ByteDance | AI Drug Discovery

ANEW THERAPEUTICS

Learning the All-Atom Equilibrium Distribution of Biomolecular Interactions at Scale

Yusong Wang[1,2,*], **Youjun Xu**[1,2,*], **Wentao Li**[1,2,*], **Haoyu Yu**[1,2,†], **Wenjuan Tan**[1,2], **Shaoning Li**[1,2], **Qiaojing Huang**[1,2], **Nanjun Chen**[1,2], **Xuan Wu**[1,2], **Qilong Wu**[1,2], **Kai Liu**[1,2,†]

[1]ByteDance AI Drug Discovery, [2]Anew Therapeutics

[*]Contributed equally, [†]Corresponding authors

Abstract

Biomolecular functions are governed by dynamic conformational ensembles rather than static structures. While models like AlphaFold have revolutionized static structure prediction, accurately capturing the equilibrium distribution of all-atom biomolecular interactions remains a significant challenge due to the high computational cost of molecular dynamics (MD). We present AnewSampling, a transferable generative foundation framework designed for the high-fidelity sampling of all-atom equilibrium distributions, which is the first model to faithfully reproduce MD at the all-atom level. It uses a novel quotient-space generative framework to ensure mathematical consistency and leverages the largest self-curated database of protein-ligand trajectories to date, with over 15 million conformations. Statistically, AnewSampling consistently outperforms all prior generative methods on the ATLAS monomer benchmark, and the all-atom capabilities of AnewSampling enable close statistical alignment with ground-truth MD for evaluating atomic biomolecular interactions in protein-ligand dynamics. Furthermore, AnewSampling successfully recovers coupled ligand and side-chain motions in CDK2 systems, overcoming a major sampling hurdle inherent to conventional MD. AnewSampling enables rapid exploration of conformational landscapes prior to intensive simulations, elucidating fundamental biophysical mechanisms and accelerating the broader design of functional biomolecules.

Date: March 11, 2026

Correspondence: Kai Liu at liukai0824@bytedance.com, Haoyu Yu at haoyu.yu@bytedance.com

1 Introduction

Biomolecular functions are fundamentally dictated by precise, atomic-scale physical interactions between molecules, which constitute the essential machinery of life and drive complex processes ranging from enzymatic catalysis [1] to intricate cellular signaling [2]. Consequently, achieving a comprehensive understanding of the biophysical mechanisms underlying these interactions, alongside developing robust methods for their artificial modulation, represents one of the most significant challenges in contemporary science. Notably, the structural characterization of these atomic interfaces has recently undergone a revolutionary transformation, largely driven by geometric deep learning architectures such as AlphaFold [3–6] and related predictive models [7–11]. While it is now possible to predict the static ground states of biomolecular assemblies with unprecedented accuracy, these rigid snapshots offer only a limited perspective. In physiological environments, biomolecules

exist in a state of constant conformational fluctuation, where biological efficacy is frequently governed by the dynamic transitions between diverse structural states rather than the geometry of a single structure [12]. Therefore, to comprehensively understand and engineer molecular function, it is imperative to move beyond static structural descriptions and accurately recover the dynamic equilibrium dictated by the Boltzmann distribution.

Experimental techniques such as cryogenic electron microscopy (cryo-EM) can resolve diverse conformational states [13], yet their application remains limited by high costs and the intensive time required for sample preparation and data collection. Molecular dynamics (MD) simulations [14] provide a more accessible alternative for characterizing these fluctuations at atomic resolution. While MD has long been the standard for studying biomolecular dynamics, the serial nature of numerical integration requires femtosecond-scale timesteps, creating a major computational bottleneck. Even relatively simple protein folding phenomena, which typically occur on microsecond-to-millisecond scales, can currently only be captured using special-purpose supercomputers [15]. Since such specialized hardware is not widely accessible for routine drug discovery, the dependence on conventional physical simulations ultimately limits the throughput and scalability of pipelines that require a detailed understanding of molecular dynamics.

Recent advances in artificial intelligence (AI) have explored multiple avenues to overcome the inherent bottlenecks of molecular dynamics. One prominent strategy involves multiple sequence alignment (MSA) perturbation [16–18]. While these heuristic methods possess the capability to sample alternative conformations, they offer limited controllability and fail to reflect rigorous thermodynamic distributions. Concurrently, generative AI has begun to bridge this gap, demonstrating that deep learning can emulate MD simulations orders of magnitude faster. These emerging generative approaches generally follow two distinct paradigms: trajectory-based transition modeling and equilibrium distribution learning. The first paradigm focuses on learning transition operators to generate kinetic trajectories from an initial frame. Although these methods have demonstrated the capability for capturing free energy landscapes [19–21] and observe specific dynamical phenomena [22–24], they face significant challenges regarding sampling reliability. Due to the inherent autocorrelation between sequential samples, ensuring statistical convergence remains difficult. Consequently, the resulting ensemble may be sensitive to the choice of the initial frames. In parallel, another paradigm [25–27] seeks to directly learn transferable equilibrium distributions. While BioEmu [27] has emerged as a landmark framework in this category by adapting an architecture similar to AlphaFold2, it frequently relies on simplified chemical representations. By utilizing backbone-based SE(3) diffusion, it models only the distributions of the backbone while ignoring side chains and ligands, thereby stripping away the atomic interactions that drive drug binding. In the post-AlphaFold2 era, structure prediction [6] has shifted toward generative all-atom modeling; however, the scarcity of MD training data often limits the resulting structural diversity. While MD-enhanced models, such as Boltz2 [10], produce more distinct conformations, the mathematical inconsistencies between training objectives and sampling procedures in their generative frameworks hinder their ability to accurately recover true equilibrium ensembles. Moreover, the practical utility of all-atom ensemble emulators remains difficult to quantify, given the lack of consensus on standard evaluation metrics for characterizing the dynamics of biomolecular interactions [28]. These limitations highlight a fundamental question: *Can generative frameworks investigate biomolecular interactions at an all-atom level while faithfully reproducing converged MD equilibrium distributions in a transferable manner, and what metrics can provide a rigorous evaluation of these atomic interactions?*

Toward this end, we introduce AnewSampling, a foundational generative framework developed to sample from the all-atom equilibrium distributions of biomolecular complexes. As shown in Fig. 1b, by introducing a generative framework that mathematically ensures consistency between training and sampling, we employ a sophisticated hybrid fine-tuning strategy to shift the objective from static structure prediction to stochastic dynamic modeling. The transferability of AnewSampling is underpinned by the construction of AnewSampling-DB (Fig. 1c), the largest database of protein-ligand dynamics to date, which encompasses over 15 million conformations. Spanning 10,297 unique protein sequences and 27,979 unique ligand SMILES, AnewSampling-DB covers a vast chemical space of both congeneric and non-congeneric series. All trajectories are generated through a unified pipeline using a consistent set of force fields to ensure rigorous physical consistency across diverse interaction profiles. Furthermore, we address the fundamental ambiguity of evaluating generative dynamics by proposing a multi-level assessment strategy (Fig. 1a). This framework rigorously validates

physical fidelity by quantifying ligand conformations, protein-ligand interaction patterns, and intrinsic protein dynamics.

We rigorously validated AnewSampling, beginning with the ATLAS [29] dataset of protein monomers to establish baseline performance. While existing generative methods are restricted to monomeric systems, AnewSampling consistently outperforms these specialized models across 13/13 evaluation metrics, confirming the fundamental efficacy and robustness of our generative framework. To assess protein-ligand dynamics, we utilized three distinct test sets: a diverse held-out set of PDB systems to demonstrate broad chemical generalizability, the JACS & Merck [30] datasets—industrial standards for structure-activity relationship (SAR) analysis—to capture subtle ligand-induced conformational changes and an in-house dataset derived from our internal drug discovery pipelines. Across these challenging tasks, AnewSampling achieves a statistical alignment with ground-truth MD distributions that far surpasses both pure static predictors and recent MD-enhanced models such as Boltz2 [10], which frequently suffer from mode collapse or fail to recover the full breadth of the equilibrium ensemble. Notably, while AnewSampling is trained exclusively on standard molecular dynamics trajectories, it demonstrates an emergent capability for enhanced sampling. By learning the underlying biophysical interactions shared across diverse chemical systems, the model effectively bypasses the high energy barriers that typically trap conventional MD. For CDK2 complexes (1H1R and 1H1S) [31], AnewSampling successfully navigates complex free energy landscapes to reproduce the multimodal ensembles observed in replica-exchange MD (REMD), recovering coupled ligand and side-chain motions that are often missed by conventional MD within practical simulation budgets.

The unprecedented computational efficiency of AnewSampling enables the rapid exploring of complex conformational landscapes, thereby facilitating the integration of the framework into diverse research and industrial pipelines. By providing a scalable method for the exploration of dynamic distributions prior to intensive physical simulations, AnewSampling catalyzes a paradigm shift toward dynamics-aware design, wherein the comprehensive characterization of biomolecular flexibility directly informs the development of adaptive inhibitors and the precision engineering of functional biomolecules.

2 Results

2.1 Overview of AnewSampling

AnewSampling is a transferable generative foundation framework designed for the high-fidelity sampling of all-atom equilibrium distributions of biomolecular complexes. To address the challenges of limited structural diversity in existing models, AnewSampling enables the faithful reproduction of MD accuracy across diverse systems, ranging from protein monomers to complex protein-ligand interactions. The framework is trained on AnewSampling-DB, the largest self-curated database of protein-ligand trajectories to date (Fig. 1c), which encompasses over 15 million conformations and 31,364 unique complexes.

The overall architecture of AnewSampling leverages a pretrained AlphaFold3-like architecture (Fig. 1b), integrating an Input Embedder, an MSA Module, and a Pairformer stack. To repurpose this architecture from deterministic static prediction to stochastic thermodynamic sampling, we implement a stratified hybrid fine-tuning strategy. Specifically, sequence representation modules are adapted via Low-Rank Adaptation (LoRA) to preserve their robust pretrained representation capabilities and mitigate the catastrophic forgetting of geometric reasoning. In contrast, the Diffusion Module undergoes full-parameter fine-tuning to accommodate the fundamental paradigm shift from deterministic structure prediction to a distributional sampler and to fully capture the complex manifold of distribution learning. A key innovation is the development of quotient-space diffusion (Fig. 1d). By mathematically factorizing out rigid-body degrees of freedom and constraining the generative process to the manifold of internal shapes, this formulation resolves the mathematical inconsistencies inherent in alignment-based training and ensures the precise recovery of the Boltzmann distribution.

To emulate the physical principle of ergodicity, we introduce a Cluster-Based Template Guidance mechanism (Fig. 1b). This approach decouples the target distribution from specific starting coordinates by randomly sampling and perturbing template structures from distinct conformational clusters, thereby enforcing an exhaustive exploration of the equilibrium ensemble from any valid initial state. Together, these advancements allow AnewSampling to investigate atomic interactions at an all-atom level while faithfully reproducing

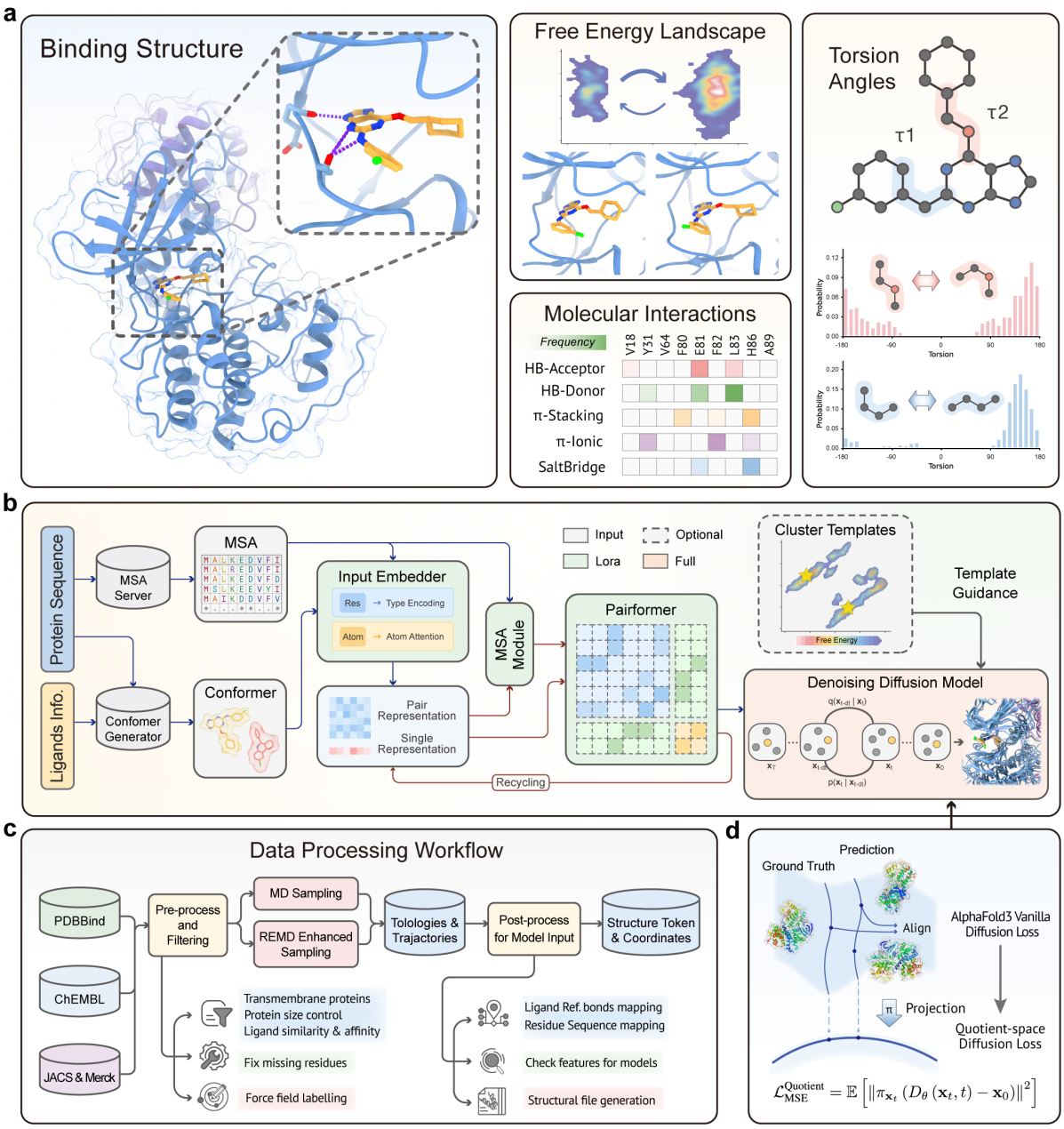

Figure 1 Overview of the AnewSampling framework for sampling biomolecular conformational ensembles. **(a)** Characterization of the dynamic binding landscape. Beyond a single static structure, protein-ligand interactions are visualized via free energy landscapes with distinct metastable states. These are further characterized by dynamic molecular interaction fingerprints and multimodal torsion angle distributions. **(b)** AnewSampling architecture and training objective. The model processes protein sequences (via MSA) and ligands through the Input Embedder, MSA Module, and Pairformer module. To shift the objective from static structure prediction to stochastic dynamic modeling, Low-Rank Adaptation (LoRA) is integrated into the sequence feature representation modules (green), while the Diffusion module undergoes full-parameter fine-tuning (red). The generation process is guided by clustered noisy templates, with training supervised by a novel quotient-space diffusion loss ($\mathcal{L}_{\text{MSE}}^{\text{Quotient}}$) that projects coordinates onto a quotient manifold to resolve symmetry invariances. **(c)** Data curation and processing workflow. Raw structural data from PDBBind, ChEMBL, and JACS & Merck undergo rigorous quality filtering—including resolution truncation and redundancy removal—prior to MD and REMD enhanced sampling. The resulting trajectories are featurized and tokenized to serve as high-fidelity inputs for model training and testing.

converged MD equilibrium distributions in a transferable manner.

2.2 Sampling Generalizes Across Unseen Conformational and Chemical Spaces

Current cutting-edge models reliably predict static structures but frequently fail to represent dynamic conformational distributions of protein-ligand complexes. To examine this systematically, we compared conformations generated by different models to a stringent ground-truth ensemble derived from MD simulations.

Table 1 Evaluation of protein dynamics modeling on the ATLAS test set. Performance comparison of AnewSampling against state-of-the-art generative baselines for generating monomeric protein conformational ensembles. Performance metrics for all baseline methods evaluated on the ATLAS dataset are sourced from the BioMD literature. Arrows (↑ or ↓) adjacent to each metric designate whether higher or lower values indicate superior performance. Evaluated statistical measures include Pearson correlation (r), root mean Wasserstein-2 (W2) distance, Jaccard index (J), and Spearman correlation (r_s). Bold text denotes the best performance achieved for each individual metric.

Category	Metrics	ESMFlow-MD (Full)	ESMFlow-MD (Distilled)	ConfDiff	BioEmu	Str2Str	MDGen	EBA	BioMD	AnewSampling
Predicting flexibility	Pairwise RMSD r(↑)	0.19	0.19	0.59	0.46	0.23	0.48	0.62	0.70	**0.85**
	Global RMSF r(↑)	0.31	0.33	0.67	0.57	0.38	0.50	0.71	0.76	**0.89**
	Per-target RMSF r(↑)	0.76	0.74	0.85	0.71	0.57	0.71	0.90	0.91	**0.93**
Distributional accuracy	Root mean W2-dist.(↓)	3.60	4.23	2.76	4.32	4.05	2.69	2.43	2.18	**1.88**
	Trans. contrib.(↓)	3.13	3.75	2.23	4.04	3.73	-	2.03	1.89	**1.63**
	Var. contrib.(↓)	1.74	1.90	1.40	1.77	1.43	-	1.20	1.10	**0.89**
	MD PCA W2-dist.(↓)	1.51	1.87	1.44	1.97	2.04	1.89	1.19	1.24	**1.16**
	Joint PCA W2-dist.(↓)	3.19	3.79	2.25	3.98	3.55	-	2.04	1.82	**1.44**
	% PC-sim > 0.5 ↑	26	33	35	**51**	12	-	44	46	**51**
Ensemble observables	Weak contacts J(↑)	0.55	0.48	0.59	0.33	0.43	0.51	0.65	0.61	**0.70**
	Transient contacts J(↑)	0.34	0.30	0.36	-	-	-	0.41	0.46	**0.52**
	Exposed residue J(↑)	0.49	0.43	0.50	-	-	0.29	0.70	0.70	**0.75**
	Exposed MI matrix r_s(↑)	0.20	0.16	0.24	0.07	0.21	-	0.36	0.34	**0.41**

ATLAS Dataset To comprehensively evaluate the capacity of AnewSampling to generate realistic conformational ensembles for protein monomers, we benchmarked our model on the ATLAS test set using standard metrics established in the AlphaFlow [32]. As detailed in Table 1, AnewSampling consistently outperforms existing state-of-the-art baseline methods—including BioMD [33], BioEmu [27]—across all major evaluation categories. Specifically, AnewSampling demonstrates unparalleled accuracy in predicting flexibility, achieving the highest Pearson correlations with reference REMD trajectories for Pairwise RMSD ($r = 0.85$) and per-target RMSF ($r = 0.93$). Furthermore, our method exhibits exceptional distributional accuracy by minimizing the root mean Wasserstein-2 (W2) distance to 1.88 (compared to 2.18 for the next best method, BioMD) and successfully capturing the principal components of the MD conformational space. Crucially, AnewSampling excels at recovering transient ensemble observables, achieving state-of-the-art Jaccard indices for predicting weak contacts ($J = 0.70$) and highly transient contacts ($J = 0.52$). Collectively, these comprehensive metrics confirm that AnewSampling is a highly robust generative model capable of accurately reproducing the thermodynamic distributions and transient biophysical microstates inherent to monomeric protein dynamics.

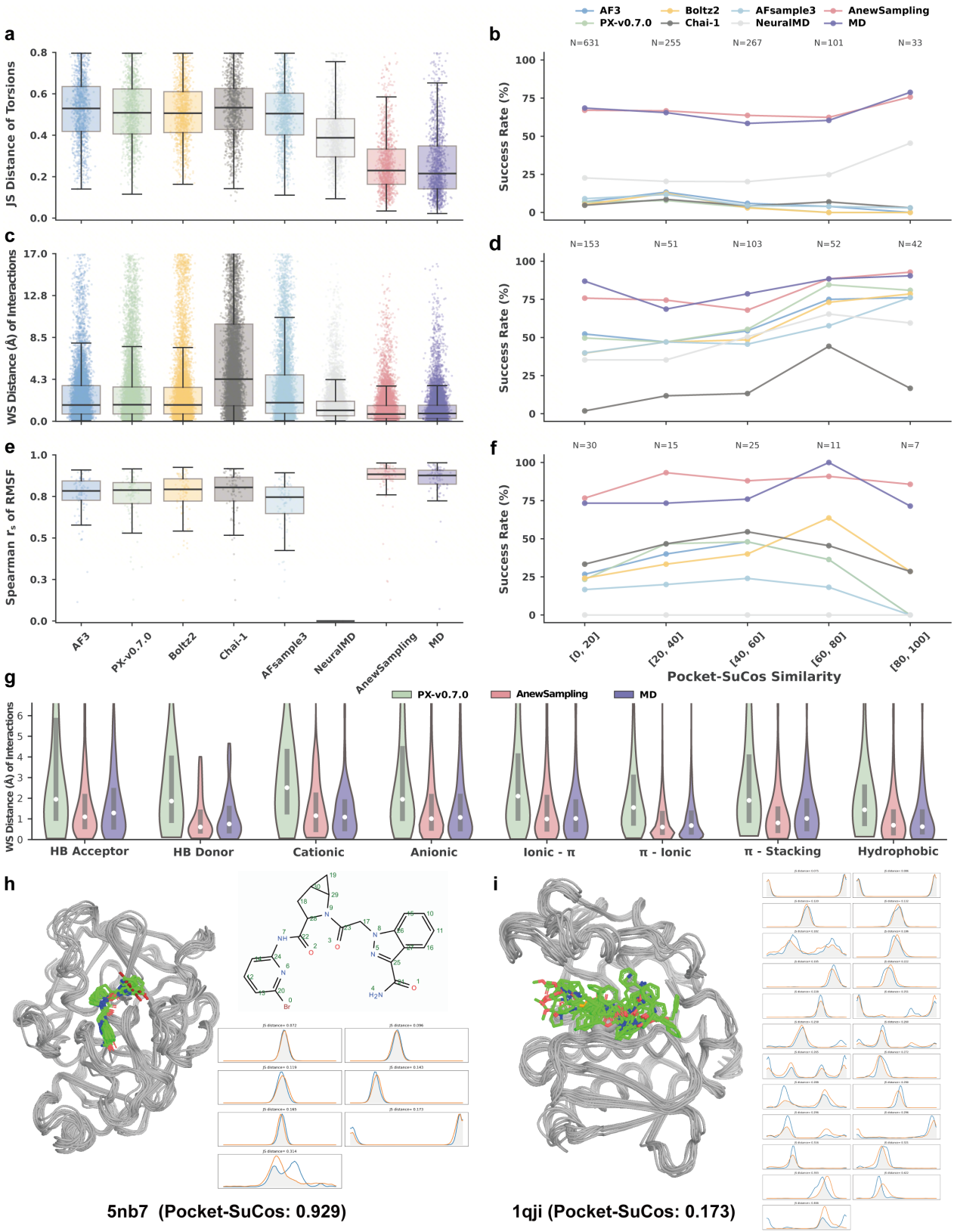
a
JS Distance of Torsions
b
Success Rate (%)
N=631
N=255
N=267
N=101
N=33
AF3
PX-v0.7.0
Boltz2
Chai-1
AFsample3
NeuralMD
AnewSampling
MD
c
WS Distance (Å) of Interactions
d
N=153
N=51
N=103
N=52
N=42
e
Spearman r_s of RMSF
f
N=30
N=15
N=25
N=11
N=7
[0, 20]
[20, 40]
[40, 60]
[60, 80]
[80, 100]
Pocket-SuCos Similarity
g
HB Acceptor
HB Donor
Cationic
Anionic
Ionic - π
π - Ionic
π - Stacking
Hydrophobic
h
5nb7 (Pocket-SuCos: 0.929)
i
1qji (Pocket-SuCos: 0.173)

Figure 2 Evaluation of generative models for protein-ligand conformational ensemble generation on the held-out test set. (a) Boxplot of JS distances for ligand torsion angle distributions across 500 generated conformations compared to a 501-frame REMD reference ensemble. **(b)** Success rate (defined as JS distance ≤ 0.3) for ligand torsion predictions, stratified by Pocket-SuCos similarity to evaluate generalization capability. N indicates the number of torsions in each bin. **(c)** Boxplot of WS distances (in Å) comparing the distributions of protein-ligand interaction distances. This metric assesses non-covalent interaction stability without relying on reference-dependent superposition. **(d)** Success rate (defined as WS distance ≤ 0.3 Å) across different Pocket-SuCos similarity bins. N indicates the number of interactions in each bin. **(e)** Boxplot of Spearman correlation coefficients (r_s) for protein backbone Cα root-mean-square fluctuation (RMSF), evaluating the consistency of global protein dynamics. **(f)** Success rate (Spearman correlation ≥ 0.85) for capturing protein backbone dynamics across similarity bins. N indicates the number of systems in each bin. **(g)** Violin plots detailing the WS distance distributions for specific non-covalent interaction types (e.g., Hydrogen Bond Acceptor/Donor, Cationic, π-Stacking). **(h, i)** Structural overlays of generated protein backbones and ligand conformations, alongside corresponding ligand torsion probability density functions, for two representative test cases: a highly similar target (5nb7, similarity: 0.929) and a highly dissimilar target (1qji, similarity: 0.173).

Held-out Test Set We assessed the intrinsic flexibility of the ligands by calculating the Jensen-Shannon (JS) distance of ligand torsion angle distributions (Fig. 2a,b and Table S1). Models trained exclusively on static structures (AlphaFold3 (AF3) [6], Protenix (PX-v0.7.0) [8], Chai-1 [11]) exhibited median JS distances of approximately 0.5, indicating a profound inability to reproduce dynamic ligand distributions. Notably, although Boltz2 [10] incorporates limited MD data and B-factor during training, it similarly failed to capture these dynamics, yielding a median JS distance of 0.5297. This observation strongly suggests that training generative models via random structural perturbation of single, static conformations is fundamentally insufficient for learning the underlying physical energy landscape. In contrast, models trained on MD conformational data showed marked improvements. Strikingly, AnewSampling achieved a conformational distribution statistically indistinguishable from the MD baseline, yielding a median JS distance of 0.2402 (MD: 0.2251). Furthermore, AnewSampling maintained a consistently high success rate (JS distance ≤ 0.3) across diverse Pocket-SuCos similarity bins (a similarity partitioning metric adopted from IsoDDE [34]), demonstrating robust generalization to novel, out-of-distribution complexes (Fig. 2b).

Because ligand torsions do not fully capture the binding mechanism, we introduced a superposition-free metric using Wasserstein (WS) distance to evaluate the stability of specific atom-pair interactions within the binding pocket. While the co-folding models yielded median WS distances around 2.0 Å, AnewSampling accurately modeled the physical stability of these non-covalent interactions, achieving a median WS distance of 0.9931 Å—comparable to the MD baseline of 1.0561 Å (Fig. 2c and Table S2). For highly conserved interactions observed during MD simulations (frequency ≥ 0.8, Table S3), AnewSampling achieved a 66.3% success rate at a stringent WS distance threshold (≤ 0.3 Å), substantially outperforming the best co-folding model (AF3, 48.8%). This high-fidelity interaction modeling held true across all physical interaction types, including hydrogen bonds, ionic, and π-stacking interactions (Fig. 2g).

Beyond the binding pocket, we evaluated global protein dynamics by computing the Spearman correlation (r_s) and root-mean-square error (RMSE) of the protein backbone Cα RMSF (Fig. 2e and Table S4). AnewSampling effectively preserved global conformational motions, yielding a median RMSE of 0.2286, which surpassed both the best co-folding model (Boltz2: 0.7864) and the MD baseline (0.2699). Impressively, AnewSampling achieved a 75% success rate for highly correlated global dynamics ($r_s \geq 0.85$), outperforming the MD reference (68%), and maintained this accuracy across all similarity bins (Fig. 2f).

In-house Dataset To further validate the real-world applicability and robustness of AnewSampling, we extended our evaluation to an independent, internal dataset comprising 17 structurally diverse protein-ligand complexes (similarity: [0.06, 0.62]) derived from active drug discovery pipelines. Consistent with our test set benchmarks, AnewSampling maintained stable and high-fidelity conformational ensemble generation across these novel complexes. The model accurately reproduced local ligand flexibility, achieving a 68.8% success rate in torsion JS distance (≤ 0.3) (Fig S1b). It also successfully captured highly conserved protein-ligand interaction networks, yielding an 87.7% success rate (WS distance ≤ 0.3 Å) that closely trails the rigorous MD reference (97.3%) (Fig S1d). Furthermore, AnewSampling exhibited exceptional consistency in predicting

global protein backbone dynamics, achieving a 100.0% success rate for highly correlated global motions ($r_s \geq 0.85$) (Fig S1f), marginally outperforming the inherent variance of the MD baseline (94.1%). These findings underscore the capacity of AnewSampling to generalize effectively to novel, therapeutically relevant targets without degradation in thermodynamic sampling accuracy.

The transition from predicting static, single-state protein-ligand structures to generating thermodynamic ensembles represents a critical frontier in structure-based drug design. Our benchmarking clearly illustrates that while leading diffusion-based co-folding models (such as AF3 and Protenix) excel at finding local energy minima, they inherently lack the underlying physics required to sample a Boltzmann-like distribution of the target complex. AnewSampling overcomes this fundamental limitation by learning directly from MD-derived conformational spaces. The results unequivocally demonstrate that AnewSampling is the first all-atom generative model capable of producing highly accurate, equilibrium-state conformational distributions that mirror extensive REMD sampling (evidence in subsection 2.3). The model's ability to maintain physical stability (evidenced by interaction WS distances) and global motion coherence (evidenced by RMSF correlation) across highly dissimilar targets highlights its robust generalization.

While AnewSampling achieves unprecedented parity with MD sampling for buried pocket interactions, our structural analysis (Fig. 2h,i) reveals minor deviations in the torsion distributions of highly solvent-exposed ligand moieties. We hypothesize that the current training paradigm captures the induced fit and pocket constraints well, but may lack sufficient sampling of unbound, solvent-driven ligand flexibility. Future iterations of AnewSampling could directly address this by incorporating large-scale, ligand-only MD datasets as a structural prior, thereby refining the generative distributions of solvent-facing functional groups and achieving true thermodynamic completeness.

2.3 Capturing perturbation-induced conformational shifts in SAR series

A fundamental challenge in lead optimization is that minor chemical modifications to a ligand—such as the addition of a methyl group, a halogen, or a subtle linker extension—can induce thermodynamically significant shifts in the conformational ensemble of the binding pocket. Conventional static structure prediction methods frequently neglect these subtleties, generally producing inflexible, fairly uniform pocket geometries over entire congeneric series. To investigate whether AnewSampling can detect and model these fine-grained structural responses, we evaluated our model on congeneric structure-activity relationship (SAR) series from the highly validated JACS and Merck benchmark datasets [30], comprising 593 complexes across 16 protein targets. Among these systems, where the median Pocket-SuCos similarity ranged from 0.53 to 0.93, our model consistently delivered outstanding performance. As shown in Fig. 3a. AnewSampling achieved the highest success rates among all evaluated generative methods across all three metrics, closely approaching the performance of the standard MD baseline.

The success rate of torsion JS distance ≤ 0.3 specifically reflects how well a model reproduces the ligand conformational sampling distributions obtained from REMD (Fig. 3a left and Table S5). Static models—including AF3, Protenix, Boltz2, and Chai-1—inherently failed to reflect dynamic distributional shifts within the binding pocket. Consequently, they performed poorly on this metric, with maximum success rates below 20% and scoring below 2% for the vast majority of systems. While AFsample3 is designed to capture protein structural diversity, it similarly exhibited poor performance in predicting ligand conformational sampling. NeuralMD, which predicts trajectories for pocket-bound ligands, demonstrated marginal distribution prediction capabilities (overall success rate of 28%). In stark contrast, AnewSampling accurately captured these distributions, achieving an overall success rate of 66%, which is highly comparable to the sampling fidelity of standard MD.

Notably, standard MD simulations exhibited relatively low success rates in torsion JS distance for two specific targets: TNKS2 and HIF-2α. This discrepancy highlights inconsistant distributions sampled from standard MD and REMD, implying that these systems pose significant sampling challenges for conventional simulations. This difficulty stems from two primary factors. First, sampling specific torsions—such as the rotatable bonds between rings in TNKS2 ligands and the rotation of decorated phenyl rings in HIF-2α ligands—is hindered by high energy barriers. While MD frequently becomes kinetically trapped in local minima, REMD successfully traverses these barriers to sample multiple minima, resulting in divergent distributions. Second, some of the ligands in HIF-2α system may contain suboptimal initail local structures, such as incorrect sulfonyl

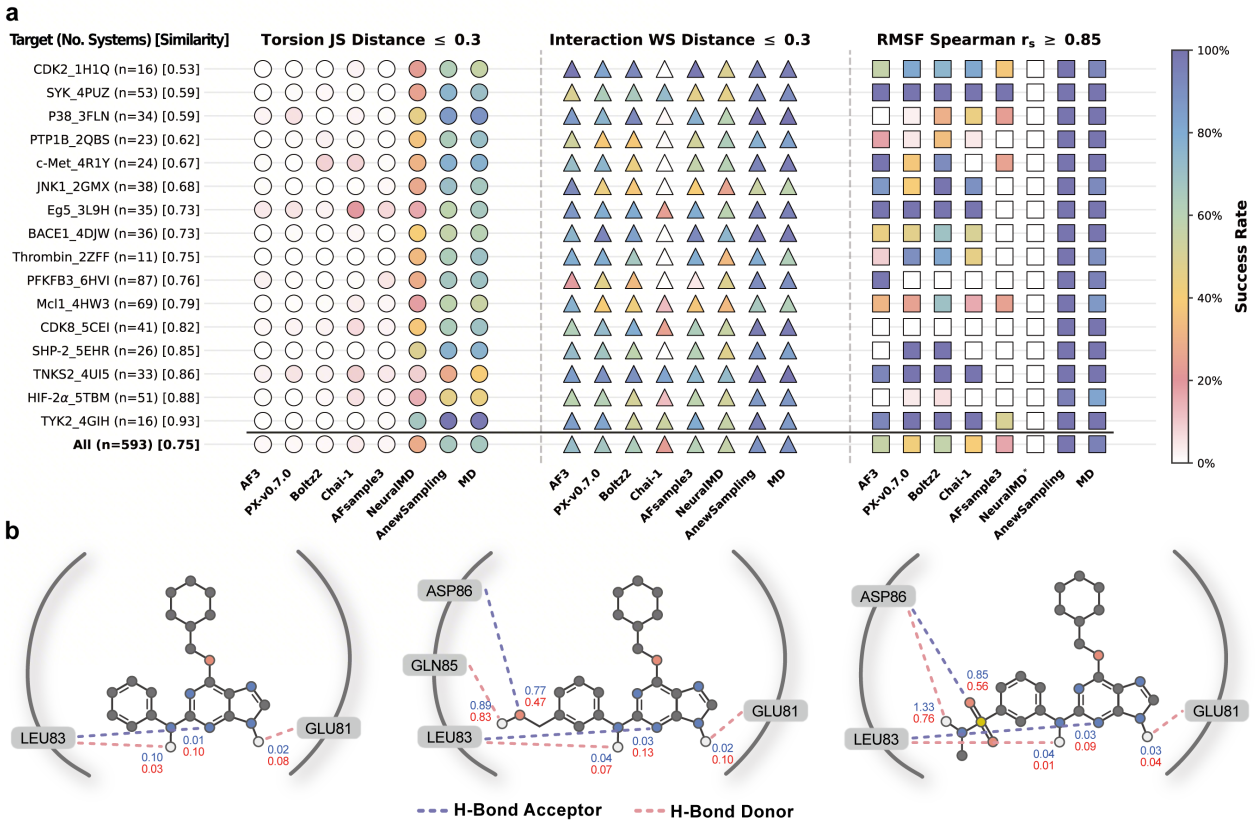

Figure 3 Evaluation of generative models for protein-ligand conformational ensemble generation on JACS and Merck datasets. (a) The heatmaps visualize the comparison between generative models and standard MD simulations in the success rates of ligand torsion JS distance ≤ 0.3 (left), protein-ligand highly conserved interaction (frequency ≥ 0.8) WS distance ≤ 0.3 Å (middle), and protein backbone Cα RMSF Spearman correlation ≥ 0.85 (right). The color gradient reflects the success rate of each method, scaling from 0% (white) to 100% (dark blue). Each row corresponds to a specific protein target, arranged by its median pocket-SuCos similarity relative to the training set. Beside each target name and PDB ID, the number of evaluated ligands is indicated in parentheses, followed by the median Pocket-SuCos similarity in square brackets. *Note that NeuralMD treats proteins as rigid stuctures, its corresponding RMSF values are uniformly zero. **(b)** Example of CDK2 ligands showing protein-ligand hydrogen bonds induced by phenyl ring substitutions and comparison in WS distances between MD (blue) and AnewSampling (red)

.

group orientations. In standard MD, the system remains trapped in these initial states for extended periods before converging to a more stable pose, artificially blending two distinct sampling distributions within the production trajectory. Conversely, REMD rapidly relaxes into a stable pose during equilibration, leading to a cleaner distribution in the production phase.

While fundamentally incapable of modeling dynamic ligand torsions, static co-folding models reproduced highly conserved protein-ligand interactions (frequency ≥ 0.8) reasonably well (Fig. 3a middle and Table S7), with AF3 performing the best among these methods. Although NeuralMD possesses some ligand conformational sampling capability, its interaction sampling performance was mediocre, likely bottlenecked by its rigid treatment of the protein. AnewSampling once again outperformed all other generative methods, achieving a remarkable success rate that closely mirrored the REMD baseline. This underscores AnewSampling's proficiency in capturing subtle interaction shifts between congeneric ligands and the protein pocket. For instance, in the CDK2 system, which exhibited the lowest median Pocket-SuCos similarity (0.53), AnewSampling successfully captured the changes in pocket hydrogen bonding induced by phenyl ring substitutions while maintaining stable interactions in conserved regions (Fig. 3b). Furthermore, standard MD exhibited a high WS distance for the hydrogen bond between the sulfonamide group of molecule 28 and Asp, caused by insufficient sampling of both the ligand moiety and the adjacent pocket residues. AnewSampling's ability to match the REMD distribution in this instance indicates a strong potential for enhanced sampling—a capability we discuss with examples in Section 2.4.

Beyond the accurate sampling of ligand torsions and pocket interactions, AnewSampling substantially surpassed other methods in predicting global protein dynamics, achieving the highest success rate for the protein Cα RMSF Spearman correlation ($r_s \geq 0.85$) across all systems (Fig. 3a right and Table S8). This demonstrates that AnewSampling maintains a high degree of biophysical realism in its global structural sampling. Finally, it is worth noting that because NeuralMD employs a strictly rigid protein representation, its calculated protein RMSF values are inherently zero, precluding it from capturing these dynamic fluctuations.

2.4 Enhanced Sampling Beyond Conventional MD

AnewSampling not only samples conformational distributions consistent with standard MD, but also explores regions that are difficult to reach within typical simulation budgets (approximately 1–100 ns). In such settings, enhanced-sampling methods such as REMD are often required to escape trapped local minima separated by high free-energy barriers. To demonstrate the capability of AnewSampling in navigating complex energy landscapes, we examined two ligands of the CDK2 receptor. CDK2 is a key member of the cyclin-dependent kinase family important in the regulation of the cell proliferation and RNA polymerase II transcription [31]. It has emerged as a critical target for tumor-selective therapy, particularly in CCNE1-amplified cancers, where its inhibition exploits synthetic lethality to eliminate tumor cells while sparing healthy tissues [35, 36]. Two of the protein ligand complexes for CDK2, 1h1r and 1h1s, are representative case necessitating REMD enhanced sampling to capture its multiple distinct binding poses [37]. As shown in the top penel of Figure 4, ligand 1h1r occupies two distinct binding poses characterized by a significant flip of the benzene ring, supported by the electron density of 1h1r in X-ray crystallography. While is challenging for standard MD escape the local minima of binding pose 1 of 1h1r, AnewSampling exhibits exceptional sampling capability by capturing both binding poses in agreement with the REMD simulations.

Furthermore, we emphasize that accurate binding mode prediction often requires the coupled sampling of both ligand and protein degrees of freedom. This is exemplified by the ligand 1h1s as illustrated in the bottom panels of Figure 4, where the transition between binding modes involves not only the rotation of the ligand's sulfonamide group but also the necessary reorganization of pocket residues (specifically Gln and Asp) to establish key hydrogen bond networks. Standard 10ns MD simulations typically fail to overcome the associated energy barriers, missing the critical Gln hydrogen bonds entirely. However, AnewSampling effectively captures these coupled side-chain fluctuations and recovers these transient hydrogen bonds with a fidelity comparable to REMD. This ability to accurately sample coupled protein-ligand fluctuations is vital for identifying bioactive conformations and improving the precision of activity predictions from both energetic and entropic perspectives. Although AnewSampling does not yet fully replicate the extensive configurational space exploration of REMD, it demonstrates remarkable enhanced sampling potential given that it was trained exclusively on standard MD data, which itself fails to capture these rare events. This suggests that the model

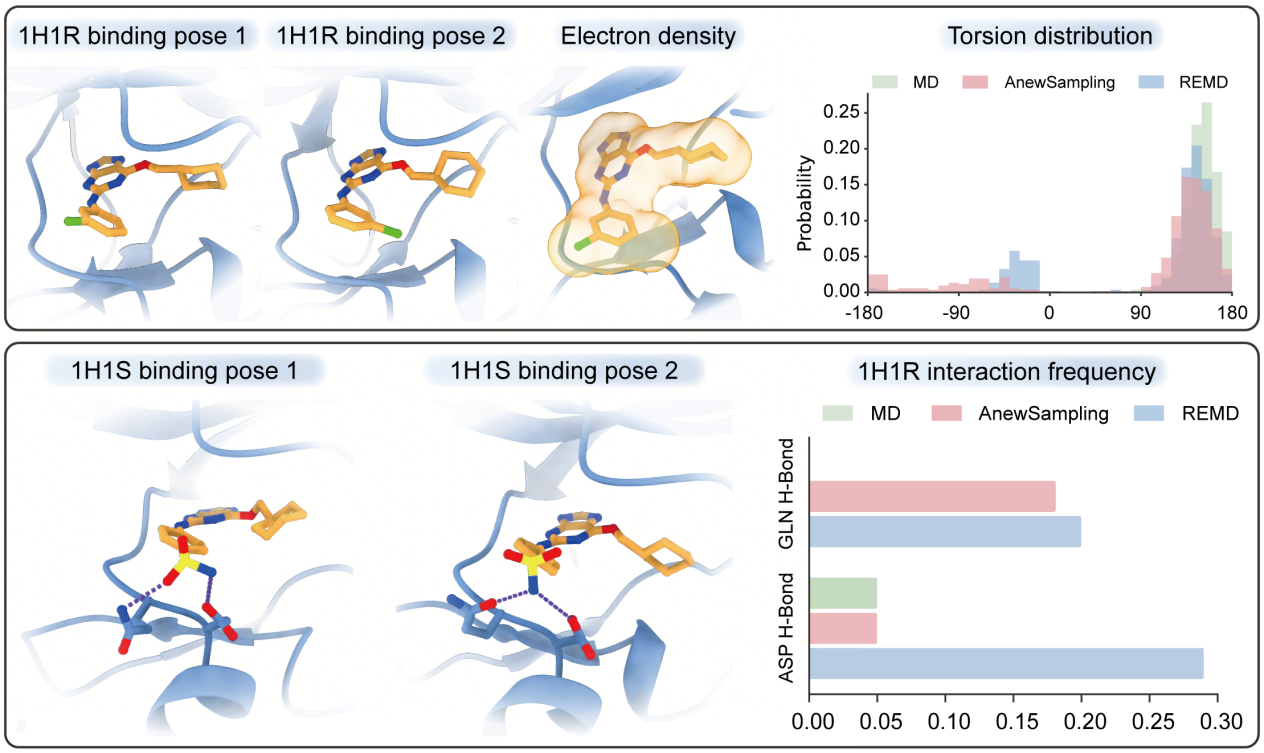

Figure 4 Enhanced conformational sampling in CDK2 complexes. (Top) The left panels illustrate two distinct binding poses of ligand 1h1r sampled from REMD and AnewSampling, alongside the experimental electron density of ligand 1h1r confirming the two alternative poses. The torsion distribution analysis (top right) demonstrates that conventional MD (green) is restricted to a single conformational basin (pose 1). In contrast, AnewSampling (pink) shows the potential of capturing the multimodal distribution observed in the REMD trajectory (blue). (Bottom) Visualization of two distinct 1h1s binding poses (left), highlighting the sulfonamide rotation and associated Gln/Asp side chain rearrangements required for hydrogen bond formation. Frequency analysis of the key interactions (right) confirms the ability of AnewSampling (pink) to capture the critical transient hydrogen bonds observed in REMD trajectory (blue), which are missed by conventional MD simulations (green).

effectively learns the underlying energy landscape rather than merely memorizing the limited conformational diversity of its training set.

3 Discussion

We have introduced AnewSampling, a generative foundation framework developed to sample the all-atom equilibrium distributions of biomolecular complexes. The framework has been demonstrated to achieve rigorous statistical alignment with ground-truth MD. And the generated ensembles yield evaluation metrics that are indistinguishable from the intrinsic variance observed between independent MD replicates. Statistically, the model establishes a new baseline by consistently outperforming all generative methods across every evaluation metric on the ATLAS monomer benchmark. Broadening the scope to protein-ligand dynamics, the framework accurately characterizes the precise atomic interactions. By capturing these fundamental interactions, the model faithfully recovers coupled ligand and side-chain motions within challenging targets such as CDK2 complexes, which are essential for industrial structure-activity relationship analysis. The computational inference operates at a fraction of the cost required for conventional MD simulations, effectively bypassing the high energy barriers that typically trap traditional sampling methods.

Despite the significant advancements introduced by AnewSampling, several limitations warrant further investigation in future research. Although the underlying architecture natively supports diverse biomolecular modalities, the primary constraint remains the scarcity of comprehensive training data. While we have

curated the largest database to date, this database is insufficient to capture the full landscape of biological complexity. Future iterations will expand the framework to support a broader range of interaction types, such as protein-nucleic acid or protein-protein interactions. The current framework relies on the incorporation of templates. While this approach aligns with the standard inputs required for conventional MD, and the framework technically supports sequence-only inputs, the absence of templates results in diminished performance. The ultimate objective is to derive highly accurate equilibrium distributions solely from primary sequences. We observe that templates provide marginal benefits for simple monomeric systems; however, for complexes, the challenge becomes more pronounced. As observed in models similar to AlphaFold 3, accurately predicting the static structures of such complexes is inherently difficult. Extending this to the task of distribution prediction further exacerbates the challenge. Consequently, scaling the training data and refining the architectural structures represent critical directions for future work.

The current framework learns thermodynamic distributions within a single, fixed thermodynamic environment. In contrast, conventional MD can be naturally extended to model varying macroscopic conditions, although it still encounter severe sampling bottlenecks. Our AnewSampling framework and conventional MD are fundamentally complementary. MD provides the essential raw training data that underpins the generative system. As previously noted, by learning fundamental physical interactions across diverse chemical systems, AnewSampling demonstrates an emergent capacity for enhanced sampling. Conversely, the generative framework can accelerate MD by providing a diverse and robust set of initial structural candidates. By utilizing these generated conformations as starting points, traditional physical simulations can efficiently bypass potential energy barriers and explore the broader conformational space with unprecedented thoroughness.

4 Method

4.1 Datasets

We constructed the AnewSampling-DB dataset, which represents a significant advancement in both scale and physical fidelity, by filtering and curating data from PDBBind v2020 [38], ChEMBL [39], JACS and Merck [30]. The dataset comprises 31,364 complexes, substantially outperforming the MISATO dataset [40] (15,592 complexes) while covering a broad chemical space of 10,297 unique protein sequences and 27,979 unique ligand SMILES. Building upon the logic of the Bindingnet v2 hierarchical template-based modeling approach [41], we employed Glide [42] to prepare ∼20,000 complexes from ChEMBL [39] to enrich our dataset with congeneric series. The inclusion of congeneric series captures the subtle differences in protein-ligand interactions between structurally similar ligands with distinct bioactivities binding to identical protein targets. The dataset exhibits robust distributional characteristics across key structural metrics. The torsion number distribution (Fig. 5B) shows a unimodal peak between 4 and 6 rotatable bonds while extending to highly flexible ligands, encapsulating over 43,000 distinct force field torsion types. Token and sequence number distributions further highlight the dataset's coverage of diverse system sizes.

To capture the complexity of molecular recognition, the dataset balances a wide array of non-covalent interaction types. The overall interaction landscape is dominated by hydrophobic and van der Waals (vdW) contacts, which collectively account for 67.7% of the total interactions. Focusing on the specific polar and aromatic interactions (excluding hydrophobic and vdW contacts), Fig. 5A illustrates a distribution primarily driven by hydrogen bonding, where H-bond acceptors and donors constitute 36.9% and 27.8% of this subset, respectively. This specific interaction profile is further enriched by essential electrostatic and aromatic contacts, including salt bridges (16.7%), π-stacking (12.4%), and π-ionic interactions (6.3%).

To ensure high data quality and physical consistency across this expansive library, all molecular dynamics (MD) trajectories were generated via a unified pipeline employing a consistent set of protein and ligand force fields. Furthermore, REMD simulations were utilized to rigorously explore the conformational landscape more efficiently. Detailed simulation protocols are provided in the Appendix.

4.2 Evaluation Metrics

To systematically quantify generative performance, we established strict success thresholds grounded in physical intuition and empirical reference distributions (Fig. 5c). For intrinsic ligand flexibility in the binding

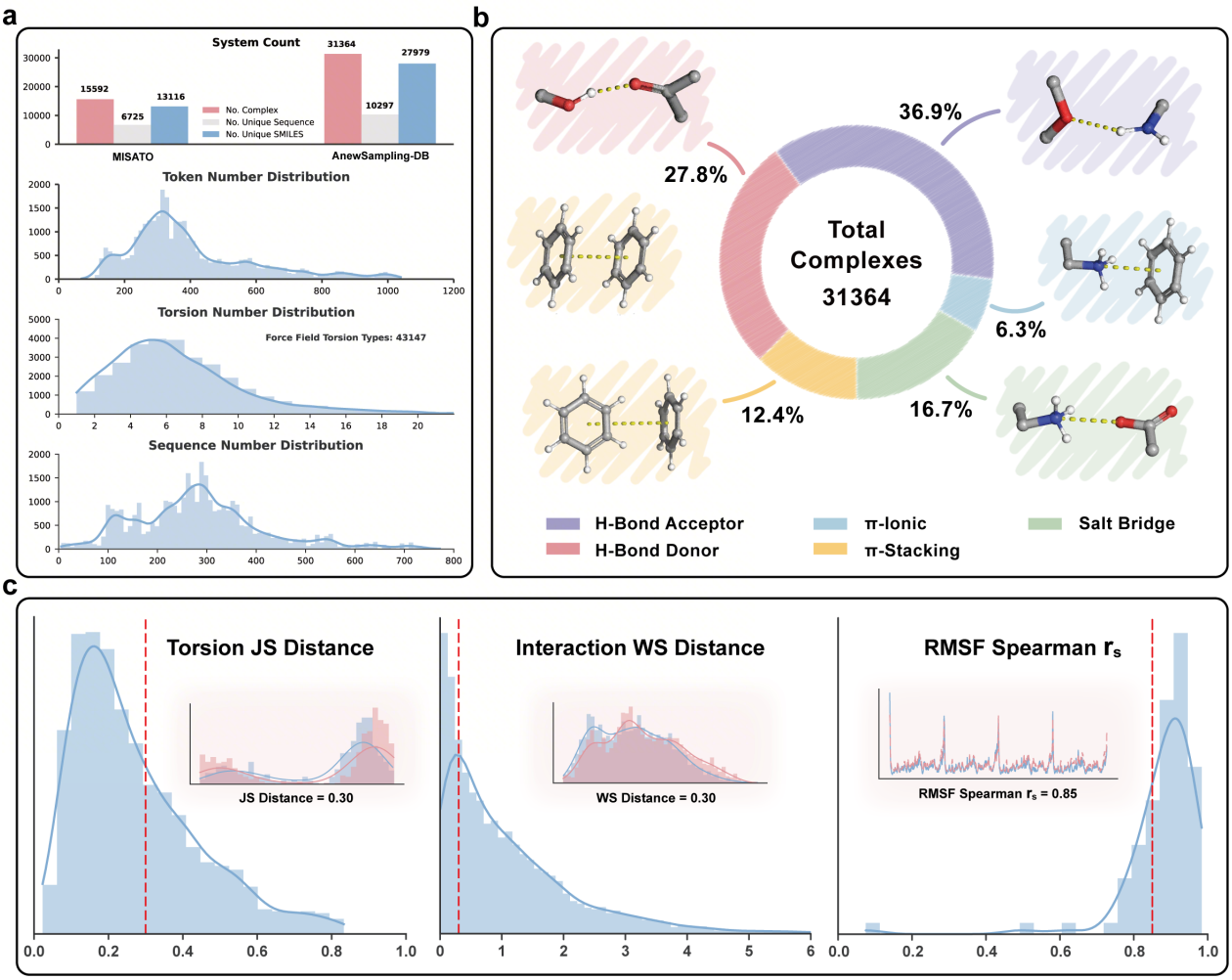

Figure 5 Statistical overview of the AnewSampling-DB dataset and Evaluation Metrics. (a) Classification of polar and aromatic interaction types within the 31,364 complexes. The percentage breakdown details specific contacts (e.g., H-bonds, π-stacking), excluding hydrophobic and Van der Waals interactions which constitute the majority (67.7%) of the total profile. **(b)** Dataset scale comparison against the MISATO benchmark, highlighting the increase in unique protein sequences and ligand SMILES. Subplots display the distributions of token counts, torsion numbers (covering 43147 types), and sequence lengths. **(c)** Examples showing the evaluation metrics of torsional JS distance (JSD), interaction WS distance, and protein Cα RMSF.

pocket, we defined successful recovery as a JS distance of ≤ 0.30 for torsion angle distributions, a threshold representing highly overlapping conformational states. To evaluate the precise stability of protein-ligand contact networks without relying on rigid-body superposition, we set a stringent WS distance threshold of ≤ 0.30 Å. Finally, to ensure the preservation of coherent global protein backbone dynamics, we required a Spearman correlation (r_s) of ≥ 0.85 for the RMSF. These rigorously defined metrics ensure that generated ensembles are evaluated not just on static geometric similarity, but on true thermodynamic and dynamic fidelity.

References

[1] Maolin Lu and H Peter Lu. Revealing multiple pathways in t4 lysozyme substep conformational motions by single-molecule enzymology and modeling. The Journal of Physical Chemistry B, 121(19):5017–5024, 2017.

[2] Arthur G Palmer III. Nmr characterization of the dynamics of biomacromolecules. Chemical reviews, 104(8): 3623–3640, 2004.

[3] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792):706–710, 2020.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. Nature, 596(7873):583–589, 2021.

[5] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. Biorxiv, pages 2021–10, 2021.

[6] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, 630(8016):493–500, 2024.

[7] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. Science, 384(6693):eadl2528, 2024.

[8] ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, et al. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. BioRxiv, pages 2025–01, 2025.

[9] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. BioRxiv, pages 2024–11, 2025.

[10] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. BioRxiv, 2025.

[11] Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. BioRxiv, pages 2024–10, 2024.

[12] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. Nature, 450(7172):964–972, 2007.

[13] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. How cryo-em is revolutionizing structural biology. Trends in biochemical sciences, 40(1):49–57, 2015.

[14] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. Nature, 253(5494):694–698, 1975.

[15] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. Science, 334(6055):517–520, 2011.

[16] Diego Del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with alphafold2. elife, 11:e75751, 2022.

[17] Hannah K Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and alphafold2. Nature, 625(7996):832–839, 2024.

[18] Yogesh Kalakoti and Björn Wallner. Afsample3: Generating and selecting multiple conformational states with alphafold3. bioRxiv, pages 2026–01, 2026.

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

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

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

[22] Shengchao Liu, Weitao Du, Hannan Xu, Yanjing Li, Zhuoxinran Li, Vignesh Bhethanabotla, Divin Yan, Christian Borgs, Anima Anandkumar, Hongyu Guo, et al. A multi-grained symmetric differential equation model for learning protein-ligand binding dynamics. Nature Communications, 2025.

[23] Bin Feng, Jiying Zhang, Xinni Zhang, Zijing Liu, and Yu Li. Biomd: All-atom generative model for biomolecular dynamics simulation. arXiv preprint arXiv:2509.02642, 2025.

[24] Bin Feng, Jiying Zhang, Xini Zhang, Ming Zhang, Patrick Barth, Zijing Liu, and Yu Li. Physically grounded generative modeling of all-atom biomolecular dynamics. bioRxiv, 2026. doi: 10.64898/2026.02.15.705956. URL https://www.biorxiv.org/content/early/2026/02/15/2026.02.15.705956.

[25] 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.

[26] Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, et al. Predicting equilibrium distributions for molecular systems with deep learning. Nature Machine Intelligence, 6(5):558–567, 2024.

[27] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. Science, 389(6761):eadv9817, 2025.

[28] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Ai-based methods for simulating, sampling, and predicting protein ensembles. arXiv preprint arXiv:2509.17224, 2025.

[29] Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. Nucleic Acids Research, 52(D1): D384–D392, 2024.

[30] Gregory A. Ross, Chao Lu, Guido Scarabelli, Steven K. Albanese, Evelyne Houang, Robert Abel, Edward D. Harder, and Lingle Wang. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. Communications Chemistry, 6(1):222, 2023. ISSN 2399-3669.

[31] Ian R Hardcastle, Christine E Arris, Johanne Bentley, F Thomas Boyle, Yuzhu Chen, Nicola J Curtin, Jane A Endicott, Ashleigh E Gibson, Bernard T Golding, Roger J Griffin, Philip Jewsbury, Jerome Menyerol, Veronique Mesguiche, David R Newell, Martin E M Noble, David J Pratt, Lan-Zhen Wang, and Hayley J Whitfield. N2-substituted o6-cyclohexylmethylguanine derivatives: potent inhibitors of cyclin-dependent kinases 1 and 2. J Med Chem, 47(15):3710–3722, 2004.

[32] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In Forty-first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id=rs8Sh2UASt.

[33] Anonymous. BioMD: All-atom generative model for biomolecular dynamics simulation. In The Fourteenth International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=LQDeJk6NOr.

[34] Isomorphic Labs Team. Accurate predictions of novel biomolecular interactions with isodde, February 2026. URL https://doi.org/10.5281/zenodo.18606681.

[35] George Au-Yeung, Franziska Lang, Walid J Azar, Chris Mitchell, Kate E Jarman, Kurt Lackovic, Diar Aziz, Carleen Cullinane, Richard B Pearson, Linda Mileshkin, Danny Rischin, Alison M Karst, Ronny Drapkin, Dariush Etemadmoghadam, and David D L Bowtell. Selective targeting of cyclin E1-Amplified High-Grade serous ovarian cancer by Cyclin-Dependent kinase 2 and AKT inhibition. Clin Cancer Res, 23(7):1862–1874, 2016.

[36] Solomon Tadesse, Abel T Anshabo, Neil Portman, Elgene Lim, Wayne Tilley, C Elizabeth Caldon, and Shudong Wang. Targeting CDK2 in cancer: challenges and opportunities for therapy. Drug Discov Today, 25(2):406–413, 2019.

[37] Lingle Wang, Yuqing Deng, Jennifer L. Knight, Yujie Wu, Byungchan Kim, Woody Sherman, John C. Shelley, Teng Lin, and Robert Abel. Modeling local structural rearrangements using fep/rest: Application to relative binding affinity predictions of cdk2 inhibitors. Journal of Chemical Theory and Computation, 9(2):1282–1293, 2013.

[38] Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. Accounts of Chemical Research, 50(2):302–309, 2017.

[39] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michal Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. Chembl: towards direct deposition of bioassay data. Nucleic acids research, 47:D930—D940, 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1075.

[40] Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Kieran Didi, André Santos Dias Mourão, Radosław Kitel, Pietro Liò, Stefan Kesselheim, Marie Piraud, Fabian J. Theis, Michael Sattler, and Grzegorz M. Popowicz. Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. Nature Computational Science, 4(5):367–378, 2024. ISSN 2662-8457.

[41] Hui Zhu, Xuelian Li, Baoquan Chen, and Niu Huang. Augmented bindingnet dataset for enhanced ligand binding pose predictions using deep learning. npj Drug Discovery, 2(1):1, 2025. ISSN 3005-1452.

[42] Richard A. Friesner, Robert B. Murphy, Matthew P. Repasky, Leah L. Frye, Jeremy R. Greenwood, Thomas A. Halgren, Paul C. Sanschagrin, and Daniel T. Mainz. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. Journal of Medicinal Chemistry, 49(21):6177–6196, 2006.

[43] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. Biophysical Journal, 109(8):1528–1532, 2015. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2015.08.015. URL https://www.sciencedirect.com/science/article/pii/S0006349515008267.

[44] Justin R. Porter, Katelyn E. Moeder, Carrie A. Sibbald, Maxwell I. Zimmerman, Kathryn M. Hart, Michael J. Greenberg, and Gregory R. Bowman. Cooperative changes in solvent exposure identify cryptic pockets, switches, and allosteric coupling. Biophysical Journal, 116(5):818–830, 2019. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2018.11.3144. URL https://www.sciencedirect.com/science/article/pii/S0006349519300530.

[45] Cédric Bouysset and Sébastien Fiorucci. Prolif: a library to encode molecular interactions as fingerprints. Journal of Cheminformatics, 13(1):72, 2021. doi: 10.1186/s13321-021-00548-6. URL https://doi.org/10.1186/s13321-021-00548-6.

[46] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX, 1-2:19–25, 2015.

[47] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics, 79(2): 926–935, 07 1983. ISSN 0021-9606.

[48] In Suk Joung and Thomas E. III Cheatham. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. The Journal of Physical Chemistry B, 112(30):9020–9041, 2008.

[49] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. Journal of Chemical Theory and Computation, 11(8):3696–3713, 2015.

[50] Tianze Zheng, Ailun Wang, Xu Han, Yu Xia, Xingyuan Xu, Jiawei Zhan, Yu Liu, Yang Chen, Zhi Wang, Xiaojie Wu, Sheng Gong, and Wen Yan. Data-driven parametrization of molecular mechanics force fields for expansive chemical space coverage. Chem. Sci., pages 2730–2740, 2025.

[51] Lingle Wang, Richard A. Friesner, and B. J. Berne. Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (rest2). The Journal of Physical Chemistry B, 115(30):9431–9438, 2011. doi: 10.1021/jp204407d.

[52] Alexandra Patriksson and David van der Spoel. A temperature predictor for parallel tempering simulations. Phys. Chem. Chem. Phys., 10:2073–2077, 2008.

[53] N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink, and H. J. C. Berendsen. Efficient algorithms for langevin and dpd dynamics. Journal of Chemical Theory and Computation, 8(10):3637–3649, 2012.

[54] Mattia Bernetti and Giovanni Bussi. Pressure control using stochastic cell rescaling. The Journal of Chemical Physics, 153(11):114107, 09 2020. ISSN 0021-9606.

[55] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. Journal of Computational Chemistry, 18(12):1463–1472, 1997.

[56] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An $n \cdot \log(n)$ method for ewald sums in large systems. The Journal of Chemical Physics, 98(12):10089–10092, 06 1993. ISSN 0021-9606.

[57] Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. Mdanalysis: A toolkit for the analysis of molecular dynamics simulations. Journal of Computational Chemistry, 32(10):2319–2327, 2011.

[58] Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domański, David L. Dotson, Sébastien Buchoux, Ian M. Kenney, and Oliver Beckstein. Mdanalysis: A python package for the rapid analysis of molecular dynamics simulations. SciPy 2016, 2016.

[59] Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology, 13(7):1–17, 07 2017.

[60] Matthew P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J.F. Day, Barry Honig, David E. Shaw, and Richard A. Friesner. A hierarchical approach to all-atom protein loop prediction. Proteins: Structure, Function, and Bioinformatics, 55(2):351–367, 2004.

[61] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. Journal of Computer-Aided Molecular Design, 27(3):221–234, 2013. ISSN 1573-4951.

[62] Ryne C. Johnston, Kun Yao, Zachary Kaplan, Monica Chelliah, Karl Leswing, Sean Seekins, Shawn Watts, David Calkins, Jackson Chief Elk, Steven V. Jerome, Matthew P. Repasky, and John C. Shelley. Epik: pka and protonation state prediction through machine learning. Journal of Chemical Theory and Computation, 19(8): 2380–2388, 2023.

[63] Chresten R. Søndergaard, Mats H. M. Olsson, Michał Rostkowski, and Jan H. Jensen. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pka values. Journal of Chemical Theory and Computation, 7(7):2284–2295, 2011.

[64] Peter Škrinjar, Jérôme Eberhardt, Gerardo Tauriello, Torsten Schwede, and Janani Durairaj. Have protein-ligand cofolding methods moved beyond memorisation? bioRxiv, 2025. doi: 10.1101/2025.02.03.636309. URL https://www.biorxiv.org/content/early/2025/08/04/2025.02.03.636309.

Appendix

A Methods

A.1 Evaluation Metrics

ATLAS Evaluation Metircs: Following the benchmarking frameworks established by AlphaFlow [32] and BioMD [33], we evaluated AnewSampling using the ATLAS dataset [29] to perform ensemble analysis (predicting flexibility, distributional accuracy, and ensemble observables). Prior to analysis, all ensembles were aligned to the initial static all-atom structure. Computations were performed using Cartesian coordinates via MDTraj [43], with trajectories subsampled to 1,000 frames to ensure computational efficiency. The 2-Wasserstein distance between two 3D Gaussian distributions is defined as:

$$\mathcal{W}_2^2(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\right) \tag{1}$$

Notably, this metric generalizes the Euclidean distance metric used for point masses in AlphaFlow. The aggregate Root Mean Squared Wasserstein Distance (RMWD) can thus be decomposed into distinct translational and variance components:

$$\mathrm{RMWD}^2(\mathcal{X}_1, \mathcal{X}_2) = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\|\mu_{1,i} - \mu_{2,i}\|^2}_{(\text{translation contribution})^2} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathrm{Tr}\left(\Sigma_{1,i} + \Sigma_{2,i} - 2(\Sigma_{1,i}\Sigma_{2,i})^{1/2}\right)}_{(\text{variance contribution})^2} \tag{2}$$

Quantitative results for these contributions are detailed in Table S1. For joint $\mathcal{W}_2$ distance calculations, we projected the data into a PCA subspace to mitigate convergence issues arising from high-dimensional thermal fluctuations. To avoid biasing the projection solely towards the reference simulation, which may obscure orthogonal deviations in the predicted ensemble, we performed PCA using both the MD reference ensemble alone and a pooled ensemble (equal weighting of MD and predicted frames). Finally, sidechain solvent-accessible surface area (SASA) was computed using the Shrake-Rupley algorithm (probe radius: 2.8 Å), with residues classified as buried if their SASA fell below 2.0 $Å^2$ [44].

Ligand Conformational Fidelity via Torsional JS Distance: To assess how accurately our model explores the internal degrees of freedom of the ligand within the binding pocket, we utilize the torsional JS distance (D_{JS}). We compute the probability distribution of all rotatable dihedrals (ϕ) for the ligand across the generated ensemble P and the MD reference ensemble Q:

$$D_{JS}(P \parallel Q) = \sqrt{\frac{1}{2}D_{KL}(P \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M)}$$

where $M = \frac{1}{2}(P + Q)$. This metric quantifies the divergence between predicted and ground-truth torsional preferences, ensuring that the model captures physical rotameric states rather than merely predicting averaged geometries. For each rotatable bond, we discretize the dihedral angle space into $N = 36$ bins (bin size = 10°) to construct a discrete probability distribution of observed orientations.

Interaction Distance Distribution Difference via WS Distance: The stability of a protein-ligand complex is governed by its interaction network (e.g., hydrogen bonds, π-stacking, and salt bridges). To evaluate the sampling of these non-bonded interactions, we characterized the dynamic protein-ligand interaction landscape using the ProLIF toolkit [45] across 500 frames. We calculate the WS distance (W_1) between the distributions of specific protein-ligand atomic distances:

$$W_1(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| \mathrm{d}x$$

The W_1 metric (or "Earth Mover's Distance") provides a continuous measure of how much "work" is required to transform the predicted interaction profile into the MD-derived profile. Crucially, the calculation accounts

for the inherent symmetry of ligands, residues, and protein chains; equivalent atoms and subunits are permuted to minimize the distance, thereby preventing structural degeneracies from inflating the error metric. By integrating across the entire ensemble, this metric penalizes both the omission of critical binding modes and the generation of energetically unfavorable clashes.

Global Protein Dynamics via Cα RMSF: To determine if the generative model recovers the intrinsic plasticity of the protein scaffold, we compare the Cα RMSF across multiple diverse systems.

$$\text{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_{i,t} - \bar{x}_i)^2}$$

By correlating the predicted RMSF profiles with MD benchmarks across a broad test set, we demonstrate the model's ability to distinguish between rigid secondary structures and disordered loops. This global assessment ensures that the ligand-binding event is coupled with realistic protein transitions, maintaining the biophysical integrity of the holo-structure.

A.2 Simulation Protocol

The GROMACS 2023 package [46] was used to conduct the all-atom MD simulations. The ligand-protein complexes were prepared by placing the molecules at the center of triclinic boxes with a minimum distance of 1.0 nm to the box boundary. The TIP3P water model [47] was used to solvate the complexes. The system was neutralized and brought to a physiological ionic strength of 0.15 M by adding sodium and chloride ions [48]. The proteins were parameterized with Amber ff14SB [49] and ligands were parameterized with the ByteFF [50]. The relaxation of each system started with energy minimization of 50,000 steps, followed by 400 ps relaxation in the NVT ensemble with a 1fs time step to increase the temperature to 298.15K and 100 ps relaxation in the NPT ensemble at 298.15 K and a pressure of 1 bar. The protein backbone atoms and the ligand heavy atoms were restrained by harmonic position restraints with a force constant of 1000 kJ mol^{-1}nm^{-2} during relaxation. Finally, the systems were further equilibrated without any restraints for 2 ns in the NPT ensemble with a 2 fs time step before collecting 10 ns production data with saving trajectory every 20 ps. For the REMD simulations, we adopted the replica exchange with solute tempering (REST2) [51] enhanced sampling scheme with all the ligand atoms selected as the hot region. A total of 16 λ windows were used for each REMD simulation to scale the nonbonded and torsional interactions involving the hot region. The alchemical schedule of effective temperature of the hot region was estimated with the expected acceptance ratio of replica exchange set to be 0.3 [52].

The temperature in the simulations was controlled using a Langevin integrator [53] with the friction coefficient set to 0.5 ps^{-1}. The stochastic cell rescaling algorithm [54] was applied for pressure coupling with a time constant of 2 ps. All the bonds with hydrogen atoms were constrained using the LINCS algorithm [55]. Periodic boundary conditions were applied, and the particle mesh-Ewald method [56] was used to treat long-range electrostatic interactions with a direct space cutoff of 1.0 nm. The Lennard-Jones interactions were switched off at 0.8 nm and a dispersion correction was applied for energy and pressure. The analysis of MD trajectories was performed using MDAnalysis [57, 58] and ProLIF [45].

A.3 Structural Refining Protocol

Our structural refining protocol emphasized the preservation of biological and chemical integrity. We employed PDBFixer [59] and Schrödinger Prime [60] to model missing loops and non-standard amino acids. To preserve the structural integrity of protein chains, we conducted global alignments between the sequences derived from the PDB and their respective RefSeq identities. Protein-ligand complexes were prepared using Schrödinger protein preparation wizard [61]. Protonation states were generated with Epik [62] then optimized with PROPKA [63].

A.4 Quantification of system similarity

We used a combined similarity measure to carefully test how well our model could generalize to different, out-of-distribution targets. This tool measures the distance between different protein-ligand complexes by

taking into account both ligand- and pocket-centric factors. It is based on the stratification method set up in the IsoDDE report [34]. This particular metric, called Pocket-SuCos [64], blends a ligand-focused SuCos score—which only looks at three-dimensional volumetric overlap and physicochemical feature similarity (feature score)—with a pocket query coverage (pocket_qcov) score to make sure that the binding site stays the same.

A.5 Baselines and Implementation Details

All benchmark evaluations were conducted on a local high-performance computing cluster using a standardized hardware baseline to ensure consistency in the inference environment. To ensure a rigorous and fair comparison, each baseline model was deployed using its official local repository, with input protocols optimized to match the information density of AnewSampling:

- **Protenix (PX-v0.7.0)**: We employed the official base version of Protenix. Similar to other co-folding models, the inference script was adapted to accept 3D ligand conformations (SDF) as spatial hints, ensuring the model had access to the same initial geometric information as AnewSampling.
- **Boltz2**: The official Boltz2 open-source codebase was used for local inference. The input protocol was modified to accept SDF conformations. During evaluation on the held-out test set, 1 system failed due to Out-of-Memory (OOM) errors and was subsequently excluded from the aggregate statistics.
- **Chai-1**: We utilized the official Chai-1 repository. The input logic was similarly modified to support 3D SDF conformations. Due to its substantial memory footprint when processing complex targets in the held-out test set, 9 systems encountered OOM errors; these cases were omitted from the final performance metrics.
- **NeuralMD**: This model was initialized with the starting frame of each system and performed continuous sampling for 500 steps. It serves as a dedicated baseline for evaluating the propagation of local molecular dynamics from a known starting point.

Sampling Strategy and Ensemble Construction: To ensure a direct comparison with the ensemble generation capabilities of AnewSampling, we applied a uniform enhanced sampling protocol to all co-folding baselines (Protenix, Boltz2, and Chai-1). For each case in the test set, we executed the models using 5 distinct random seeds, with each seed generating 100 independent conformations. The resulting 500 frames were aggregated to construct the structural ensemble for each model, which was then used to calculate distribution-based metrics, including JS distance of ligand torsions, WS distance of interaction pairs, and RMSF of protein Cα.

Table S1 Statistical summary of JS distances for ligand torsion angle distributions on the held-out testset. The table provides the mean, median, standard deviation, and cumulative success rates (binned from ≤ 0.00 to ≤ 0.55) of JS distances for conformations generated by the co-folding models, MD-trained generative models, and the MD baseline.

	AF3	PX-v0.7.0	Boltz2	Chai-1	AFsample3	NerualMD	AnewSampling	MD
JS distance(↓)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)
mean=	0.5441	0.5355	0.5344	0.5492	0.5216	0.4152	0.2744	**0.2726**
median=	0.5537	0.5317	0.5297	0.5576	0.5275	0.4055	0.2402	0.2251
std=	0.1551	0.1514	0.1439	0.1484	0.1506	0.1441	0.1439	0.1677
≤0.00	0/1287(0.0000)	0/1287(0.0000)	0/1276(0.0000)	0/1105(0.0000)	0/1287(0.0000)	0/1287(0.0000)	0/1287(0.0000)	0/1287(0.0000)
≤0.05	0/1287(0.0000)	0/1287(0.0000)	0/1276(0.0000)	0/1105(0.0000)	0/1287(0.0000)	0/1287(0.0000)	2/1287(0.1554)	**7/1287(0.5439)**
≤0.10	0/1287(0.0000)	0/1287(0.0000)	0/1276(0.0000)	1/1105(0.0905)	0/1287(0.0000)	1/1287(0.0777)	69/1287(5.3613)	**134/1287(10.4118)**
≤0.15	2/1287(0.1554)	2/1287(0.1554)	0/1276(0.0000)	5/1105(0.4525)	1/1287(0.0777)	15/1287(1.1655)	235/1287(18.2595)	**332/1287(25.7964)**
≤0.20	16/1287(1.2432)	11/1287(0.8547)	7/1276(0.5486)	11/1105(0.9955)	19/1287(1.4763)	57/1287(4.4289)	461/1287(35.8197)	**552/1287(42.8904)**
≤0.25	49/1287(3.8073)	33/1287(2.5641)	28/1276(2.1944)	32/1105(2.8959)	57/1287(4.4289)	148/1287(11.4996)	677/1287(52.6030)	**708/1287(55.0117)**
≤**0.30**	98/1287(7.6146)	72/1287(5.5944)	72/1276(5.6426)	63/1105(5.7014)	104/1287(8.0808)	289/1287(22.4553)	**851/1287(66.1228)**	842/1287(65.4235)
≤0.35	157/1287(12.1989)	155/1287(12.0435)	132/1276(10.3448)	117/1105(10.5882)	172/1287(13.3644)	466/1287(36.2082)	**969/1287(75.2914)**	943/1287(73.2712)
≤0.40	233/1287(18.1041)	253/1287(19.6581)	233/1276(18.2602)	189/1105(17.1041)	286/1287(22.2222)	622/1287(48.3294)	**1060/1287(82.3621)**	1019/1287(79.1764)
≤0.45	355/1287(27.5835)	402/1287(31.2354)	367/1276(28.7618)	281/1105(25.4299)	414/1287(32.1678)	815/1287(63.3256)	**1131/1287(87.8788)**	1088/1287(84.5377)
≤0.50	488/1287(37.9176)	534/1287(41.4918)	537/1276(42.0846)	383/1105(34.6606)	564/1287(43.8228)	961/1287(74.6698)	**1180/1287(91.6861)**	1142/1287(88.7335)
≤0.55	629/1287(48.8733)	705/1287(54.7786)	706/1276(55.3292)	529/1105(47.8733)	719/1287(55.8664)	1076/1287(83.6053)	**1217/1287(94.5610)**	1187/1287(92.2300)

Table S2 Wasserstein distance statistics for all overall protein-ligand interactions on the held-out testset. Data includes mean, median, standard deviation, and binned cumulative frequencies for WS distances mapping the continuous stability of non-covalent interactions within the binding pocket.

	AF3	PX-v0.7.0	Boltz2	Chai-1	AFsample3	NeuralMD	AnewSampling	MD
WS distance(↓)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)
mean=	4.6885	3.4663	4.6713	10.5476	5.0469	2.0563	**1.6646**	1.6732
median=	1.9585	1.969	2.053	5.1482	2.2687	1.4588	0.9931	1.0561
std=	8.2437	3.9623	7.1907	12.8107	6.738	2.1407	2.0347	2.0104
≤0.00	0/3342(0.0000)	0/3342(0.0000)	0/3314(0.0000)	0/2906(0.0000)	0/3342(0.0000)	0/3342(0.0000)	0/3342(0.0000)	0/3342(0.0000)
≤0.05	3/3342(0.0898)	4/3342(0.1197)	5/3314(0.1509)	3/2906(0.1032)	1/3342(0.0299)	4/3342(0.1197)	31/3342(0.9276)	**189/3342(5.6553)**
≤0.10	37/3342(1.1071)	35/3342(1.0473)	31/3314(0.9354)	16/2906(0.5506)	37/3342(1.1071)	36/3342(1.0772)	124/3342(3.7104)	**308/3342(9.2160)**
≤0.15	89/3342(2.6631)	81/3342(2.4237)	77/3314(2.3235)	27/2906(0.9291)	83/3342(2.4835)	80/3342(2.3938)	241/3342(7.2113)	**403/3342(12.0586)**
≤0.20	144/3342(4.3088)	125/3342(3.7403)	131/3314(3.9529)	40/2906(1.3765)	129/3342(3.8600)	139/3342(4.1592)	385/3342(11.5200)	**495/3342(14.8115)**
≤0.25	194/3342(5.8049)	179/3342(5.3561)	180/3314(5.4315)	53/2906(1.8238)	176/3342(5.2663)	197/3342(5.8947)	505/3342(15.1107)	**584/3342(17.4746)**
≤0.30	255/3342(7.6302)	239/3342(7.1514)	231/3314(6.9704)	67/2906(2.3056)	231/3342(6.9120)	268/3342(8.0192)	630/3342(18.8510)	**670/3342(20.0479)**
≤0.35	314/3342(9.3956)	313/3342(9.3656)	283/3314(8.5395)	82/2906(2.8217)	284/3342(8.4979)	348/3342(10.4129)	727/3342(21.7534)	**771/3342(23.0700)**
≤0.40	361/3342(10.8019)	355/3342(10.6224)	346/3314(10.4406)	102/2906(3.5100)	327/3342(9.7846)	420/3342(12.5673)	827/3342(24.7457)	**853/3342(25.5236)**
≤0.45	411/3342(12.2980)	413/3342(12.3579)	399/3314(12.0398)	124/2906(4.2670)	376/3342(11.2507)	483/3342(14.4524)	**935/3342(27.9773)**	928/3342(27.7678)
≤0.50	464/3342(13.8839)	461/3342(13.7941)	445/3314(13.4279)	144/2906(4.9553)	421/3342(12.5972)	540/3342(16.1580)	**1011/3342(30.2513)**	1002/3342(29.9820)
≤0.55	514/3342(15.3800)	514/3342(15.3800)	496/3314(14.9668)	160/2906(5.5058)	488/3342(14.6020)	610/3342(18.2525)	**1098/3342(32.8546)**	1073/3342(32.1065)
≤0.60	555/3342(16.6068)	559/3342(16.7265)	539/3314(16.2643)	182/2906(6.2629)	520/3342(15.5595)	678/3342(20.2873)	**1190/3342(35.6074)**	1141/3342(34.1412)
≤0.65	593/3342(17.7439)	611/3342(18.2825)	578/3314(17.4412)	204/2906(7.0200)	555/3342(16.6068)	749/3342(22.4117)	**1266/3342(37.8815)**	1201/3342(35.9366)
≤0.70	635/3342(19.0006)	656/3342(19.6290)	619/3314(18.6783)	227/2906(7.8114)	599/3342(17.9234)	823/3342(24.6260)	**1332/3342(39.8564)**	1282/3342(38.3603)
≤0.75	681/3342(20.3770)	698/3342(20.8857)	664/3314(20.0362)	249/2906(8.5685)	635/3342(19.0006)	891/3342(26.6607)	**1392/3342(41.6517)**	1344/3342(40.2154)
≤0.80	729/3342(21.8133)	749/3342(22.4117)	699/3314(21.0923)	270/2906(9.2911)	681/3342(20.3770)	965/3342(28.8749)	**1456/3342(43.5667)**	1395/3342(41.7415)
≤0.85	775/3342(23.1897)	785/3342(23.4889)	740/3314(22.3295)	287/2906(9.8761)	728/3342(21.7834)	1028/3342(30.7600)	**1521/3342(45.5117)**	1459/3342(43.6565)
≤0.90	824/3342(24.6559)	834/3342(24.9551)	784/3314(23.6572)	314/2906(10.8052)	767/3342(22.9503)	1083/3342(32.4057)	**1578/3342(47.2172)**	1523/3342(45.5715)
≤0.95	871/3342(26.0622)	894/3342(26.7504)	813/3314(24.5323)	328/2906(11.2870)	816/3342(24.4165)	1148/3342(34.3507)	**1623/3342(48.5637)**	1577/3342(47.1873)
≤1.00	921/3342(27.5583)	931/3342(27.8576)	855/3314(25.7996)	343/2906(11.8032)	850/3342(25.4339)	1206/3342(36.0862)	**1681/3342(50.2992)**	1625/3342(48.6236)
≤1.05	963/3342(28.8151)	988/3342(29.5631)	900/3314(27.1575)	368/2906(12.6635)	903/3342(27.0197)	1258/3342(37.6421)	**1735/3342(51.9150)**	1663/3342(49.7606)

Table S3 Wasserstein distance statistics for highly conserved protein-ligand interactions on the held-out testset. This table isolates structural interactions that occur with high frequency (≥ 0.8) in the reference REMD trajectories, highlighting the capacity of each model to capture strongly bound, dominant interaction states.

WS distance(↓)	AF3 Fraction(%)	PX-v0.7.0 Fraction(%)	Boltz2 Fraction(%)	Chai-1 Fraction(%)	AFsample3 Fraction(%)	NeuralMD Fraction(%)	AnewSampling Fraction(%)	MD Fraction(%)
≤0.00	0/297(0.0000)	0/297(0.0000)	0/291(0.0000)	0/242(0.0000)	0/297(0.0000)	0/297(0.0000)	0/297(0.0000)	0/297(0.0000)
≤0.05	2/297(0.6734)	2/297(0.6734)	5/291(1.7182)	1/242(0.4132)	1/297(0.3367)	2/297(0.6734)	14/297(4.7138)	**114/297(38.3838)**
≤0.10	27/297(9.0909)	24/297(8.0808)	19/291(6.5292)	9/242(3.7190)	24/297(8.0808)	21/297(7.0707)	54/297(18.1818)	**158/297(53.1987)**
≤0.15	67/297(22.5589)	57/297(19.1919)	50/291(17.1821)	17/242(7.0248)	56/297(18.8552)	32/297(10.7744)	104/297(35.0168)	**187/297(62.9630)**
≤0.20	95/297(31.9865)	81/297(27.2727)	81/291(27.8351)	22/242(9.0909)	77/297(25.9259)	59/297(19.8653)	151/297(50.8418)	**206/297(69.3603)**
≤0.25	121/297(40.7407)	108/297(36.3636)	104/291(35.7388)	28/242(11.5702)	102/297(34.3434)	82/297(27.6094)	176/297(59.2593)	**215/297(72.3906)**
≤0.30	145/297(48.8215)	135/297(45.4545)	120/291(41.2371)	31/242(12.8099)	123/297(41.4141)	106/297(35.6902)	197/297(66.3300)	**224/297(75.4209)**
≤0.35	164/297(55.2189)	168/297(56.5657)	138/291(47.4227)	34/242(14.0496)	138/297(46.4646)	131/297(44.1077)	210/297(70.7071)	**235/297(79.1246)**
≤0.40	177/297(59.5960)	177/297(59.5960)	159/291(54.6392)	40/242(16.5289)	150/297(50.5051)	155/297(52.1886)	222/297(74.7475)	**238/297(80.1347)**
≤0.45	183/297(61.6162)	184/297(61.9529)	168/291(57.7320)	44/242(18.1818)	155/297(52.1886)	168/297(56.5657)	232/297(78.1145)	**242/297(81.4815)**
≤0.50	194/297(65.3199)	195/297(65.6566)	178/291(61.1684)	47/242(19.4215)	164/297(55.2189)	177/297(59.5960)	242/297(81.4815)	**247/297(83.1650)**
≤0.55	199/297(67.0034)	207/297(69.6970)	189/291(64.9485)	49/242(20.2479)	172/297(57.9125)	188/297(63.2997)	248/297(83.5017)	**250/297(84.1751)**
≤0.60	202/297(68.0135)	209/297(70.3704)	196/291(67.3540)	54/242(22.3140)	177/297(59.5960)	201/297(67.6768)	254/297(85.5219)	**256/297(86.1953)**
≤0.65	211/297(71.0438)	215/297(72.3906)	199/291(68.3849)	56/242(23.1405)	181/297(60.9428)	206/297(69.3603)	**258/297(86.8687)**	257/297(86.5320)
≤0.70	213/297(71.7172)	218/297(73.4007)	201/291(69.0722)	57/242(23.5537)	186/297(62.6263)	217/297(73.0640)	258/297(86.8687)	**260/297(87.5421)**
≤0.75	216/297(72.7273)	221/297(74.4108)	202/291(69.4158)	61/242(25.2066)	189/297(63.6364)	222/297(74.7475)	260/297(87.5421)	**264/297(88.8889)**
≤0.80	220/297(74.0741)	225/297(75.7576)	206/291(70.7904)	61/242(25.2066)	192/297(64.6465)	234/297(78.7879)	260/297(87.5421)	**265/297(89.2256)**
≤0.85	224/297(75.4209)	227/297(76.4310)	209/291(71.8213)	61/242(25.2066)	194/297(65.3199)	237/297(79.7980)	262/297(88.2155)	**267/297(89.8990)**
≤0.90	226/297(76.0943)	233/297(78.4512)	214/291(73.5395)	62/242(25.6198)	195/297(65.6566)	242/297(81.4815)	265/297(89.2256)	**271/297(91.2458)**
≤0.95	228/297(76.7677)	237/297(79.7980)	215/291(73.8832)	62/242(25.6198)	202/297(68.0135)	245/297(82.4916)	265/297(89.2256)	**272/297(91.5825)**
≤1.00	229/297(77.1044)	238/297(80.1347)	216/291(74.2268)	63/242(26.0331)	203/297(68.3502)	255/297(85.8586)	267/297(89.8990)	**274/297(92.2559)**
≤1.05	229/297(77.1044)	242/297(81.4815)	217/291(74.5704)	64/242(26.4463)	204/297(68.6869)	255/297(85.8586)	269/297(90.5724)	**275/297(92.5926)**

Table S4 Evaluation metrics for protein backbone structural dynamics on the held-out testset. Comparative performance of each model in predicting global structural flexibility, quantified via Spearman correlation r_s and RMSE of the protein backbone Cα RMSF relative to the REMD ground truth.

	AF3	PX-v0.7.0	Boltz2	Chai-1	AFsample3	NeuralMD	AnewSampling	MD
r_s (↑)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)
mean=	0.7854	0.7788	0.791	0.801	0.7337	nan	0.8847	**0.8862**
median=	0.8106	0.8161	0.8188	0.8377	0.771	nan	**0.913**	0.9058
std=	0.1184	0.1334	0.1294	0.1311	0.1448	nan	0.1323	0.0754
≥1.00	0/88(0.0000)	0/88(0.0000)	0/87(0.0000)	0/79(0.0000)	0/88(0.0000)	0/88(0.0000)	0/88(0.0000)	0/88(0.0000)
≥0.95	0/88(0.0000)	0/88(0.0000)	1/87(1.1494)	0/79(0.0000)	0/88(0.0000)	0/88(0.0000)	**22/88(25.0000)**	11/88(12.5000)
≥0.90	8/88(9.0909)	8/88(9.0909)	16/87(18.3908)	16/79(20.2532)	3/88(3.4091)	0/88(0.0000)	**58/88(65.9091)**	48/88(54.5455)
≥**0.85**	30/88(34.0909)	30/88(34.0909)	31/87(35.6322)	34/79(43.0380)	16/88(18.1818)	0/88(0.0000)	**75/88(85.2273)**	68/88(77.2727)
≥0.80	48/88(54.5455)	48/88(54.5455)	51/87(58.6207)	53/79(67.0886)	32/88(36.3636)	0/88(0.0000)	**80/88(90.9091)**	79/88(89.7727)
≥0.75	66/88(75.0000)	64/88(72.7273)	66/87(75.8621)	59/79(74.6835)	53/88(60.2273)	0/88(0.0000)	82/88(93.1818)	**85/88(96.5909)**
≥0.70	72/88(81.8182)	71/88(80.6818)	73/87(83.9080)	65/79(82.2785)	60/88(68.1818)	0/88(0.0000)	84/88(95.4545)	**86/88(97.7273)**
≥0.65	76/88(86.3636)	76/88(86.3636)	76/87(87.3563)	68/79(86.0759)	68/88(77.2727)	0/88(0.0000)	84/88(95.4545)	**86/88(97.7273)**
≥0.60	79/88(89.7727)	79/88(89.7727)	77/87(88.5057)	73/79(92.4051)	75/88(85.2273)	0/88(0.0000)	84/88(95.4545)	**87/88(98.8636)**
rmse(↓)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)	Value / Fraction(%)
mean=	1.8854	1.9283	1.8812	1.8842	2.2459	nan	0.5523	**0.4621**
median=	0.7921	0.7842	0.7864	0.768	0.7905	nan	**0.2286**	0.2699
std=	3.2245	2.7312	2.9326	2.6836	3.6016	nan	0.9039	0.5881
≤0.00	0/88(0.0000)	0/88(0.0000)	0/87(0.0000)	0/79(0.0000)	0/88(0.0000)	0/88(0.0000)	0/88(0.0000)	0/88(0.0000)
≤0.50	1/88(1.1364)	0/88(0.0000)	2/87(2.2989)	2/79(2.5316)	4/88(4.5455)	0/88(0.0000)	66/88(75.0000)	**73/88(82.9545)**
≤1.00	65/88(73.8636)	62/88(70.4545)	62/87(71.2644)	53/79(67.0886)	59/88(67.0455)	0/88(0.0000)	74/88(84.0909)	**80/88(90.9091)**
≤1.50	71/88(80.6818)	69/88(78.4091)	69/87(79.3103)	59/79(74.6835)	69/88(78.4091)	0/88(0.0000)	80/88(90.9091)	**82/88(93.1818)**
≤2.00	75/88(85.2273)	71/88(80.6818)	74/87(85.0575)	62/79(78.4810)	72/88(81.8182)	0/88(0.0000)	83/88(94.3182)	**84/88(95.4545)**
≤2.50	78/88(88.6364)	71/88(80.6818)	74/87(85.0575)	65/79(82.2785)	72/88(81.8182)	0/88(0.0000)	**86/88(97.7273)**	**86/88(97.7273)**
≤3.00	79/88(89.7727)	73/88(82.9545)	75/87(86.2069)	67/79(84.8101)	72/88(81.8182)	0/88(0.0000)	**87/88(98.8636)**	**87/88(98.8636)**
≤3.50	79/88(89.7727)	75/88(85.2273)	75/87(86.2069)	68/79(86.0759)	75/88(85.2273)	0/88(0.0000)	**87/88(98.8636)**	**87/88(98.8636)**
≤4.00	80/88(90.9091)	77/88(87.5000)	78/87(89.6552)	69/79(87.3418)	75/88(85.2273)	0/88(0.0000)	87/88(98.8636)	**88/88(100.0000)**
≤4.50	81/88(92.0455)	79/88(89.7727)	79/87(90.8046)	71/79(89.8734)	76/88(86.3636)	0/88(0.0000)	87/88(98.8636)	**88/88(100.0000)**
≤5.00	81/88(92.0455)	79/88(89.7727)	79/87(90.8046)	71/79(89.8734)	76/88(86.3636)	0/88(0.0000)	87/88(98.8636)	**88/88(100.0000)**

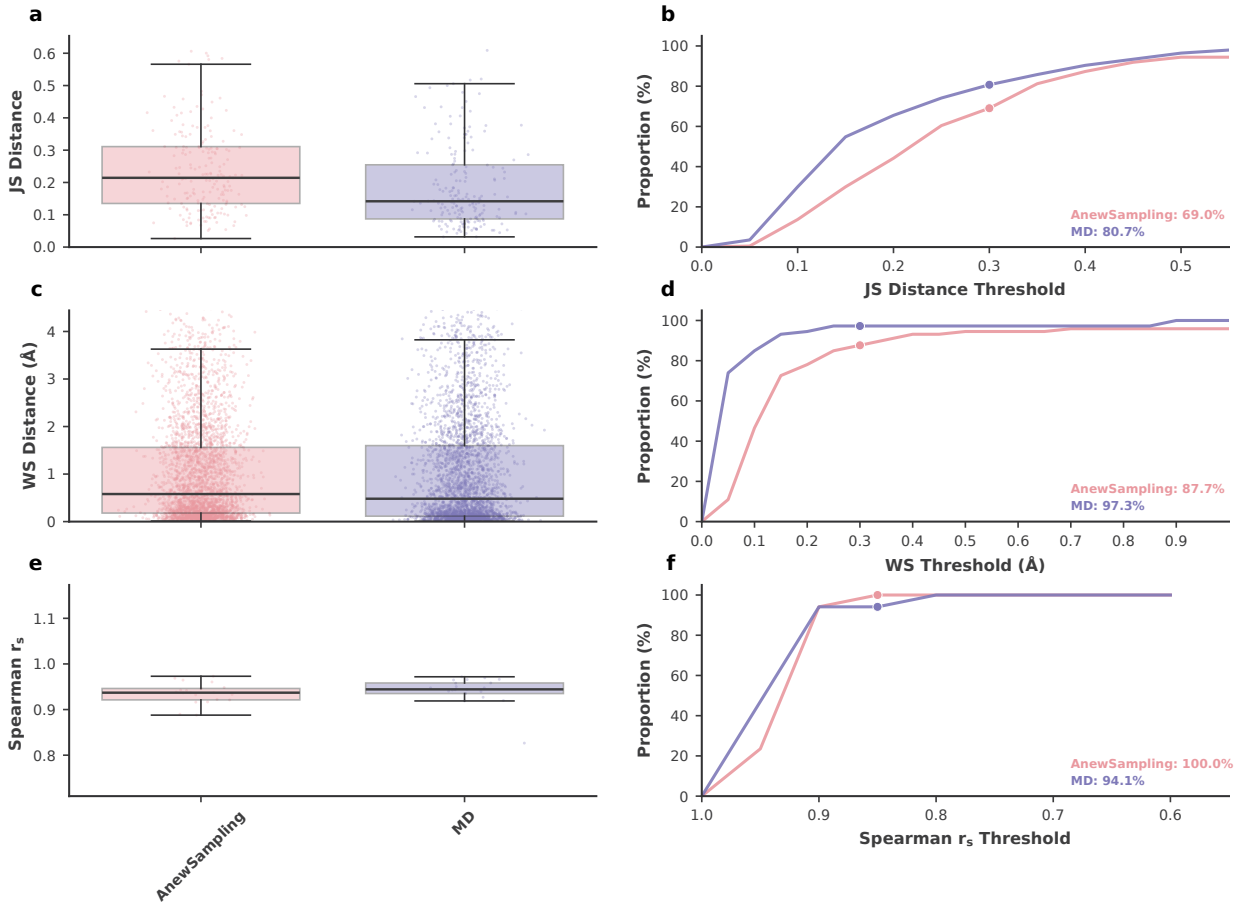

Figure S1 Evaluation of generative models for protein-ligand conformational ensemble on an independent internal pipeline dataset. (a) Boxplot comparing the JS distance of ligand torsion angle distributions between AnewSampling and the MD baseline. **(b)** Cumulative success rate of torsion JS distances; AnewSampling achieves a 68.8% success rate at a threshold of ≤ 0.3, compared to 80.2% for the MD baseline.**(c)** Boxplot detailing the WS distance (in Å) for protein-ligand interaction distributions. **(d)** Cumulative success rate for highly conserved, stable interactions (MD frequency ≥ 0.8), with AnewSampling reaching an 87.7% success rate at a stringent WS threshold of ≤ 0.3 Å. **(e)** Boxplot of RMSE for protein backbone Cα RMSF, demonstrating the consistent prediction of global protein flexibility. **(f)** Cumulative success rate of r_s for protein RMSF; AnewSampling exhibits a 100.0% success rate at $r_s \geq 0.85$, demonstrating robust recovery of coherent structural dynamics across the internal dataset.

Table S5 Statistical summary of JS distances for ligand torsion angle distributions on the jacsmerck dataset. The table provides the mean, median, standard deviation, and cumulative success rates (binned from ≤ 0.00 to ≤ 0.55) of JS distances for conformations generated by the co-folding models, MD-trained generative models, and the MD baseline.

	AF3	PX-v0.7.0	Boltz2	Chai-1	AFsample3	NeuralMD	AnewSampling	MD
JS distance(↓)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)
mean=	0.5844	0.5910	0.5869	0.5589	0.5653	0.3854	0.2659	0.2640
median=	0.5918	0.5937	0.5934	0.5676	0.5753	0.3720	0.2395	0.2251
std=	0.1165	0.1199	0.1217	0.1321	0.1208	0.1329	0.1421	0.1562
≤0.00	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)
≤0.05	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	12/3060(0.3922)	23/3060(0.7516)
≤0.10	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	0/3060(0.0000)	4/3060(0.1307)	251/3060(8.2026)	335/3060(10.9477)
≤0.15	0/3060(0.0000)	0/3060(0.0000)	1/3060(0.0327)	2/3060(0.0654)	0/3060(0.0000)	48/3060(1.5686)	697/3060(22.7778)	842/3060(27.5163)
≤0.20	2/3060(0.0654)	5/3060(0.1634)	8/3060(0.2614)	18/3060(0.5882)	3/3060(0.0980)	191/3060(6.2418)	1195/3060(39.0523)	1328/3060(43.3987)
≤0.25	10/3060(0.3268)	12/3060(0.3922)	17/3060(0.5556)	53/3060(1.7320)	14/3060(0.4575)	465/3060(15.1961)	1614/3060(52.7451)	1706/3060(55.7516)
≤0.30	36/3060(1.1765)	39/3060(1.2745)	42/3060(1.3725)	114/3060(3.7255)	57/3060(1.8627)	872/3060(28.4967)	2027/3060(66.2418)	2037/3060(66.5686)
≤0.35	101/3060(3.3007)	89/3060(2.9085)	105/3060(3.4314)	225/3060(7.3529)	145/3060(4.7386)	1302/3060(42.5490)	2314/3060(75.6209)	2269/3060(74.1503)
≤0.40	204/3060(6.6667)	188/3060(6.1438)	218/3060(7.1242)	373/3060(12.1895)	314/3060(10.2614)	1788/3060(58.4314)	2540/3060(83.0065)	2474/3060(80.8497)
≤0.45	434/3060(14.1830)	366/3060(11.9608)	417/3060(13.6275)	622/3060(20.3268)	543/3060(17.7451)	2173/3060(71.0131)	2697/3060(88.1373)	2644/3060(86.4052)
≤0.50	685/3060(22.3856)	672/3060(21.9608)	701/3060(22.9085)	951/3060(31.0784)	882/3060(28.8235)	2492/3060(81.4379)	2820/3060(92.1569)	2768/3060(90.4575)
≤0.55	1088/3060(35.5556)	1082/3060(35.3595)	1104/3060(36.0784)	1393/3060(45.5229)	1288/3060(42.0915)	2710/3060(88.5621)	2925/3060(95.5882)	2874/3060(93.9216)

Table S6 Wasserstein distance statistics for all overall protein-ligand interactions on the jacsmerck dataset. Data includes mean, median, standard deviation, and binned cumulative frequencies for WS distances mapping the continuous stability of non-covalent interactions within the binding pocket.

	AF3	PX-v0.7.0	Boltz2	Chai-1	AFsample3	NeuralMD	AnewSampling	MD
WS distance(↓)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)
mean=	1.0557	1.1332	1.1827	4.6657	1.2792	1.1607	0.7528	0.8221
median=	0.6742	0.8422	0.8559	2.4410	0.9267	0.8001	0.4506	0.4750
std=	1.1443	1.0661	1.1192	7.4278	1.1838	1.1276	0.8928	0.9971
≤0.00	0/6276(0.0000)	0/7736(0.0000)	0/7736(0.0000)	0/7736(0.0000)	0/7736(0.0000)	0/7736(0.0000)	0/7736(0.0000)	0/7736(0.0000)
≤0.05	26/6276(0.4143)	1/7736(0.0129)	2/7736(0.0259)	2/7736(0.0259)	33/7736(0.4266)	25/7736(0.3232)	342/7736(4.4209)	776/7736(10.0310)
≤0.10	228/6276(3.6329)	127/7736(1.6417)	107/7736(1.3831)	18/7736(0.2327)	201/7736(2.5982)	191/7736(2.4690)	1069/7736(13.8185)	1405/7736(18.1618)
≤0.15	470/6276(7.4888)	316/7736(4.0848)	265/7736(3.4255)	72/7736(0.9307)	355/7736(4.5889)	453/7736(5.8557)	1673/7736(21.6262)	1865/7736(24.1081)
≤0.20	767/6276(12.2212)	684/7736(8.8418)	584/7736(7.5491)	191/7736(2.4690)	651/7736(8.4152)	781/7736(10.0957)	2106/7736(27.2234)	2252/7736(29.1107)
≤0.25	1143/6276(18.2122)	1019/7736(13.1722)	985/7736(12.7327)	332/7736(4.2916)	971/7736(12.5517)	1123/7736(14.5165)	2546/7736(32.9111)	2630/7736(33.9969)
≤0.30	1505/6276(23.9802)	1389/7736(17.9550)	1344/7736(17.3733)	427/7736(5.5196)	1291/7736(16.6882)	1473/7736(19.0408)	2914/7736(37.6680)	2950/7736(38.1334)
≤0.35	1781/6276(28.3779)	1664/7736(21.5098)	1638/7736(21.1737)	526/7736(6.7994)	1574/7736(20.3464)	1802/7736(23.2937)	3268/7736(42.2441)	3260/7736(42.1406)
≤0.40	2026/6276(32.2817)	1960/7736(25.3361)	1888/7736(24.4054)	610/7736(7.8852)	1838/7736(23.7590)	2062/7736(26.6546)	3601/7736(46.5486)	3527/7736(45.5920)
≤0.45	2270/6276(36.1695)	2230/7736(28.8263)	2152/7736(27.8180)	710/7736(9.1779)	2052/7736(26.5253)	2315/7736(29.9250)	3865/7736(49.9612)	3749/7736(48.4617)
≤0.50	2495/6276(39.7546)	2470/7736(31.9286)	2365/7736(30.5714)	789/7736(10.1991)	2275/7736(29.4080)	2583/7736(33.3893)	4123/7736(53.2963)	3983/7736(51.4866)
≤0.55	2690/6276(42.8617)	2688/7736(34.7466)	2584/7736(33.4023)	876/7736(11.3237)	2482/7736(32.0838)	2831/7736(36.5951)	4315/7736(55.7782)	4183/7736(54.0719)
≤0.60	2855/6276(45.4908)	2933/7736(37.9137)	2843/7736(36.7503)	964/7736(12.4612)	2698/7736(34.8759)	3074/7736(39.7363)	4554/7736(58.8676)	4370/7736(56.4891)
≤0.65	3057/6276(48.7094)	3121/7736(40.3438)	3066/7736(39.6329)	1064/7736(13.7539)	2883/7736(37.2673)	3286/7736(42.4767)	4731/7736(61.1556)	4540/7736(58.6867)
≤0.70	3233/6276(51.5137)	3322/7736(42.9421)	3275/7736(42.3345)	1158/7736(14.9690)	3093/7736(39.9819)	3501/7736(45.2559)	4932/7736(63.7539)	4702/7736(60.7808)
≤0.75	3386/6276(53.9516)	3514/7736(45.4240)	3484/7736(45.0362)	1256/7736(16.2358)	3268/7736(42.2441)	3696/7736(47.7766)	5085/7736(65.7316)	4886/7736(63.1593)
≤0.80	3549/6276(56.5488)	3712/7736(47.9835)	3681/7736(47.5827)	1338/7736(17.2958)	3453/7736(44.6355)	3866/7736(49.9741)	5245/7736(67.7999)	5058/7736(65.3826)
≤0.85	3683/6276(58.6839)	3898/7736(50.3878)	3847/7736(49.7285)	1415/7736(18.2911)	3618/7736(46.7684)	4048/7736(52.3268)	5379/7736(69.5321)	5205/7736(67.2828)
≤0.90	3815/6276(60.7871)	4065/7736(52.5465)	4011/7736(51.8485)	1509/7736(19.5062)	3778/7736(48.8366)	4223/7736(54.5889)	5504/7736(71.1479)	5347/7736(69.1184)
≤0.95	3956/6276(63.0338)	4204/7736(54.3433)	4173/7736(53.9426)	1582/7736(20.4498)	3947/7736(51.0212)	4376/7736(56.5667)	5621/7736(72.6603)	5469/7736(70.6954)
≤1.00	4068/6276(64.8184)	4384/7736(56.6701)	4308/7736(55.6877)	1664/7736(21.5098)	4090/7736(52.8697)	4522/7736(58.4540)	5732/7736(74.0951)	5567/7736(71.9623)
≤1.05	4178/6276(66.5711)	4562/7736(58.9710)	4452/7736(57.5491)	1758/7736(22.7249)	4252/7736(54.9638)	4651/7736(60.1215)	5845/7736(75.5558)	5663/7736(73.2032)

Table S7 Wasserstein distance statistics for highly conserved protein-ligand interactions on the jacsmerck dataset. This table isolates structural interactions that occur with high frequency (≥ 0.8) in the reference REMD trajectories, highlighting the capacity of each model to capture strongly bound, dominant interaction states.

	AF3	PX-v0.7.0	Boltz2	Chai-1	AFsample3	NeuralMD	AnewSampling	MD
WS distance(↓)	Fraction(%)	Fraction(%)	Fraction(%)	Fraction(%)	Fraction(%)	Fraction(%)	Fraction(%)	Fraction(%)
≤0.00	0/1174(0.0000)	0/1058(0.0000)	0/1058(0.0000)	0/1058(0.0000)	0/1058(0.0000)	0/1058(0.0000)	0/1058(0.0000)	0/1058(0.0000)
≤0.05	26/1174(2.2147)	1/1058(0.0945)	2/1058(0.1890)	1/1058(0.0945)	33/1058(3.1191)	21/1058(1.9849)	168/1058(15.8790)	492/1058(46.5028)
≤0.10	197/1174(16.7802)	113/1058(10.6805)	95/1058(8.9792)	7/1058(0.6616)	175/1058(16.5406)	121/1058(11.4367)	533/1058(50.3781)	678/1058(64.0832)
≤0.15	375/1174(31.9421)	261/1058(24.6692)	229/1058(21.6446)	39/1058(3.6862)	255/1058(24.1021)	251/1058(23.7240)	722/1058(68.2420)	743/1058(70.2268)
≤0.20	559/1174(47.6150)	487/1058(46.0302)	455/1058(43.0057)	93/1058(8.7902)	388/1058(36.6730)	358/1058(33.8374)	792/1058(74.8582)	808/1058(76.3705)
≤0.25	714/1174(60.8177)	607/1058(57.3724)	589/1058(55.6711)	150/1058(14.1777)	498/1058(47.0699)	457/1058(43.1947)	845/1058(79.8677)	845/1058(79.8677)
≤0.30	847/1174(72.1465)	687/1058(64.9338)	688/1058(65.0284)	184/1058(17.3913)	609/1058(57.5614)	551/1058(52.0794)	881/1058(83.2703)	877/1058(82.8922)
≤0.35	905/1174(77.0869)	744/1058(70.3214)	743/1058(70.2268)	210/1058(19.8488)	685/1058(64.7448)	642/1058(60.6805)	907/1058(85.7278)	903/1058(85.3497)
≤0.40	949/1174(80.8348)	797/1058(75.3308)	787/1058(74.3856)	227/1058(21.4556)	748/1058(70.6994)	685/1058(64.7448)	936/1058(88.4688)	929/1058(87.8072)
≤0.45	990/1174(84.3271)	827/1058(78.1664)	826/1058(78.0718)	245/1058(23.1569)	778/1058(73.5350)	740/1058(69.9433)	944/1058(89.2250)	941/1058(88.9414)
≤0.50	1022/1174(87.0528)	865/1058(81.7580)	843/1058(79.6786)	256/1058(24.1966)	806/1058(76.1815)	797/1058(75.3308)	965/1058(91.2098)	954/1058(90.1701)
≤0.55	1039/1174(88.5009)	888/1058(83.9319)	863/1058(81.5690)	269/1058(25.4253)	835/1058(78.9225)	834/1058(78.8280)	973/1058(91.9660)	961/1058(90.8318)
≤0.60	1061/1174(90.3748)	920/1058(86.9565)	892/1058(84.3100)	284/1058(26.8431)	858/1058(81.0964)	866/1058(81.8526)	987/1058(93.2892)	972/1058(91.8715)
≤0.65	1084/1174(92.3339)	933/1058(88.1853)	926/1058(87.5236)	301/1058(28.4499)	880/1058(83.1758)	893/1058(84.4045)	997/1058(94.2344)	980/1058(92.6276)
≤0.70	1103/1174(93.9523)	950/1058(89.7921)	946/1058(89.4140)	305/1058(28.8280)	897/1058(84.7826)	917/1058(86.6730)	1007/1058(95.1796)	985/1058(93.1002)
≤0.75	1110/1174(94.5486)	966/1058(91.3043)	961/1058(90.8318)	317/1058(29.9622)	909/1058(85.9168)	941/1058(88.9414)	1013/1058(95.7467)	994/1058(93.9509)
≤0.80	1118/1174(95.2300)	976/1058(92.2495)	977/1058(92.3440)	328/1058(31.0019)	914/1058(86.3894)	960/1058(90.7372)	1018/1058(96.2193)	1004/1058(94.8960)
≤0.85	1121/1174(95.4855)	989/1058(93.4783)	986/1058(93.1947)	332/1058(31.3800)	921/1058(87.0510)	975/1058(92.1550)	1022/1058(96.5974)	1011/1058(95.5577)
≤0.90	1124/1174(95.7411)	1001/1058(94.6125)	996/1058(94.1399)	342/1058(32.3251)	933/1058(88.1853)	988/1058(93.3837)	1024/1058(96.7864)	1014/1058(95.8412)
≤0.95	1130/1174(96.2521)	1006/1058(95.0851)	1002/1058(94.7070)	351/1058(33.1758)	943/1058(89.1304)	999/1058(94.4234)	1027/1058(97.0699)	1021/1058(96.5028)
≤1.00	1137/1174(96.8484)	1017/1058(96.1248)	1007/1058(95.1796)	362/1058(34.2155)	947/1058(89.5085)	1002/1058(94.7070)	1029/1058(97.2590)	1023/1058(96.6919)
≤1.05	1140/1174(97.1039)	1023/1058(96.6919)	1012/1058(95.6522)	370/1058(34.9716)	954/1058(90.1701)	1008/1058(95.2741)	1030/1058(97.3535)	1026/1058(96.9754)

Table S8 Evaluation metrics for protein backbone structural dynamics on the jacsmerck dataset. Comparative performance of each model in predicting global structural flexibility, quantified via r_s and RMSE of the protein backbone Cα RMSF relative to the REMD ground truth.

	AF3	PX-v0.7.0	Boltz2	Chai-1	AFsample3	NeuralMD	AnewSampling	MD
r_s(↑)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)
mean=	0.8422	0.8201	0.7795	0.7995	0.6208	0.0288	0.9490	0.9237
median=	0.8649	0.8364	0.8655	0.8217	0.6722	0.0203	0.9529	0.9312
std=	0.0964	0.0866	0.2229	0.1198	0.2333	0.1273	0.0258	0.0365
≥1.00	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)
≥0.95	2/593(0.3373)	0/593(0.0000)	5/593(0.8432)	32/593(5.3963)	2/593(0.3373)	0/593(0.0000)	328/593(55.3120)	129/593(21.7538)
≥0.90	201/593(33.8954)	112/593(18.8870)	201/593(33.8954)	121/593(20.4047)	58/593(9.7808)	0/593(0.0000)	569/593(95.9528)	485/593(81.7875)
≥0.85	329/593(55.4806)	248/593(41.8212)	338/593(56.9983)	234/593(39.4604)	98/593(16.5261)	0/593(0.0000)	590/593(99.4941)	566/593(95.4469)
≥0.80	441/593(74.3676)	391/593(65.9359)	423/593(71.3322)	330/593(55.6492)	182/593(30.6914)	0/593(0.0000)	592/593(99.8314)	586/593(98.8196)
≥0.75	495/593(83.4739)	499/593(84.1484)	468/593(78.9207)	440/593(74.1990)	221/593(37.2681)	0/593(0.0000)	592/593(99.8314)	590/593(99.4941)
≥0.70	563/593(94.9410)	541/593(91.2310)	502/593(84.6543)	507/593(85.4975)	266/593(44.8567)	0/593(0.0000)	592/593(99.8314)	592/593(99.8314)
≥0.65	584/593(98.4823)	558/593(94.0978)	506/593(85.3288)	536/593(90.3879)	317/593(53.4570)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≥0.60	585/593(98.6509)	572/593(96.4587)	506/593(85.3288)	551/593(92.9174)	351/593(59.1906)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
rmse(↓)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)	Value/Fraction(%)
mean=	0.8600	0.8210	0.9440	2.3285	1.8907	NaN	0.2359	0.3042
median=	0.9032	0.8112	0.8584	0.8654	0.9212	NaN	0.2273	0.2996
std=	0.1441	0.1261	0.4008	3.2047	1.9783	NaN	0.0906	0.1100
≤0.00	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)	0/593(0.0000)
≤0.50	1/593(0.1686)	0/593(0.0000)	1/593(0.1686)	1/593(0.1686)	3/593(0.5059)	0/593(0.0000)	584/593(98.4823)	569/593(95.9528)
≤1.00	514/593(86.6779)	555/593(93.5919)	491/593(82.7993)	443/593(74.7049)	372/593(62.7319)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤1.50	593/593(100.0000)	593/593(100.0000)	525/593(88.5329)	465/593(78.4148)	439/593(74.0304)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤2.00	593/593(100.0000)	593/593(100.0000)	564/593(95.1096)	480/593(80.9444)	454/593(76.5599)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤2.50	593/593(100.0000)	593/593(100.0000)	588/593(99.1568)	480/593(80.9444)	461/593(77.7403)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤3.00	593/593(100.0000)	593/593(100.0000)	593/593(100.0000)	480/593(80.9444)	462/593(77.9089)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤3.50	593/593(100.0000)	593/593(100.0000)	593/593(100.0000)	480/593(80.9444)	470/593(79.2580)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤4.00	593/593(100.0000)	593/593(100.0000)	593/593(100.0000)	480/593(80.9444)	484/593(81.6189)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤4.50	593/593(100.0000)	593/593(100.0000)	593/593(100.0000)	480/593(80.9444)	492/593(82.9680)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)
≤5.00	593/593(100.0000)	593/593(100.0000)	593/593(100.0000)	480/593(80.9444)	511/593(86.1720)	0/593(0.0000)	593/593(100.0000)	593/593(100.0000)

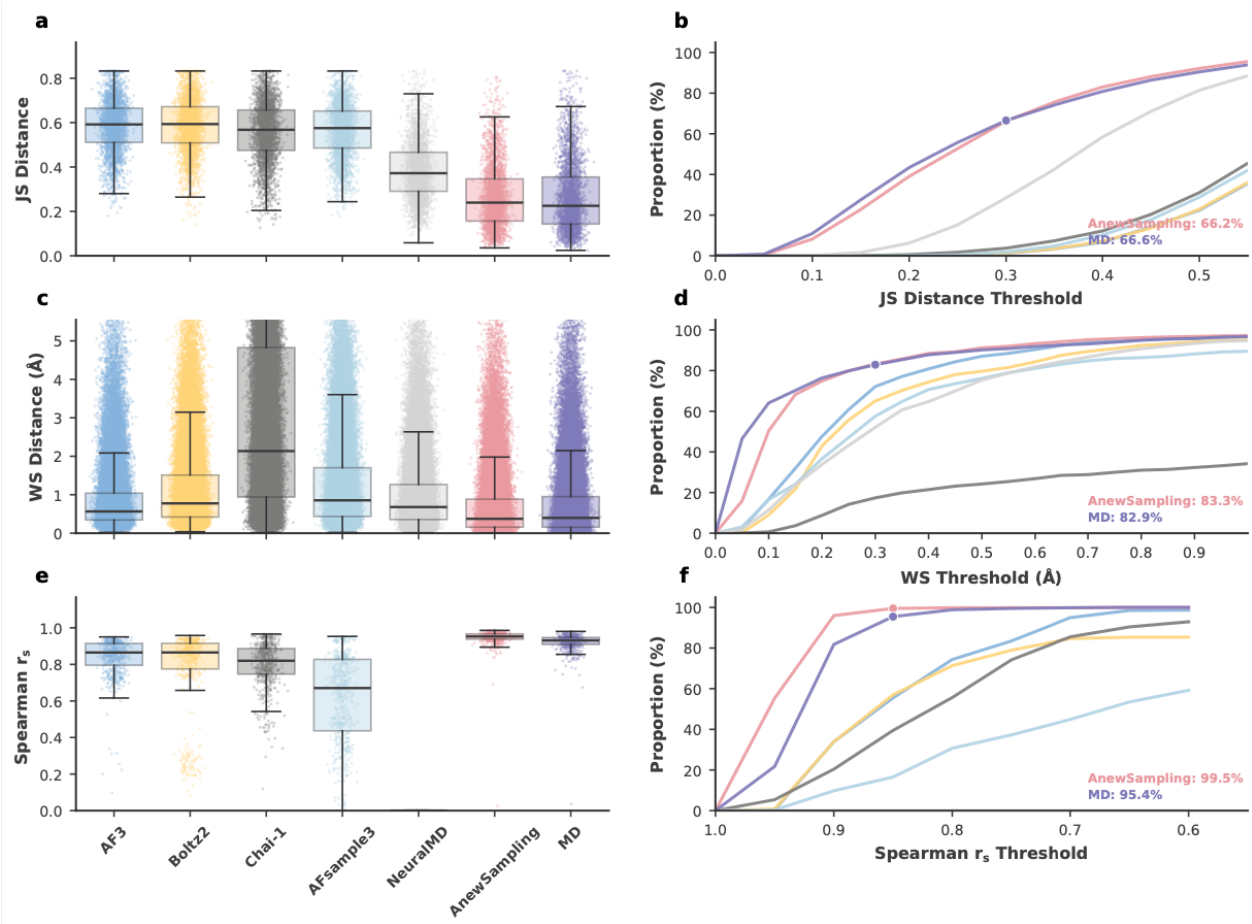

Figure S2 Comparative performance of generative models in capturing protein-ligand conformational dynamics for the jacsmerck dataset. (a) Boxplot detailing the JS distance of ligand torsion angle distributions across generated ensembles compared to the MD reference. Evaluated models include AF3, Boltz2, Chai-1, AFsample3, NeuralMD, and AnewSampling. **(b)** Cumulative proportion curves for ligand torsion JS distances. AnewSampling successfully captures ligand flexibility, achieving a 66.2% success rate that closely mirrors the MD baseline (66.6%) at the focal threshold. **(c)** Boxplot of WS distances (in Å) assessing the distributional accuracy and stability of protein-ligand interactions. **(d)** Cumulative success rate curves for interaction WS distances. AnewSampling demonstrates robust recovery of physical interaction networks, matching the rigorous MD reference (83.3% vs. 82.9%) at a stringent ≤ 0.3 Å threshold. **(e)** Boxplot evaluating the consistency of global protein backbone structural dynamics across the tested models. **(f)** Cumulative proportion curves for the r_s of protein backbone RMSF. AnewSampling exhibits superior preservation of coherent global motions, achieving a near-perfect 99.5% success rate compared to 95.4% for the MD baseline.