DynaPPI: A large-scale dynamic protein dataset for AI-driven advances in protein interactomics

Anonymous Author(s)
Affiliation
Address
email

Abstract

Diffusion models have been widely explored in protein backbone generation due to their powerful generation capabilities. However, in today's AI-driven biological research, predicting the structure of unknown multi-chain protein aggregates (called "*complexes*" in biology) remains an unsolved challenge. This is because existing static or dynamic protein datasets focus solely on static snapshots or single-entity trajectories, neglecting the dynamic process of multiple monomers forming complexes. To alleviate this dilemma, we present **DynaPPI**, a dynamic protein dataset comprising molecular dynamics (MD) trajectories of protein complex formation from dissociated chains to the bound state, as a pivotal resource to bridge the gap between static structural biology and the inherently temporal nature of dynamic molecular interactions. Benefiting from this dataset, diffusion models can explicitly learn the dynamic binding trajectories of known complexes and accurately predict the structures of unknown complexes based on their diverse generative properties, thereby further catalyzing AI-driven structural biology and protein interactomics.

1 AI Task Definition

The primary AI task associated with our dataset is defined as a ***conditional generative prediction task***, wherein diffusion models are trained to predict and generate realistic dynamic binding trajectories conditioned on the initial dissociated state and contextual biophysical parameters. This task amalgamates elements of generation—generating novel trajectories that explore diverse binding modes—and prediction—predicting physically plausible outcomes based on empirical or simulated groundtruths. Specifically, given the sequences, structures, or representations of these unbound protein chains, along with environmental conditions (*e.g.*, temperature, ionic strength, or pH), the model predicts the time-resolved sequence of intermediate states and then generates the final composite structure.

2 Dataset Rationale

Bottleneck: Our dataset addresses a critical unmet need in AI-driven biology, where the transition from dissociated entities to the bound state remains poorly represented in existing resources. Current datasets, such as PDB [3] or Dynamic PDB [16], predominantly focus on static snapshots or single-entity trajectories, capturing equilibrium structures but failing to elucidate the kinetic pathways, intermediate states, and non-additive emergent behaviors inherent to complex formation. Our DynaPPI dataset is envisioned as a comprehensive repository of time-resolved recombination trajectories, encompassing not only multi-chain protein assemblies but also protein-ligand, protein-nucleic acid, and intra-molecular domain rearrangements. By integrating MD simulations with experimental validations, this dataset promotes diffusion models to generate unknown-complex structures, ultimately enabling breakthroughs in fields such as drug design, synthetic biology, and systems pharmacology.

Submitted to AI for Science workshop (NeurIPS 2025).

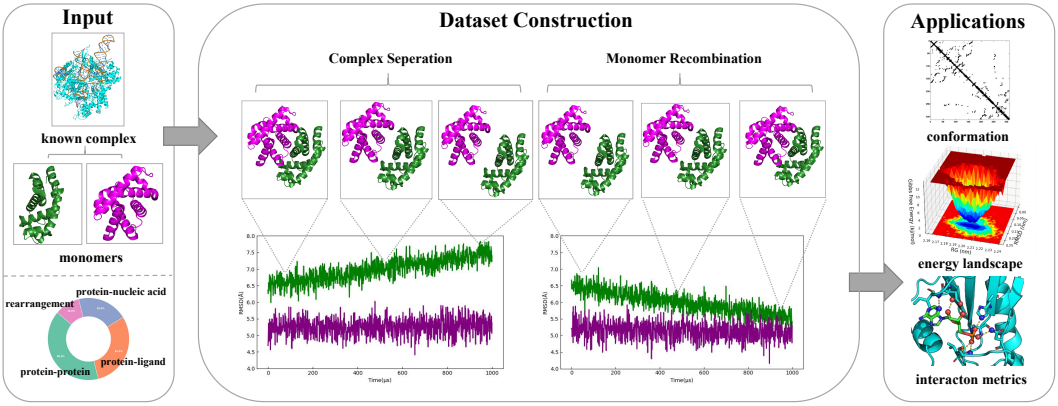

Figure 1: Overview of our dataset—input, construction, and applications. We disassemble known complexes from the PDB as references and record their MD recombination trajectories. Meanwhile, our dataset covers multiple categories of protein-related complexes, including protein-protein, protein-ligand, protein-nucleic acid, and intra-molecular rearrangements. We first explore the feasibility and effectiveness of the protein-protein split and provide a complete construction pipeline in Section A.

Data types: The dataset encompasses multi-modal data types, including time-series trajectories of atomic coordinates and physical properties, structural snapshots of clustered intermediate states, and auxiliary representations such as contact maps and free energy profiles, all formatted for efficient storage and AI compatibility (*e.g.*, DCD/NetCDF for trajectories and HDF5 for properties).

Scale: The dataset targets about $10,000$ complexes initially, with $5 \sim 20$ replicas each spanning from 10 ns to 1 ms, yielding totally $10^8 \sim 10^{11}$ frames, covering diverse biomolecular interactions. Moreover, the dataset ensures broad coverage across categories like protein-protein (40%), protein-ligand (30%), protein-nucleic acid (20%), and intra-molecular rearrangements (10%), with low sequence redundancy ($< 30\%$) and storage in the $10 \sim 100$ TB range for scalable open access.

Resolution: Temporal resolution is maintained at $1 \sim 10$ ps per frame to capture rapid conformational changes while controlling data volume, complemented by all-atom spatial precision (sub-Å accuracy in coordinates and forces) for detailed modeling of non-additive interactions, with optional coarser variants (*e.g.*, Cα-only backbones at 100 ps) for computational efficiency.

Labels and metadata needed: Essential labels include state indicators (*e.g.*, unbound/bound classifications), kinetic parameters, and thermodynamic metrics, supported by metadata such as source identifiers, environmental conditions, validation benchmarks, and AI-ready annotations like train/test splits and pre-computed embeddings to achieve robust, physics-informed model training.

3 Acceleration Potential

The far-reaching implications of our proposed dataset will extend to transformative applications in biomedicine and materials science. In drug discovery, it could rapidly simulate ligand-binding pathways, effectively identify transient pockets for allosteric inhibitors and accurately predict off-target interactions in polypharmacology. For synthetic biology, generative predictions could guide the design of self-assembling nanostructures or engineered enzymes with tunable affinities. Moreover, by incorporating multi-modal data (*e.g.*, integrating cryo-EM snapshots or fluorescence resonance energy transfer kinetics), the dataset further fosters interdisciplinary advancements, such as AI-accelerated virtual screening for pandemic preparedness or the rational engineering of biomolecular machines.

4 Scalability

The dataset exhibits strong scalability via leveraging parallelized MD simulations across distributed computing frameworks or using neural network potentials to approximate force fields. Key enablers include modular curation (automated PDB filtering and preprocessing), batch processing for replicas, and ***incremental growth by incorporating diverse biomolecular types***. At scale, it could encompass millions of trajectories, akin to the expanded version of Dynamic PDB, supporting broad AI training.

References

[1] Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617, 2025.

[2] 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.

[3] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[4] Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C Richardson. Molprobity: all-atom structure validation for macromolecular crystallography. *Biological crystallography*, 66(1):12–21, 2010.

[5] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

[6] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. protein crystallogr*, 40(1):82–92, 2002.

[7] Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon, and Nathan A Baker. Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic acids research*, 32(suppl_2):W665–W667, 2004.

[8] 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, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

[9] 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.

[10] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[11] Joël Janin, Kim Henrick, John Moult, Lynn Ten Eyck, Michael JE Sternberg, Sandor Vajda, Ilya Vakser, and Shoshana J Wodak. Capri: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2–9, 2003.

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

[13] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of computational chemistry*, 13(8):1011–1021, 1992.

[14] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the national academy of sciences*, 99(20):12562–12566, 2002.

[15] Roman A Laskowski, Malcolm W MacArthur, David S Moss, and Janet M Thornton. Procheck: a program to check the stereochemical quality of protein structures. *Applied Crystallography*, 26(2):283–291, 1993.

[16] Ce Liu, Jun Wang, Zhiqiang Cai, Yingxu Wang, Huizhen Kuang, Kaihui Cheng, Liwei Zhang, Qingkun Su, Yining Tang, Fenglei Cao, et al. Dynamic pdb: A new dataset and a se (3) model extension by integrating dynamic behaviors and physical properties in protein structures. *arXiv preprint arXiv:2408.12413*, 2024.

[17] 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.

[18] 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.

[19] Yusuke Naritomi and Sotaro Fuchigami. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *The Journal of chemical physics*, 134(6), 2011.

[20] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, 2011.

[21] Mats HM Olsson, Chresten R Søndergaard, Michal Rostkowski, and Jan H Jensen. Propka3: consistent treatment of internal and surface residues in empirical p k a predictions. *Journal of chemical theory and computation*, 7(2):525–537, 2011.

[22] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics*, 23(3):327–341, 1977.

[23] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, 1993.

[24] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of chemical theory and computation*, 11(11):5525–5542, 2015.

[25] Min-yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein science*, 15(11):2507–2524, 2006.

[26] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[27] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

[28] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of computational physics*, 23(2):187–199, 1977.

[29] Mihaly Varadi, John Berrisford, Mandar Deshpande, Sreenath S Nair, Aleksandras Gutmanas, David Armstrong, Lukas Pravda, Bissan Al-Lazikani, Stephen Anyango, Geoffrey J Barton, et al. Pdbe-kb: a community-driven resource for structural and functional annotations. *Nucleic Acids Research*, 48(D1):D344–D353, 2020.

[30] Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A Bates, Juan Fernandez-Recio, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, 427(19):3031–3041, 2015.

[31] Jianyi Yang, Ambrish Roy, and Yang Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.

[32] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

A Dynamic Recombination Trajectory Dataset for Protein Complexes

Inspired by Dynamic PDB [16], we propose the construction of a ***Dynamic Recombination Trajectory Dataset for Protein Complexes*** to capture the temporal evolution of protein complexes from separated chains to the bound state. This innovative dataset extends the pipeline of Dynamic PDB [16] to multi-chain systems while addressing the non-additivity of the binding process. Specifically, the dataset focuses on the preliminary feasibility of about 100 to $1,000$ complexes, simulated utilizing all-atom molecular dynamics (MD) simulation with OpenMM [8] for GPU acceleration and parallel optimization. Moreover, production runs are set from 10 ns to 1 ms per replica, inconsistent with Dynamic PDB's [16] 1 μs simulations, since binding events are typically uncertain, but monitored in real time and dynamically adjustable. Multiple replicas ensure stochastic sampling of binding paths.

In this section, we introduce the proposed dataset, detailing the preparation process, the molecular dynamics simulation, and the analysis of dynamic behaviors, respectively.

A.1 Source Selection and Data Curation

The original structures for the Dynamic Recombination Trajectory Dataset for Protein Complexes are sourced from the Protein Data Bank (PDB) [3], focusing on multi-chain entries with an experimental resolution of no greater than 2.5Å determined via X-ray diffraction to ensure high structural reliability. And in order to enhance annotation and diversity, we supplement PDB data [3] with specialized interaction databases, including PDBe-KB [29] for biological assemblies, BioLiP [31] for protein-protein interfaces (adapted from ligand contexts), and the Protein-Protein Docking Benchmark (version 5.0) [30], which provides over 230 non-redundant complexes with validated binding modes.

Selection criteria prioritize systems amenable to molecular dynamics (MD) while promoting diversity: To maintain computational tractability, protein complexes are restricted to $2 \sim 4$ chains, with individual chain length ≤ 300 residues and a total system size $\leq 50,000$ atoms (including solvent). The dataset composition consists of 40% homo-oligomers (*e.g.*, symmetrical dimers such as HIV-1 protease), 40% hetero-complexes (*e.g.*, barnase-barstar complex for high-affinity binding), and 20% transient interactions (*e.g.*, ubiquitin-conjugating enzyme pairs). This balance ensures representation across interface types, such as hydrophobic cores and electrostatic complementarity, as well as functional classes like enzyme-inhibitor and signaling interactions. Furthermore, we exclude transmembrane complexes, large assemblies (> 4 chains), and covalently linked systems to minimize simulation artifacts, focusing instead on non-covalent, soluble proteins.

For each selected complex, individual chains are first extracted from the bound PDB structure [3] and isolated to simulate the unbound state (more details in Section B). Then these separate chains are translated $5 \sim 20$ nm apart with randomized orientations per replica, employing some toolkits like MDAnalysis [18] or VMD [10] for coordinate manipulation. This configuration can effectively mimic diffusive encounters, achieving clear observation of the approach, collision, and binding phases.

The pilot phase targets 500 complexes (*e.g.*, 250 dimers and 250 trimers), with the final dataset expanding to about $5,000$ entries. To avoid redundancy, complexes are further clustered by sequence similarity ($< 30\%$) using MMseqs2 [26]. Additionally, metadata, including PDB IDs, chain counts, and literature-derived affinities, are tracked in a SQLite database. And the original structures can be downloaded via the PDB API or the official website[1], with Biopython [5] for parsing.

A.2 Preprocessing of Protein Data

Prior to simulations, the original structures must undergo rigorous preprocessing to address PDB limitations such as missing residues and non-standard elements. The specific gap ***repair*** scheme is as follows: for segments ≤ 5 residues, we utilize MODELLER [23] for homology-based loop modeling, combined with DOPE scoring [25] for energetic evaluation; for larger gaps (> 5 residues) or entire loops, AlphaFold-Multimer [9] can provide context-aware predictions, accessible via ColabFold.

Cleaning involves: ***1)*** removing all heteroatoms from the protein structures, including water molecules, ligands, and metal ions, to focus on the dynamic behavior between proteins; ***2)*** mapping non-standard residues (*e.g.*, selenomethionine to methionine) with Open Babel [20]; and ***3)*** protonat-

[1] https://www.rcsb.org/

Table 1: Attributes of the proposed dataset.

Name	Data Type	Shape	Description	Unit
Structural Information				
Replica ID	int8	(1)	Identifier of Replica	-
Position	float32	$(N_{frames}, N_{atoms}, 3)$	Trajectory Coordinates	Å
Distance	float32	$(N_{frames}, N_{chains} * (N_{chains} - 1)/2)$	Inter-Chain Distances[2]	Å
Contact Map	bool	$(N_{frames}, N_{residues}, N_{residues})$	Binary Contact Maps[3]	-
Dynamic and Physical Property				
Velocity	float32	$(N_{frames}, N_{atoms}, 3)$	Trajectory Velocities	Å/ps
Force	float32	$(N_{frames}, N_{atoms}, 3)$	Trajectory Forces	kcal/mol · Å
Potential Energy	float32	(N_{frames})	System Potential Energy	kJ/mol
Kinetic Energy	float32	(N_{frames})	System Kinetic Energy	kJ/mol
Total Energy	float32	(N_{frames})	System Total Energy	kJ/mol
Temperature	float32	(N_{frames})	System Temperature	K
Pressure	float32	(N_{frames})	System pressure	bar
Box Volume	float32	$(N_{frames}, 3)$	System Volume Forces	nm^3
Density	float32	(N_{frames})	System Density	g/ml
Binding-Specific Metric				
Interaction Energy	float32	$(N_{frames}, N_{chains} * (N_{chains} - 1)/2)$	Inter-Chain Energy[4]	kJ/mol
Hydrogen Bond	int16	(N_{frames})	Number of Hydrogen Bonds	-
Salt Bridge	int16	(N_{frames})	Number of Salt Bridges	-
Binding Status	bool	(N_{frames})	Binding Status Flag	-
ΔSASA[5]	float32	(N_{frames}, N_{chains})	SASA Changes	$Å^2$
Final Status	bool	(1)	Status for Prolongation[6]	-

ing at physiological pH 7 using PROPKA3 [21] or PDB2PQR [7] for accurate charge assignment. Moreover, explicit hydrogen atoms are added via MODELLER [23] to achieve all-atom completeness.

We further ***validate*** the preprocessing statistics, including completion rates (*e.g.*, percentage relying on AlphaFold *vs*. MODELLER) and gap size distributions, aiming for a rejection rate $< 10\%$ due to irreparable issues. After cleaning, isolated chains may undergo a brief energy minimization to relax strains. And initial configurations for the separated state incorporate random rotations and translations via PyMOL scripts [6], with clash detection leveraging MolProbity [4] to confirm no structural overlaps. Structural integrity also need to be validated with PROCHECK [15], ensuring Ramachandran plot outliers $< 1\%$; structures exceeding 5% issues are discarded.

A.3 MD Simulation Setup and Execution

The simulation environment is designed to mimic physiological conditions, extending Dynamic PDB's setup [16] for multi-chain mobility. Specifically, all-atom molecular dynamics simulations are conducted using OpenMM [8] in conjunction with the Amber-ff14SB force field [17], which effectively governs protein interactions and further enhances the accuracy of protein side chain and backbone parameters. Meanwhile, isolated protein chains are solvated in a truncated octahedral or cubic periodic box with a padding thickness of ≥ 1.5 nm to prevent self-interaction artifacts during diffusion. And this box is filled with TIP3P water molecules for hydration accuracy, and subsequently neutralized and salted with Na^+/Cl^- ions at a concentration of $150\ mM$.

Equilibration begins with an energy minimization process via steepest descent followed by conjugate gradient, targeting a force tolerance of $2.39\ kcal/mol \cdot Å$ (up to $10,000$ steps) to resolve bad contacts. The canonical ensemble (NVT) phase runs for 1 ns at $300\ K$ using the *LangevinMiddleIntegrator* (friction coefficient of $1.0\ ps^{-1}$ and time step of 2 fs with SHAKE constraints [22] for hydrogen atoms), initially restraining protein heavy atoms (force constant of $10\ kcal/mol \cdot Å^2$ with gradual

[2]The distance denotes pairwise center-of-mass distances between chains.

[3]Binary contact maps (1 if distance ≤ 4.5Å, 0 otherwise) for inter- and intra-chain interactions.

[4]This energy denotes decomposed inter-chain interaction energy via MMPBSA.

[5]SASA denotes solvent-accessible surface area.

[6]This status indicates whether the simulation has extended beyond initial duration.

release). And the isothermal-isobaric ensemble (NPT) phase follows for 1 ns at 1 bar, employing the *Monte Carlo Barostat* (updates every 100 steps) with the same integrator to equilibrate density.

Production runs are set from 10 ns to 1 ms per replica, scaled based on complex size and anticipated binding timescales (*e.g.*, fast binders like barnase-barstar might stabilize in 10 ns). The hydrogen mass redistribution in OpenMM [8] is implemented with a time step of 2 fs to achieve numerical stability. Moreover, to account for energy barriers during binding and capture non-additive effects such as induced fit, enhanced sampling techniques are innovatively integrated: umbrella sampling [28] applies bias potentials along the center-of-mass distance reaction coordinate, while metadynamics [14] (via the PLUMED plugin in OpenMM [8]) adds history-dependent biases. For diffusion-limited phases, $5 \sim 20$ replicas per complex are seeded with different random seeds, with Temperature replica-exchange (T-REMD) [27] performed in the $300 \sim 350$ K range to facilitate barrier crossing.

Real-time monitoring detects binding (*e.g.*, > 20 inter-chain contacts) via custom scripts, allowing for early termination of converged runs to optimize computation. The simulations are run on NVIDIA A100 GPUs with 80 GB of memory and parallelized with MPI in OpenMM [8]. During the pilot phase, the total GPU hours per machine are expected to be $50 \sim 200$.

A.4 Data Recording and Physical Properties

Data recording intervals balance resolution and storage: atomic coordinates are recorded every 10 ps and physical properties are recorded every 1 ps, yielding 1K$\sim$ 10K frames per ns (approximately 10 GB per complex). Table 1 provides a detailed overview of the data attributes associated with each replica, containing structural information, dynamic and physical properties, and binding-specific metrics. This structured format supports subsequent analyses and interpretations of the inter-chain binding dynamic behaviors and properties of the complexes within the dataset.

Metadata identifies binding events (*e.g.*, timestamps of stable contacts), replica IDs, and prolongation status for extended runs. Trajectories are stored in DCD or NetCDF formats, and attributes are stored in HDF5 for efficient access. Ultimately, all data is compressed with gzip.

A.5 Post-Processing, Analysis, and Validation

Post-simulation processing aligns trajectories to reference bound structures using MDAnalysis [18], thereby correcting periodic boundary artifacts. States (i.e., unbound, transient intermediates, and bound) are identified via time-lagged independent component analysis (tICA) [19] and k-means clustering with PyEMMA [24], enabling quantification of transition rates using Markov State Models.

The analysis emphasizes non-additivity: per-chain and complex-wide RMSD/RMSF track conformational changes; evolution of the radius of gyration illustrates compactness shifts; free energy landscapes are reconstructed via WHAM [13] on biased simulations; and interface remodeling is assessed through variations in residue contact frequency.

Further quantitative validation ensures the dataset's internal quality (*e.g.*, energy drift $< 1\%$, RMSF alignment with Dynamic PDB single-chain benchmarks), as well as external quality (*e.g.*, bound-state RMSD < 3Å *vs.* experimental PDBs; binding affinities from MMPBSA correlating > 0.7 *vs.* experimental values from literature; and association rates from Markov State Models *vs.* experimental values from literature). And subsets are benchmarked against CAPRI [11] and Docking Benchmark 5.0 [30] to assess pose accuracy. Additionally, unstable runs (*e.g.*, DSSP [12] secondary structure loss $> 20\%$) are filtered, with a target success rate of $> 80\%$.

The compiled dataset is split into train/validation/test $(80/10/10)$, including raw trajectories, processed states, and access APIs, and will be hosted on Zenodo or Dryad with a DOI.

A.6 Preliminary Timetable and Rough Budget

The implementation unfolds in phases: selection and preprocessing ($6 \sim 8$ weeks, emphasizing manual curation), MD simulations ($8 \sim 16$ weeks, via automated batching), and analysis/validation ($4 \sim 6$ weeks, iterative refinement). The rough budget ranges from $9,000$ \$ $\sim 15,000$ \$ for cloud computing (*e.g.*, AWS p3.2xlarge at 3 \$ per hour), with open tools minimizing additional expenses; academic resources (*e.g.*, NSF/NIH clusters with SLURM) could further reduce this for scale-up.

B Initial Separation of Reference Complexes

The initial separation of protein chains from bound complex structures in the PDB [3] is critical for simulating realistic recombination dynamics. However, direct extraction from bound PDB entries retains conformational biases induced by inter-chain interactions, such as interface remodeling or induced-fit changes, which deviate from the native unbound (apo) states. This can underestimate energy barriers and non-additive effects during binding, leading to artificially accelerated or biased trajectories. To mitigate this, we prioritize or approximate unbound conformations, ensuring starting structures more accurately reflect free-chain dynamics. The procedure comprises three sequential steps: (1) chain extraction, (2) unbound state approximation, and (3) geometric separation.

Step 1: Chain Extraction. First, load the bound PDB structure into a molecular analysis framework. After that, parse and isolate individual chains by selecting atoms based on chain identifiers (*e.g.*, 'A', 'B'). Then export each chain as a separate PDB file, preserving atomic coordinates, residues, and any existing hydrogen atoms. Finally, validate completeness by checking for gaps or clashes using stereochemical assessment tools, rejecting structures with $> 5\%$ unresolved residues.

Step 2: Unbound State Approximation. Approximate native unbound conformations to remove bound-state bias: First, query databases (*e.g.*, PDBe-KB [29], UniProt [1]) for experimental unbound structures of each chain (sequence identity $> 95\%$). If available, align to the bound reference using TM-align structural superposition [32] and adopt as the starting model. If unbound structures are unavailable, extract chains from the bound PDB and relax via short molecular dynamics (MD) simulations: Solvate each chain individually in a TIP3P water box with $150\ mM$ NaCl, minimize energy (tolerance $2.39\ kcal/mol{\cdot}$Å), equilibrate (NVT/NPT, 1 ns each at $300\ K$), and run production for $5 \sim 10$ ns with Langevin integration (2 fs timestep). Cluster resulting frames (*e.g.*, using tICA [19] and k-means) and select the dominant conformation (RMSD > 1Å from bound state). Moreover, as an alternative or augmentation for incomplete chains, predict unbound models from sequences using AlphaFold3 [2] in monomer mode, aligning outputs to bound equivalents for coordinate consistency.

Step 3: Geometric Separation. Following the previous step, recombine approximated unbound chains into a single structure and compute centers of mass for each chain. After that, apply random rotations (Euler angles uniformly sampled from $0 \sim 360^{\circ}$) and translations (vectors yielding $5 \sim 20$ nm inter-chain separation, randomized per replica) to mimic diffusive encounters. Then assign initial velocities from a Maxwell-Boltzmann distribution at $300\ K$. Ultimately, validate final configurations for clashes (minimum distance >0.5 nm) and export as a merged PDB file for solvation.

In conclusion, this elaborately designed process of complex separation effectively enhances simulation fidelity by starting from unbound-like states, better capturing non-additive binding effects, at a modest computational cost ($1 \sim 2$ GPU-hours per chain for relaxation/prediction).