwary. Path sampling of recurrent neural networks by incorporating known physics. *Nature communications*, 13(1): 7231, 2022.

[31] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[32] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

[33] Fang Wu and Stan Z Li. Diffmd: a geometric diffusion model for molecular dynamics simulations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 5321–5329, 2023.

[34] Tim Hsu, Babak Sadigh, Vasily Bulatov, and Fei Zhou. Score dynamics: Scaling molecular dynamics with picoseconds time steps via conditional diffusion model. *Journal of Chemical Theory and Computation*, 20(6):2335–2348, 2024.

[35] Kai J Kohlhoff, Diwakar Shukla, Morgan Lawrenz, Gregory R Bowman, David E Konerding, Dan Belov, Russ B Altman, and Vijay S Pande. Cloud-based simulations on google exacycle reveal ligand modulation of gpcr activation pathways. *Nature chemistry*, 6(1):15–21, 2014.

[36] Søren GF Rasmussen, Brian T DeVree, Yaozhong Zou, Andrew C Kruse, Ka Young Chung, Tong Sun Kobilka, Foon Sun Thian, Pil Seok Chae, Els Pardon, Diane Calinski, et al. Crystal structure of the β2 adrenergic receptor–gs protein complex. *Nature*, 477(7366):549–555, 2011.

[37] Yinglong Miao and J Andrew McCammon. G-protein coupled receptors: advances in simulation and drug discovery. *Current opinion in structural biology*, 41:83–89, 2016.

[38] Aashish Manglik, Tae Hun Kim, Matthieu Masureel, Christian Altenbach, Zhongyu Yang, Daniel Hilger, Michael T Lerch, Tong Sun Kobilka, Foon Sun Thian, Wayne L Hubbell, et al. Structural insights into the dynamic process of β2-adrenergic receptor signaling. *Cell*, 161(5): 1101–1111, 2015.

[39] Aiveliagaram J Venkatakrishnan, Xavier Deupi, Guillaume Lebon, Franziska M Heydenreich, Tilman Flock, Tamara Miljus, Santhanam Balaji, Michel Bouvier, Dmitry B Veprintsev, Christopher G Tate, et al. Diverse activation pathways in class a gpcrs converge near the g-protein-coupling region. *Nature*, 536(7617):484–487, 2016.

[40] Jiaqi Han, Minkai Xu, Aaron Lou, Haotian Ye, and Stefano Ermon. Geometric trajectory diffusion models. *Advances in Neural Information Processing Systems*, 37:25628–25662, 2024.

[41] Shaoning Li, Yusong Wang, Mingyu Li, Jian Zhang, Bin Shao, Nanning Zheng, and Jian Tang. F3low: Frame-to-frame coarse-grained molecular dynamics with se (3) guided flow matching. *arXiv preprint arXiv:2405.00751*, 2024.

[42] Minkai Xu, Jiaqi Han, Aaron Lou, Jean Kossaifi, Arvind Ramanathan, Kamyar Azizzadenesheli, Jure Leskovec, Stefano Ermon, and Anima Anandkumar. Equivariant graph neural operator for modeling 3d dynamics. *arXiv preprint arXiv:2401.11037*, 2024.

[43] Hao Tian, Xi Jiang, Sian Xiao, Hunter La Force, Eric C Larson, and Peng Tao. Last: Latent space-assisted adaptive sampling for protein trajectories. *Journal of chemical information and modeling*, 63(1):67–75, 2022.

[44] Yuning Shen, Lihao Wang, Huizhuo Yuan, Yan Wang, Bangji Yang, and Quanquan Gu. Simultaneous modeling of protein conformation and dynamics via autoregression. *arXiv preprint arXiv:2505.17478*, 2025.

[45] Shengchao Liu, Weitao Du, Yanjing Li, Zhuoxinran Li, Vignesh Bhethanabotla, Nakul Rampal, Omar Yaghi, Christian Borgs, Anima Anandkumar, Hongyu Guo, et al. A multi-grained symmetric differential equation model for learning protein-ligand binding dynamics. *arXiv preprint arXiv:2401.15122*, 2024.

[46] Richard John, Lukas Herron, and Pratyush Tiwary. A comparison of probabilistic generative frameworks for molecular simulations. *The Journal of Chemical Physics*, 162(11), 2025.

[47] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021.

S1 Supplementary Information

S1.1 Supplementary Figures for A2AR

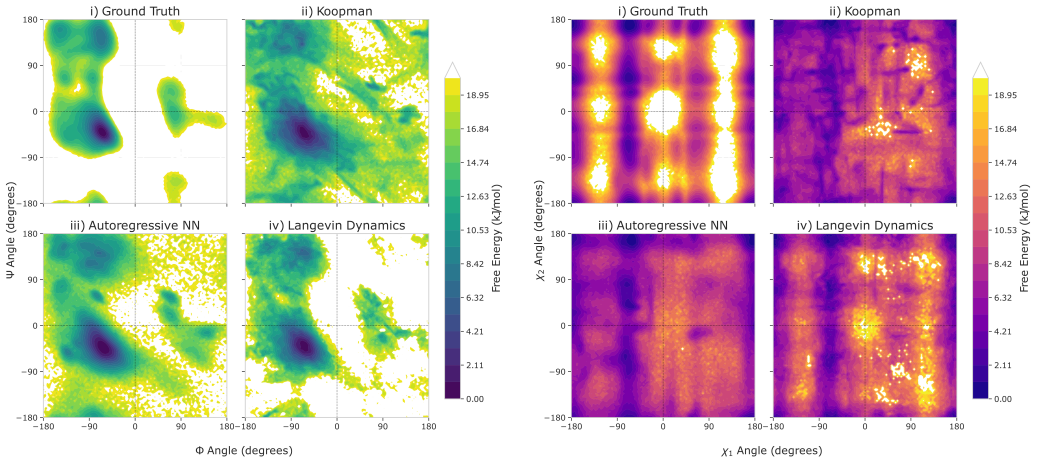

Figure S1: **A2AR dihedral statistics (SI).** Backbone (ϕ, ψ) and side-chain (χ_1, χ_2) free-energy maps for A2AR. Trends mirror the main text: Autoregressive NN provides the best backbone match among learning-based models, while score-guided Langevin attains the lowest divergence for side-chains (side-chain JSD aggregated over all residues).

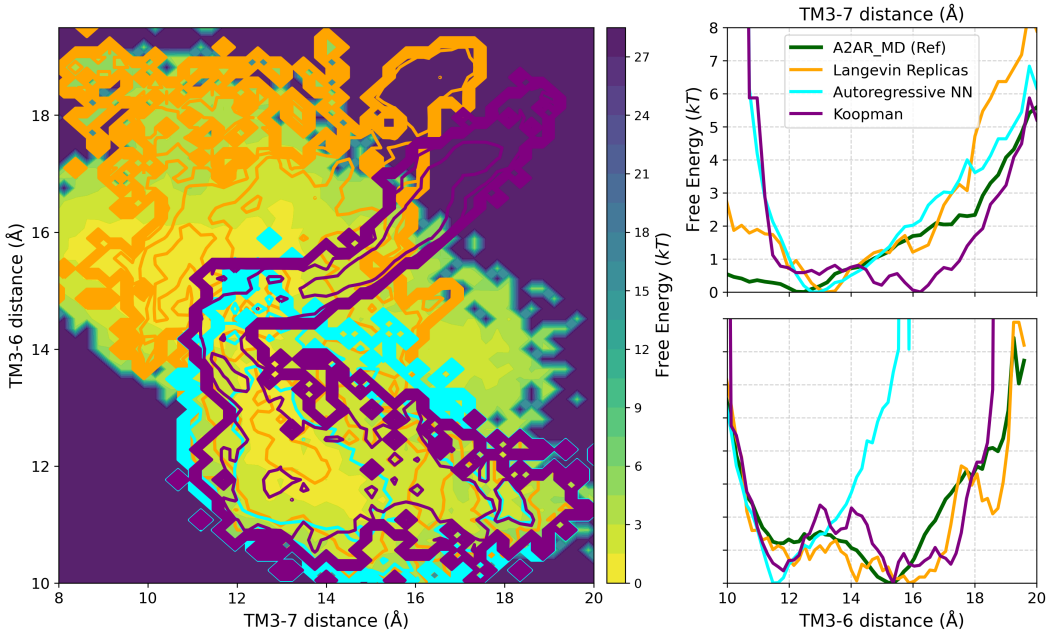

Figure S2: **A2AR TM-distance free-energy (SI).** Two-dimensional free-energy over TM3–6 ("TM36") and TM3–7 ("TM37") distances. The *background* is the ground-truth surface; *three overlapping contour sets* show Koopman (purple), Autoregressive NN (cyan), and Langevin (orange). Right panels: 1D free-energy profiles along each coordinate. Consistent with A1AR, NN and Langevin follow the principal basin and curvature more closely than Koopman.

S1.2 Simulation Datasets and Code availability

The complete code for the LD-FPG framework is publicly available on GitHub: https://github.com/adityasengar/LD-FPG/ and the code for the latent propagator is available here: https://github.com/adityasengar/latent-dynamics-propagators

The latent trajectories used to train the propagators were derived from three publicly available Molecular Dynamics (MD) simulation datasets. For each system, the original source provided a structure file (PDB) and a coordinate trajectory (XTC). These files were then processed using the LD-FPG framework's preprocessing scripts to generate the inputs for our models, including a `condensed.json` file that provides consistent, zero-based atom indexing and defines the atom quadruplets required for calculating all backbone and side-chain dihedral angles.

Alanine Dipeptide. This dataset features N-acetyl-L-alanine-N'-methylamide, a 22-atom molecule commonly known as alanine dipeptide. It is a canonical benchmark for developing and testing new simulation methods due to its simple yet non-trivial conformational landscape, which is primarily described by its two backbone dihedral angles (ϕ, ψ). The data was sourced from the CMB data repository at `ftp.imp.fu-berlin.de` and consists of a 250 ns simulation trajectory with solvent molecules removed.

Adenosine A1 Receptor (A1AR). To test our method on a complex, biologically relevant system, we used a 1 μs simulation of the human adenosine A1 receptor, a prototypical Class A GPCR involved in cardiovascular and neurological signaling. The dataset, containing the trajectory and initial PDB structure for chain A of the receptor, was derived from the simulation data available on Zenodo (DOI: 10.5281/zenodo.7944479).

Adenosine A2A Receptor (A2AR). Our second GPCR test case was the human adenosine A2A receptor, another Class A GPCR that serves as a key model system for studying receptor activation mechanisms. The data corresponds to a simulation of the receptor in its apo (ligand-free) state and was sourced from a Zenodo record (DOI: 10.5281/zenodo.13460724) supplementary to a detailed study on its dynamics.

S1.3 Qualitative rollouts: dihedral flips in a dipeptide and TM6 motion in a GPCR

To complement the quantitative metrics, Fig. S3 shows *structural snapshots* taken directly from our latent-space rollouts. The top panel illustrates alanine–dipeptide; the bottom panel shows the A_1AR GPCR. These images link the latent dynamics to familiar structural changes: local backbone dihedral flips in a small molecule and the hallmark outward displacement of transmembrane helix 6 (TM6) in a receptor.

Alanine–dipeptide (Koopman). The upper snapshots are sampled at evenly spaced steps from a **Koopman** rollout. Across frames, the molecule visits distinct conformers driven by rotations about the backbone dihedrals (ϕ, ψ), covering multiple metastable Ramachandran regions rather than collapsing to a single geometry. This visual diversity is consistent with the small backbone JSD reported in Table 1 and the ensemble analysis in Sec. 4.2, while also reflecting Koopman's tendency to slightly damp fluctuations.

A_1AR (Autoregressive NN). The lower snapshots are taken from a single long rollout of the **autoregressive neural** propagator, again at evenly spaced frames. The dashed circle marks the intracellular end of TM6. Over time TM6 swings *outward* from the receptor core while the seven-helix bundle remains well-folded, matching the activation-associated opening on the cytoplasmic side. This qualitative progression mirrors the quantitative TM3–6/TM3–7 free-energy analysis in Fig. 4, where the NN tracks the principal low-energy valley and samples the transition corridor without structural collapse.

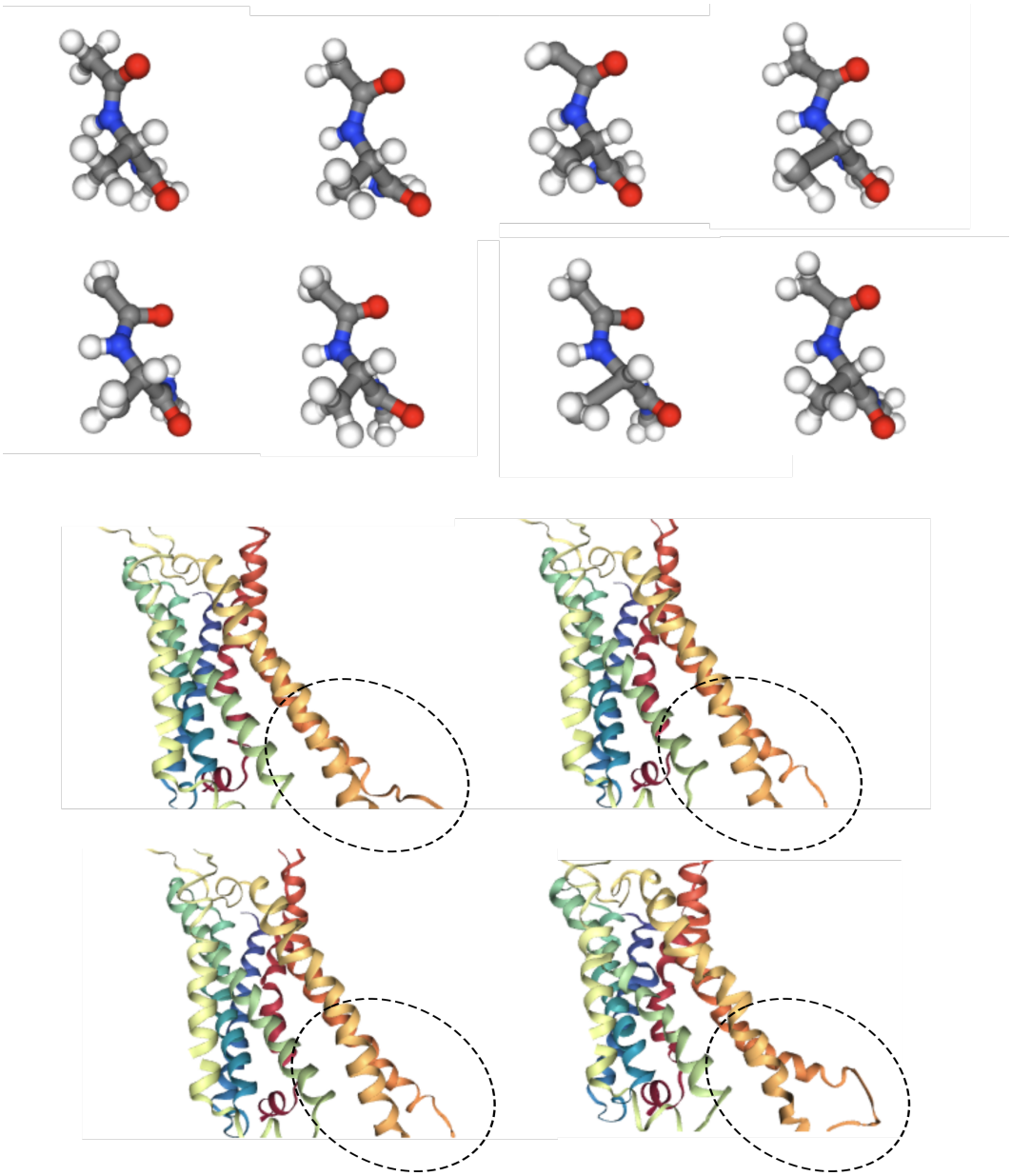

Figure S3: **Representative structural snapshots from latent rollouts.** *Top:* Alanine–dipeptide conformers sampled from a **Koopman** rollout show backbone dihedral changes across frames. *Bottom:* A_1AR snapshots from an **autoregressive NN** rollout; the dashed circle highlights the intracellular end of TM6, which moves outward over time—a hallmark of GPCR activation. These qualitative views align with the ensemble statistics in Table 1 and the TM-distance thermodynamics in Fig. 4.

S1.4 Supplementary Tables

Table S1: Summary of notation used in the description of the propagator models.

Symbol	Description
$z_t \in \mathbb{R}^d$	Latent space vector at time t.
$\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d\times(M-1)}$	Snapshot matrices with *columns* as time: $\mathbf{X} = [z_0, \ldots, z_{M-2}]$, $\mathbf{Y} = [z_1, \ldots, z_{M-1}]$.
$A \in \mathbb{R}^{d\times d}$	Koopman linear operator.
f_θ	Autoregressive neural network with parameters θ.
$p(z)$	Equilibrium probability distribution of the latent variable z.
$s(z) = \nabla_z \log p(z)$	Score function of the equilibrium distribution.
$\epsilon_\theta(z_t, \tau)$	Pre-trained LD-FPG diffusion model (denoiser).
σ_τ	Noise schedule standard deviation from the diffusion model at step τ.
Δt	Integration time step for Langevin dynamics.
T	Temperature parameter for Langevin dynamics.
η_t	Stochastic noise term, typically $\eta_t \sim \mathcal{N}(0, \sigma^2 I)$.

Table S2: **Glossary of Biophysical Terms.** Definitions of key concepts from molecular dynamics and structural biology used in this work, tailored for a machine learning audience.

Term	**Description for an ML Audience**
Collective Variable (CV)	A low-dimensional function of atomic coordinates (e.g., a distance or angle) designed to capture a specific, slow dynamic process like protein folding. Traditional simulation methods often require pre-defining good CVs; our work uses a learned latent space to discover them automatically.
Potential of Mean Force (PMF)	Essentially, an effective "energy landscape" for a molecule in solution. Lower values correspond to more probable (stable) conformations. It is the target distribution that score-guided Langevin dynamics aims to sample from, often visualized as a "free-energy surface".
Dihedral Angles (ϕ, ψ, χ)	Rotational angles around covalent bonds that define the geometry of a molecule. • ϕ **(phi),** ψ **(psi):** Define the rotation of the protein **backbone**. • χ **(chi):** Define the rotation of the amino acid **side-chains**. Their statistical distributions are a sensitive measure of structural fidelity.
Ramachandran Plot	A 2D plot of the backbone dihedral angles (ϕ, ψ). Certain regions of this plot are "allowed" based on steric constraints, leading to characteristic high-probability basins that correspond to stable secondary structures like alpha-helices and beta-sheets.
Rotamer	A discrete, low-energy, and therefore highly probable conformation of a protein's side-chain, defined by its set of χ angles. A key test for generative models is whether they can reproduce the correct statistical distribution of these rotameric states.
GPCR	*G protein-coupled receptor*. A large and important family of transmembrane proteins that act as cellular signal transducers. They are highly dynamic and switch between different functional states (e.g., inactive, active), making them an ideal and challenging test system for dynamic models.
lDDT	*local Distance Difference Test*. A metric for assessing the quality of a protein structure prediction by evaluating how well local inter-atomic distances are preserved relative to a reference structure. Unlike RMSD, it is less sensitive to global rotations and more focused on local geometric accuracy.
RMSF	*Root-Mean-Square Fluctuation*. For each atom, this metric calculates the standard deviation of its position over time in a simulation trajectory. It measures the "amplitude of motion" or flexibility of different parts of the protein.