DINO: dynamics-informed dataset to overcome the limitations of static molecular data in AI-driven drug discovery

Eva Smorodina
Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway
eva.smorodina@medisin.uio.no

Rahmad Akbar
InSilico Antibody Discovery, AI and Digital Innovation, Novo Nordisk Park 2, 2760, Måløv, Denmark
rdak@novonordisk.com

Victor Greiff
Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway
Imprint Labs, LLC, New York, NY, USA
victor.greiff@medisin.uio.no

Abstract

DINO, a dataset of molecular movements, will bridge the gap between static structural models and real molecular function, enabling AI-driven drug discovery to move beyond rigid docking toward biophysically accurate, dynamic design. By integrating experimental and synthetic MD data, this dataset will provide the missing link for mechanistic AI models, unlocking a new era of high-success-rate, computationally driven therapeutic development.

1 AI task definition

1.1 Background

This paper is submitted to the Dataset Proposal Competition at the NeurIPS 2025 AI for Science workshop. The competition challenges the community to envision the next foundational scientific dataset (similar in impact to the Protein Data Bank, ImageNet, the Human Cell Atlas, or the UK Biobank) that could catalyze a major leap in AI-accelerated research. In this spirit, we propose DINO, a large-scale dataset of experimentally grounded and simulation-augmented molecular movements designed to update today's static structural information with biophysically realistic dynamics and advance the next stage of molecular modeling, computational biophysics, and drug design.

1.2 Problem statement

Current computational drug discovery relies overwhelmingly on static molecular structures from, for example, Protein Data Bank (PDB) [6] or PubChem [18]. Despite the ability of the latest AI models (such as BioEmu [19], AlphaFlow [16]) to generate conformational ensembles that aim to represent molecular movements, the resulting dynamics suffer from critical limitations:

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science workshop.

1. **State bias**: overrepresentation of a limited number of biophysical states (e.g., disordered molecules are underrepresented) [24, 34, 28, 12, 14]
2. **Dynamics blind spots**: omission of functionally essential motions (e.g., only one conformation per structure) [30, 27]
3. **Disconnected sampling**: conformational diversity is captured without meaningful connection to real-world behavior (e.g., sampled rearrangements or large motions inconsistent with actual molecular conformations) [25, 26]
4. **Property-function disconnect**: Modeled drug-related biophysical properties (e.g., expression, solubility, stability, immunogenicity, etc.) often fail to reflect experimental (e.g., computationally predicted as non-toxic molecule can be toxic experimentally) [9, 2]

Due to these limitations, current AI models for drug design often generate unrealistic molecules [5], requiring the production of massive numbers of candidates. Their functional potential can only be evaluated in experimental assays, where true feasibility and biochemical functionality are revealed [29, 9].

1.3 Proposed solution

Previously, it was shown that including dynamics improves functional representation of biomolecules, making the computational results in line with experiments [26]. Here, we propose DINO (short for "dynamics," and who doesn't love dinosaurs?) – a dataset of atomistic movements designed to capture the fundamental principles of molecular dynamics (MD) to represent real-world behavior of the molecules and incorporate it into AI models for drug discovery (Appendix A). This dataset will provide multi-scale, residue-to system-level insights into structural flexibility, conformational ensembles, binding energetics, and functional kinetics of proteins, lipids, nucleic acids, small molecules, and complexes, enabling models to learn biologically meaningful motion and function.

DINO will overcome the problem of functionally irrelevant drug candidates by ensuring that molecular designs are filtered through precise thermodynamic and dynamic constraints from the beginning of the drug discovery campaign. By embedding realistic movement and energetic feasibility into the training process, AI models will be able to predict biochemically plausible molecules rather than relying solely on static assumptions (Appendix B).

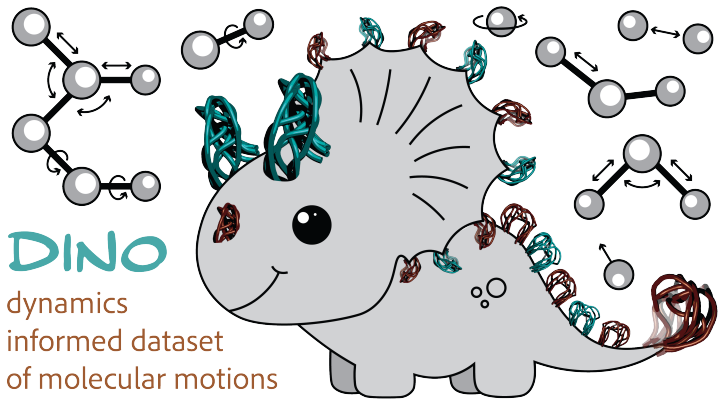

Figure 1: DINO (dynamics-informed dataset of molecular motions) captures the essential atomistic motions that underlie realistic molecular behavior, represented here as dynamic conformational ensembles decorating the cartoon dinosaur. The surrounding schematic motifs highlight fundamental modes of molecular motion sampled through molecular dynamics (MD). By incorporating these dynamic trajectories, DINO aims to provide multi-scale information on structural flexibility, conformational ensembles, binding energetics, and functional kinetics across biomolecular systems. This dataset is designed to address the limitations of static modeling and enable AI frameworks to learn biologically meaningful motions, ensuring that computationally generated molecules and drug candidates reflect physically and thermodynamically plausible behavior.

2 Dataset rationale

2.1 Limitations of the current data

To learn the fundamental principles of biomolecular movement, we need a comprehensive, multi-scale dataset integrating experimental and synthetic dynamics across all major classes of biomolecules (proteins, nucleic acids, small molecules, lipids), capturing local, intermediate, global, and mesoscale motions, with standardized labels for flexibility, kinetics, energetics, and function (Appendix C, Appendix D).

Currently available data is insufficient because:

- **Static and biased resources**: Existing datasets (e.g., PDB [6]) overrepresent bound, rigid conformations while underrepresenting unbound states, intermediates, and disordered ensembles
- **Fragmented MD coverage**: Current simulations are scattered across studies, with no systematic, multi-class collection (even though there are attempts to create a united MD database [1]) spanning proteins, nucleic acids, small molecules, membranes, and assemblies. This leaves large gaps in capturing the full landscape of molecular motions
- **Lack of experimental validation**: Available MD data are rarely paired with time-resolved experimental dynamics (cryogenic electron microscopy (cryo-EM), hydrogen deuterium exchange mass spectrometry (HDX-MS), nuclear magnetic resonance (NMR), Förster resonance energy transfer (FRET), etc.), limiting thermodynamic realism
- **AI blind to motion and molecular interactions**: Without dynamic constraints, models generate rigid, unrealistic candidates, inflating false positives in binding, catalysis, and drug design

2.2 What kind of data is needed?

1. **Synthetic MD Data**: Generated internally and from open-source workflows, integrated with public archives (MDDB [1], mdCATCH [22], MoDEL [21], etc.), including: a) All-atom MD simulations (μs–ms timescales) for 1000+ systems, b) Quantum MD simulations (QM/MM, Ab Initio) for high-resolution dynamics, c) Coarse-grained simulations for large assemblies, d) ML-generated conformational ensembles
2. **Experimental MD Data**: Drawn from public archives (PDB [6], PubChem [18], ZINC [15], NDB [15], SAbDab [10], EMDB [31], etc.) and new collaborative experiments, including: a) Time-resolved HDX-MS, b) Cryo-EM ensembles, c) Surface plasmon resonance (SPR) and biolayer Interferometry (BLI) kinetics, d) NMR relaxation, e) Single-molecule FRET
3. **Scale**: 1000+ diverse molecular all-atom systems
4. **Resolution**: a) Atomic-level: all-atom MD, QM/MM, b) Residue-level: coarse-grained, cryo-EM, c) Functional: kinetic rates, binding affinities, protections, etc.
5. **Labels/Metadata**: Standardized structural, energetic, and functional metrics, including: root mean square fluctuations (RMSF), solvent-accessible surface area (SASA), binding free energy (ΔG), kinetic rates (k_{on}, k_{off}, k_{D})), developability properties (stability, solubility, charge, etc.)

Appendix

A Acceleration potential

DINO will shift AI in drug discovery from static averages to dynamics-driven design, enabling models to propose candidates that are not only synthetically possible but also thermodynamically and functionally realistic (Appendix E).

- **Biophysical interpretability**: Provide AI with explanations grounded in dynamics, such as predicting failure due to over-constrained target flexibility

- **Mechanistic AI models**: Move beyond rigid docking to dynamic interaction prediction, improving success rates in binder (RNA, DNA, antibody, peptide, small molecule, etc.) design
- **Generalizable biological principles**: Reveal quantitative structure–dynamics–function rules for rational engineering across molecular modalities (proteins, nucleic acids, small molecules, etc.)
- **Hybrid physics-AI approaches**: Fine-tune generative models with dynamics-aware constraints, reducing brute-force screening by orders of magnitude

B What scientific question (prediction, generation, or other task) will the dataset enable?

1. **Prediction tasks**:
   - Binding affinity, specificity, and kinetics across proteins, nucleic acids, small molecules, lipids, and complexes
   - Developability metrics such as stability, solubility, aggregation, informed by flexibility and disorder
   - Functional descriptors including allosteric coupling, catalytic rates, and conformational switching
2. **Generation tasks**:
   - Designing protein, antibody, DNA, RNA, small molecule, and peptide drugs with native-like flexibility and realistic conformational ensembles
   - Proposing small molecules and cofactors that respect entropic penalties, solvation effects, and binding dynamics
   - Creating membrane proteins and assemblies that account for lipid interactions and mesoscale dynamics
   - Incorporating and reflecting the effects of post-translational modifications (PTMs) on the molecular behavior
3. **Hybrid tasks**:
   - AI-guided refinement of static structural predictions (from crystallography, AlphaFold [17], etc.) into dynamic conformational landscapes
   - Integrating quantum-level dynamics for metalloproteins and small-molecule catalysis with larger-scale MD ensembles
   - Linking local motions (rotamers, base flipping) to global motions (domain rearrangements, folding, phase separation) for cross-class generalization

C Data types, scale, resolution, labels, and metadata needed to solve the problem

1. **Synthetic MD data**:
   - Generated internally and through open-source simulation workflows, standardized, and stored in a shared repository. Public archives such as MDDB [1], mdCATCH [22], and MoDEL [21], etc. will also be integrated.
   - All-atom MD simulations (μs–ms timescales): 1000+ systems across proteins (and here subdivided to protein switches, membrane proteins, antibodies, intrinsically disordered proteins, and enzymes), nucleic acids, lipids, small molecules, and complexes, covering epitopes, catalytic sites, allosteric networks, and other types of structural and dynamic arrangements. Generated using GPU-accelerated engines (OpenMM [11], AMBER [8], GROMACS [4]).
   - Quantum MD simulations (QM/MM, Ab Initio): High-resolution dynamics of small molecules, cofactors, and metalloproteins, leveraging platforms such as Gaussian [13], GAMESS [3], PLUMED [7], and ORCA [23].

- Coarse-grained simulations (Martini [20], CALVADOS [32], custom CG): Scalable sampling for large assemblies, membrane proteins, and phase-separating systems.
- ML-generated conformational ensembles (BioEmu [19], AlphaFlow [16], AI2BMD [33] new models that will emerge): Augment experimental and simulated data with probabilistic conformational states.

2. **Experimental MD data**:
   - Drawn from both public archives and new experiments conducted with collaborators. Sources include curated entries from PDB [6], PubChem [18], ZINC [15], NDB [15], SAbDab [10], EMDB [31], etc., and dynamic datasets such as MDDB [1] (curated from publications and Zenodo).
   - Time-resolved HDX-MS: Residue-level mapping of flexibility and solvent exposure.
   - Cryo-EM ensembles: Multiple conformational states (bound, intermediate, unbound) for protein, RNA, and antibody complexes.
   - SPR/BLI: Binding kinetics (k_{on}, k_{off}), k_{D})) linking motion to affinity.
   - NMR relaxation: Side-chain and backbone flexibility, particularly for disordered proteins and small molecules.
   - Single-molecule FRET: Direct measurement of conformational transitions in solution.
   - New experiments: Focused on high-value therapeutic and biologically diverse systems (e.g., HER2, PD-1, TNF-α, RNA aptamers, enzyme families) in collaboration with academic and industrial labs.
3. **Metadata and labels**:
   - Derived from both synthetic and experimental pipelines, standardized and synchronized across datasets.
   - Structural and dynamic metrics: RMSF, SASA, hydrogen bond lifetimes, interaction networks (H-bonds, salt bridges, hydrophobic clusters, $\pi - \pi$ stacking).
   - Energy and kinetic metrics: Binding free energies (ΔG), kinetic rates (k_{on}, k_{off}, k_{D})), catalytic turnover, entropic penalties, and solvation energies.
   - Functional and developability metrics: Aggregation propensity, solubility, stability, allosteric coupling, and conformational switching, validated where possible by experiment.
4. **Data standardization across experiments**:
   - MD simulation setup will use consistent or explicitly compatible parameter sets across biomolecular classes (e.g., matching force fields, water models, ion parameters, box types, temperature/pressure controls), while still allowing system-specific (membrane proteins, IDPs, nucleic acids, or small molecules, etc.) optimizations needed to reflect precise biophysical behavior
   - Unified postprocessing pipelines will standardize all computational outputs into common formats, including coordinate files (mmCIF), trajectories (XTC), residue and chain numbering conventions, protonation states, and metadata schemas describing simulation conditions and their accuracy.
   - Experimental data will be curated to preserve comparable experimental settings within each biomolecular class (e.g., consistent buffers, temperature, pH, ionic strength), or accompanied by reference controls that permit cross-experiment normalization.
   - Heterogeneous experimental modalities (HDX-MS, smFRET, cryo-EM ensembles, NMR relaxation, SPR/BLI) will be harmonized by defining shared reference conditions, uncertainty estimates, and mapping rules that link experimental observables to structural and dynamic features.
   - Both computational and experimental data will include confidence metrics (low, medium, high) that reflect the accuracy and compatibility of each parameter, as well as the estimated noise and error relative to similar public datasets.
   - A unified ontology will connect sequence, structure, dynamics, energetics, and experimental conditions, enabling multi-modal mapping between synthetic MD data and experimental measurements at the residue, domain, and whole-molecule levels.

Overall, the data will be generated in-house and supplemented with publicly available archives, while a consortium approach drives cross-lab sharing of molecular dynamics datasets to cover diverse biomolecular classes and dynamics. Pharma and biotech partners, such as Genentech and AstraZeneca, contribute proprietary data under collaborative agreements. Initiatives like the AISB Network illustrate how academia and industry can work together toward a greater goal: leveraging secure, federated learning to combine private and public datasets, apply advanced AI algorithms on high-performance infrastructure, and accelerate molecule design while overcoming traditional IP barriers.

Existing collaborations similar to the AISB Network, such as Apheris, Diffuse Project, Anthropic, and OpenBind, demonstrate the feasibility of uniting diverse organizations to advance AI-driven drug discovery while creating a win-win for both academia and industry. Academics gain access to large, high-quality datasets generated in laboratories equipped with industry-scale instruments, enabling them to train AI models for tasks such as antibody structure prediction. In turn, pharmaceutical and biotech partners benefit from expert analysis and interpretation of these massive datasets. This mutually beneficial framework represents an exciting and practical model for collaborative innovation in AI-driven molecular design.

D Cost and scalability

The generation of multi-scale MD datasets and complementary experimental data is resource-intensive but can be made cost-effective using scalable computing, shared infrastructure, and consortium-based collaborations. Recent studies have shown that computationally derived flexible data from MD simulations provides mechanistic insights into functional dynamics and that biochemical properties extracted from MD closely align with wet-lab results, unlike rigid structural data. This allows a staged approach: large-scale MD simulations can form the first wave of data generation, and wet-lab experiments can be scaled up later depending on budget and priority targets (Table 1). In summary, the total estimated budget would be equal to $4-6M for initial dataset generation, scalable with additional computational and experimental resources.

Table 1: Estimated budget to create DINO, including both computational and experimental waves.

Component	Approach	Cost	Scalability
All-atom MD simulations	GROMACS [4], OpenMM [11], AMBER [8], PLUMED [7]	$1-2M	Can scale linearly with additional GPUs or cloud HPC, checkpointing enables multi-month simulations across hundreds of systems in parallel.
Coarse-grained simulations	CALVADOS [32], Martini [20], custom CG	$300-500K	Highly parallelizable, allows simulation of very large assemblies (membranes, condensates) with modest computing, easily integrated into high-throughput pipelines.
Quantum MD (QM/MM, Ab Initio)	GAMESS [3], Gaussian [13], ORCA [23]	$500K	Highly resource-consuming, scales to selected high-priority molecules, GPU-accelerated and cloud resources reduce runtime, batch calculations possible.
AI-generated ensembles	AlphaFlow [16], BioEmu [19], AI2BMD [33], emerging techniques	$200-300K	Computationally cheap after training, scales with additional training data, generative models enable rapid augmentation without additional wet-lab or MD costs.
Experimental dynamics	HDX-MS, Cryo-EM, NMR, smFRET	$2-3M	Scalable via multi-lab consortium, focus on representative high-value targets, automation reduces per-sample cost, federated protocols enable distributed data collection.
Data management and storage	Shared repositories, federated learning	$100-200K	Scales with cloud storage, standardized pipelines allow easy integration of new MD and experimental datasets, cross-lab access reduces duplication of effort.

Cost reduction strategies include leveraging consortium collaborations to share instrumentation and labor, using coarse-grained and ML-augmented simulations to reduce high-cost all-atom or quantum

calculations, and automating experimental workflows. Importantly, the strong correlation between MD-derived functional insights and experimental data enables a phased approach, where large-scale computational datasets are generated first, and experimental validation is added later as budget allows.

E Conclusion

Computational drug discovery is limited by the reliance on static molecular structures, leading AI models to suffer from state bias, dynamics blind spots, and unrealistic conformational sampling. As a result, generated candidates often lack the biological plausibility necessary for a molecule to be a drug and require exhaustive and costly experimental validation. To address this, we propose DINO, a dataset designed to embed molecular motion into AI-driven design. DINO integrates experimental and synthetic molecular dynamics data across membrane proteins, antibodies, nucleic acids, small molecules, and complexes, spanning atom- to system-level motions. By capturing biophysical conformational ensembles, binding energetics, and functional kinetics, the dataset provides biologically meaningful representations of molecular flexibility and function. Such data will support prediction of binding affinity, specificity, stability, and disorder, generation of flexible biomolecules and realistic small molecule binders, and hybrid tasks such as integrating static structural models into dynamic landscapes. By grounding molecular design in thermodynamic principles, DINO enables AI models to move beyond static assumptions and generate biochemically plausible candidates with higher therapeutic potential.

References

[1] Rommie Amaro, Johan Aqvist, Ivet Bahar, Federica Battistini, Adam Bellaiche, Daniel Beltran, Philip Biggin, Massimiliano Bonomi, Gregory Bowman, Richard Bryce, Giovanni Bussi, Paolo Carloni, David Case, Andrea Cavalli, Chia-En Chang, Thomas Cheatham, Margaret Cheung, Christophe Chipot, Lillian Chong, and Modesto Orozco. The need to implement fair principles in biomolecular simulations. *Nature methods*, 22, 04 2025. doi: 10.1038/s41592-025-02635-0.

[2] Catherine Baranowski, Hector Martin, Diego Oyarzún, Aviv Spinner, Bijoy Desai, Christopher Petzold, Evangelos-Marios Nikolados, Sebastian Jaaks-Kraatz, Aljaž Gaber, Robert Chalkley, Devin Scannell, Rachel Sevey, Michael Jewett, Peter Kelly, and Erika DeBenedictis. Can protein expression be 'solved'? *Trends in Biotechnology*, 06 2025. doi: 10.1016/j.tibtech.2025.04.021.

[3] Giuseppe M. J. Barca, Colleen Bertoni, Laura Carrington, Dipayan Datta, Nuwan De Silva, J. Emiliano Deustua, Dmitri G. Fedorov, Jeffrey R. Gour, Anastasia O. Gunina, Emilie Guidez, Taylor Harville, Stephan Irle, Joe Ivanic, Karol Kowalski, Sarom S. Leang, Hui Li, Wei Li, Jesse J. Lutz, Ilias Magoulas, Joani Mato, Vladimir Mironov, Hiroya Nakata, Buu Q. Pham, Piotr Piecuch, David Poole, Spencer R. Pruitt, Alistair P. Rendell, Luke B. Roskop, Klaus Ruedenberg, Tosaporn Sattasathuchana, Michael W. Schmidt, Jun Shen, Lyudmila Slipchenko, Masha Sosonkina, Vaibhav Sundriyal, Ananta Tiwari, Jorge L. Galvez Vallejo, Bryce Westheimer, Marta Wloch, Peng Xu, Federico Zahariev, and Mark S. Gordon. Recent developments in the general atomic and molecular electronic structure system. *The Journal of Chemical Physics*, 152(15):154102, April 2020. ISSN 0021-9606, 1089-7690. doi: 10.1063/5.0005188. URL `http://aip.scitation.org/doi/10.1063/5.0005188`.

[4] Henk Bekker, Herman Berendsen, E.J. Dijkstra, S. Achterop, Rudi Drunen, David van der Spoel, A. Sijbers, H. Keegstra, B. Reitsma, and M.K.R. Renardus. Gromacs: A parallel computer for molecular dynamics simulations. *Physics Computing*, 92:252–256, 01 1993.

[5] Andreas Bender and Isidro Cortés-Ciriano. Artificial intelligence in drug discovery: what is realistic, what are illusions? part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today*, 26(2):511–524, 2021. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2020.12.009. URL `https://www.sciencedirect.com/science/article/pii/S1359644620305274`.

[6] Helen Berman. The protein data bank. *Nucleic Acids Research*, 28:235–242, 01 2000. doi: 10.1093/nar/28.1.235.

[7] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth Tribello, Fabrizio Marinelli, Riccardo Capelli, and Jakub Rydzewski. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods*, 16:670, 07 2019. doi: 10.1038/s41592-019-0506-8.

[8] David Case, Thomas Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26:1668–88, 12 2005. doi: 10.1002/jcc.20290.

[9] Uddalak Das. Generative ai for drug discovery and protein design: the next frontier in ai-driven molecular science. *Medicine in Drug Discovery*, 27:100213, 2025. ISSN 2590-0986. doi: https://doi.org/10.1016/j.medidd.2025.100213. URL https://www.sciencedirect.com/science/article/pii/S2590098625000107.

[10] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte Deane. Sabdab: The structural antibody database. *Nucleic acids research*, 42, 11 2013. doi: 10.1093/nar/gkt1043.

[11] Peter Eastman, John Chodera, Robert McGibbon, Yutong Zhao, Kyle Beauchamp, Lee-Ping Wang, Andrew Simmonett, Matthew Harrigan, Chaya Stern, Rafal Wiewiora, Bernard Brooks, and Vijay Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13:e1005659, 07 2017. doi: 10.1371/journal.pcbi.1005659.

[12] Yago Ferreira e Silva, Harold Hilarion Fokoue, and Paulo Ricardo Batista. Exploring the intrinsic structural plasticity and conformational dynamics of human beta coronavirus spike glycoproteins. *Journal of Chemical Information and Modeling*, 65(14):7712–7733, 2025. doi: 10.1021/acs.jcim.5c00990. URL https://doi.org/10.1021/acs.jcim.5c00990. PMID: 40673918.

[13] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian~16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

[14] Noah Herrington, Yan Chak Li, David Stein, Gaurav Pandey, and Avner Schlessinger. A comprehensive exploration of the druggable conformational space of protein kinases using ai-predicted structures. *PLOS Computational Biology*, 20, 07 2024. doi: 10.1371/journal.pcbi.1012302.

[15] John Irwin, Khanh Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin Wong, Munkhzul Khurelbaatar, Yurii Moroz, John Mayfield, and Roger Sayle. Zinc20-a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling*, XXXX, 10 2020. doi: 10.1021/acs.jcim.0c00675.

[16] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles, 2024. URL https://arxiv.org/abs/2402.04845.

[17] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 07 2021. doi: 10.1038/s41586-021-03819-2.

[18] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Qingliang Li, Benjamin Shoemaker, Paul Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan Bolton. Pubchem 2025 update. *Nucleic acids research*, 53, 11 2024. doi: 10.1093/nar/gkae1059.

[19] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Foong, Victor Satorras, Osama Abdin, Bastiaan Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389, 07 2025. doi: 10.1126/science.adv9817.

[20] Siewert Marrink, Herre Risselada, Sergey Yefimov, D Tieleman, and Alex Vries. The martini force field: Coarse grained model for biomolecular simulations. *The journal of physical chemistry. B*, 111:7812–24, 08 2007. doi: 10.1021/jp071097f.

[21] Tim Meyer, Marco D'Abramo, Adam Hospital, Manuel Rueda, Carles Ferrer-Costa, Alberto Pérez, Oliver Carrillo, Jordi Camps, Carles Fenollosa, Dmitry Repchevsky, Josep Gelpí, and Modesto Orozco. Model (molecular dynamics extended library): A database of atomistic molecular dynamics trajectories. *Structure (London, England : 1993)*, 18:1399–409, 11 2010. doi: 10.1016/j.str.2010.07.013.

[22] Antonio Mirarchi, Toni Giorgino, and Gianni Fabritiis. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11, 11 2024. doi: 10.1038/s41597-024-04140-z.

[23] F. Neese. The orca program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(1):73–78, 2012. doi: 10.1002/wcms.81.

[24] K Peng, Z Obradovic, and S Vucetic. Exploring bias in the protein data bank using contrast classifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 435–46, 02 2004. doi: 10.1142/9789812704856_0041.

[25] Jakob R. Riccabona, Fabian C. Spoendlin, Anna-Lena M. Fischer, Johannes R. Loeffler, Patrick K. Quoika, Timothy P. Jenkins, James A. Ferguson, Eva Smorodina, Andreas H. Laustsen, Victor Greiff, Stefano Forli, Andrew B. Ward, Charlotte M. Deane, and Monica L. Fernández-Quintero. Assessing af2's ability to predict structural ensembles of proteins. 32(11):2147–2159.e2, 2024. ISSN 0969-2126. doi: 10.1016/j.str.2024.09.001. URL `https://doi.org/10.1016/j.str.2024.09.001`.

[26] Eva Smorodina, Oliver Crook, Johannes R. Loeffler, Monica Lisa Fernandez Quintero, Lucas Matthias Weissenborn, Hannah L Turner, Aleksandar Antanasijevic, Rahmad Akbar, Puneet Rawat, Khang Lê Quý, Brij Bhushan Mehta, Ole Magnus Fløgstad, Dario Segura-Peña, Nikolina Sekulić, Andrew B. Ward, Fridtjof Lund-Johansen, Jan Terje Andersen, and Victor Greiff. Structural modeling of antibody variant epitope specificity with complementary experimental and computational techniques. In *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025. URL `https://openreview.net/forum?id=d6ExTOu00A`.

[27] Kacper Szczepski and Lukasz Jaremko. Alphafold and what is next: bridging functional, systems and structural biology. *Expert Review of Proteomics*, 22, 01 2025. doi: 10.1080/14789450.2025.2456046.

[28] Elaine Tao and Ben Corry. Alphafold2 captures conformational transitions in the voltage-gated channel superfamily. *bioRxiv*, 2025. doi: 10.1101/2025.03.12.642934. URL `https://www.biorxiv.org/content/early/2025/03/14/2025.03.12.642934`.

[29] Chai Discovery Team, Jacques Boitreaud, Jack Dent, Danny Geisz, Matthew McPartlon, Joshua Meier, Zhuoran Qiao, Alex Rogozhnikov, Nathan Rollins, Paul Wollenhaupt, and Kevin Wu. Zero-shot antibody design in a 24-well plate. *bioRxiv*, 2025. doi: 10.1101/2025.07.05.663018. URL `https://www.biorxiv.org/content/early/2025/07/06/2025.07.05.663018`.

[30] Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, and Paul D. Adams. Alphafold predictions are valuable hypotheses and accelerate but do not

replace experimental structure determination. 21(1):110–116, 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02087-4. URL https://doi.org/10.1038/s41592-023-02087-4.

[31] Jack Turner, Sanja Abbott, Néli da Fonseca, Lucas Carrijo, Amudha Duraisamy, Osman Salih, Zhe Wang, Gerard Kleywegt, Kyle Morris, Ardan Patwardhan, Stephen Burley, Gregg Crichlow, Zukang Feng, Justin Flatt, Sutapa Ghosh, Brian Hudson, Catherine Lawson, Yuhe Liang, Ezra Peisach, and Xiaodan Ma. Emdb—the electron microscopy data bank. *Nucleic Acids Research*, 52, 11 2023. doi: 10.1093/nar/gkad1019.

[32] Sören von Bülow, Ikki Yasuda, Fan Cao, Thea K. Schulze, Anna Ida Trolle, Arriën Symon Rauh, Ramon Crehuet, Kresten Lindorff-Larsen, and Giulio Tesei. Software package for simulations using the coarse-grained calvados model, 2025. URL https://arxiv.org/abs/2504.10408.

[33] Tong Wang, Xinheng He, Mingyu Li, Yatao Li, Ran Bi, Yusong Wang, Chaoran Cheng, Xiangzhen Shen, Jiawei Meng, He Zhang, Haiguang Liu, Zun Wang, Shaoning Li, Bin Shao, and Tie-Yan Liu. Ab initio characterization of protein molecular dynamics with aibmd. *Nature*, 635:1019–1027, 11 2024. doi: 10.1038/s41586-024-08127-z.

[34] Stephanie Wankowicz. Modeling bias toward binding sites in pdb structural models, 12 2024.