DynPro: A Large-Scale Dataset of Molecular Dynamics Simulations for Protein Conformational Ensembles and Transitions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Protein dynamics underpin critical biological processes, yet existing datasets for AI-driven modeling are limited to short timescales and local fluctuations, failing to capture broad conformational ensembles, transitions, and complex interactions essential for drug design and biomedicine. We propose DynPro, a large-scale, openly shareable dataset comprising enhanced molecular dynamics (MD) simulations for tens of thousands of protein systems. Each system features at least 100 μs of effective simulation time via adaptive sampling techniques, providing atomistic trajectories, Boltzmann-weighted free energies, and kinetic metadata in mmCIF format. DynPro enables generative AI to capture long-timescale ensembles and rare transitions, addressing computational and data bottlenecks. Built from public PDB structures with advanced MD simulations on HPC clusters ($50-100M), it provides a transformative resource for drug design, disease mechanism studies, and synthetic biology, establishing a new paradigm in AI-driven structural dynamics.

1 Al task definition

The proposed dataset, DynPro, aims to enable AI models to address the fundamental scientific question: How can we accurately generate broad conformational ensembles and transition pathways for proteins and their complexes in biologically relevant timescales? This is primarily a generation task, where AI models (e.g., diffusion-based [1, 2, 3, 4]or autoregressive generative models[5, 6, 7]) would predict atomistic trajectories of protein dynamics[8, 9, 10], including equilibrium ensembles[11, 12], metastable states[13, 14], and rare transitions between conformations[15, 16]. Secondary tasks could include prediction of free energy landscapes from partial trajectories or classification of functional states (e.g., active vs. inactive conformations)[17, 18]. By providing ground-truth data from long-timescale simulations, DynPro would empower models to simulate protein flexibility far beyond current capabilities, akin to how AlphaFold[19, 20] revolutionized static structure prediction. The task is critical for fields like structural biology and drug design, where protein dynamics underpin functions such as enzyme catalysis, signaling, and ligand binding.

2 Dataset rationale

Access to DynPro would transform AI model development by providing the "ImageNet" equivalent for protein dynamics—high-quality, diverse training data for generative models. At scale, DynPro lets models learn to extrapolate from short seeds to long-horizon ensembles and rare transitions, converting days of MD per target into minutes of inference, which shifts the cost profile of dynamics from compute-bound to data-bound. This would accelerate downstream science in:

Submitted to the AI for Science workshop(NeurIPS 2025). Do not distribute.

- **Drug design at population scale:** Ensemble-aware docking and generative design over previously undruggable targets, capturing cryptic pocket emergence, allosteric pathways, and ligand-induced vs. conformational-selection mechanisms. Expected outcomes: higher virtual screening hit rates, better selectivity, and earlier go/no-go decisions.
- **Synthetic Biology:**Rapid evaluation of engineered loops, domain swaps, and sensor designs by predicting flexibility windows and switching kinetics, turning months of iterative MD into a compile-time check.
- **Cross-Disciplinary Impact:**Systematic inclusion of membrane proteins, intrinsically disordered regions, multi-protein assemblies, and nucleic-acid complexes—domains historically starved of long-timescale data—yields foundation models that generalize beyond soluble monomers.

By open-sharing DynPro, we anticipate rapid adoption, fostering competitions for dynamics prediction and integrating with tools like OpenFold[21] or DiffDock[22] for end-to-end pipelines.

3 Acceleration potential

Current datasets for protein dynamics are bottlenecks due to insufficient timescale, diversity, and coverage of conformational space. ATLAS[23] (100 ns) captures only local fluctuations, while MISATO[24] (19,443 protein–ligand complexes) is confined to nanosecond dynamics. Critically, protein–protein complexes—central to signaling and cellular function—remain without large-scale, long-timescale datasets. DynPro addresses this by providing μs-to-ms timescale simulations with enhanced sampling to ensure broad conformational coverage. Data types include:

- **Trajectories:** Multi-resolution output, featuring compressed coordinates at ns intervals for the complete dataset and ps-resolution raw data for rapid conformational transitions.
- **Scale:** Tens of thousands of systems (according to PDB database cluster of 30% similarity of 42,096 structures), each with at least 100 μs simulation time (via enhanced sampling).
- **Resolution:**All-atom with explicit solvent/ions (and lipid bilayers where relevant), standardized force-field stacks, and pinned engine versions for reproducibility.
- **Labels:**Boltzmann-weighted free energies for conformations, kinetic rates for transitions. To facilitate integration with related models, data will be formatted in mmCIF for static snapshots, with .nc or .xtc extensions for trajectories.

This dataset is the bottleneck because existing ML models for dynamics[25, 26, 27] suffer from poor generalization to long-timescale events, leading to inaccurate predictions of druggable states or protein interactions.

4 Data-creation pathway

Data will be generated using enhanced and accelerated sampling methods such as adaptive sampling, metadynamics, replica exchange, or Gaussian accelerated MD on computing clusters[28]. All simulations will be conducted at the all-atom level to capture detailed atomic interactions and dynamics. Sources include protein-ligand complexes starting from PDB structures (e.g., expanding MISATO), protein-protein complexes from PDB or predicted via AlphaFold-Multimer [29], focusing on biologically relevant interfaces from the STRING [30] database, and nucleic acid structures complexes from PDB or predicted models, emphasizing functional motifs like binding sites or regulatory elements. Simulations will use AMBER [31] or GROMACS [32] with state-of-the-art force fields , ensuring comprehensive coverage of conformational transition states via collective variables that facilitate barrier crossing and exploration of metastable states. For each system, aggregate simulation trajectories will exceed 100 μs to achieve sufficient sampling of rare events and equilibrium distributions.

5 Cost & Scalability

Generating 1 μs of enhanced sampling per system costs $10-20 on cloud HPC. For 50,000 systems of 100 μs, total budget is $50-100 million, scalable via parallelization. Cost reductions could come from emerging technologies like AI-accelerated MD . Phased rollout (e.g., 1,000 systems initially) allows iterative validation.

References

[1] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

[4] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

[5] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

[6] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7327–7347, 2021.

[7] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[8] Shiru Wu, Xiaowei Yang, Xun Zhao, Zhipu Li, Min Lu, Xiaoji Xie, and Jiaxu Yan. Applications and advances in machine learning force fields. *Journal of Chemical Information and Modeling*, 63(22):6972–6985, 2023.

[9] Yuanqing Wang, Kenichiro Takaba, Michael S Chen, Marcus Wieder, Yuzhi Xu, Tong Zhu, John ZH Zhang, Arnav Nagle, Kuang Yu, Xinyan Wang, et al. On the design space between molecular mechanics and machine learning force fields. *Applied Physics Reviews*, 12(2), 2025.

[10] Mingan Chen, Xinyu Jiang, Lehan Zhang, Xiaoxu Chen, Yiming Wen, Zhiyong Gu, Xutong Li, and Mingyue Zheng. The emergence of machine learning force fields in drug design. *Medicinal Research Reviews*, 44(3):1147–1182, 2024.

[11] Yaowei Jin, Qi Huang, Ziyang Song, Mingyue Zheng, Dan Teng, and Qian Shi. P2dflow: A protein ensemble generative model with se (3) flow matching. *Journal of Chemical Theory and Computation*, 21(6):3288–3296, 2025.

[12] Giacomo Janson, Gilberto Valdes-Garcia, Lim Heo, and Michael Feig. Direct generation of protein conformational ensembles via machine learning. *Nature Communications*, 14(1):774, 2023.

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

[14] Li-E Zheng, Shrishti Barethiya, Erik Nordquist, and Jianhan Chen. Machine learning generation of dynamic protein conformational ensembles. *Molecules*, 28(10):4047, 2023.

[15] Wei Lu, Jixian Zhang, Weifeng Huang, Ziqiao Zhang, Xiangyu Jia, Zhenyu Wang, Leilei Shi, Chengtao Li, Peter G Wolynes, and Shuangjia Zheng. Dynamicbind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1):1071, 2024.

[16] Heng Ma, Debsindhu Bhowmik, Hyungro Lee, Matteo Turilli, Michael T Young, Shantenu Jha, and Arvind Ramanathan. Deep generative model driven protein folding simulation. *arXiv preprint arXiv:1908.00496*, 2019.

[17] 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*, page eadv9817, 2025.

[18] Amanda A Volk, Robert W Epps, Daniel T Yonemoto, Benjamin S Masters, Felix N Castellano, Kristofer G Reyes, and Milad Abolhasani. Alphaflow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nature Communications*, 14(1):1403, 2023.

[19] 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.

[20] 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.

[21] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature methods*, 21(8):1514–1524, 2024.

[22] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

[23] 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.

[24] Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Kieran Didi, André Santos Dias Mourão, Radosław Kitel, Pietro Liò, Stefan Kesselheim, Marie Piraud, et al. Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature computational science*, 4(5):367–378, 2024.

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

[26] Aniketh Iyengar, Jiaqi Han, Pengwei Sun, Mingjian Jiang, Jianwen Xie, and Stefano Ermon. Align your structures: Generating trajectories with structure pretraining for molecular dynamics. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*.

[27] Ziyang Yu, Wenbing Huang, and Yang Liu. Unisim: A unified simulator for time-coarsened dynamics of biomolecules. *arXiv preprint arXiv:2506.03157*, 2025.

[28] Wenhui Shen, Tong Zhou, and Xinghua Shi. Enhanced sampling in molecular dynamics simulations and their latest applications—a review. *Nano Research*, 16(12):13474–13497, 2023.

[29] 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.

[30] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[31] David A Case, H Metin Aktulga, Kellon Belfon, Ido Y Ben-Shalom, Joshua T Berryman, Scott R Brozell, David S Cerutti, Thomas E Cheatham III, G Andrés Cisneros, Vinícius Wilian D Cruzeiro, et al. *Amber 2023*. University of California, San Francisco, 2023.

[32] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.