Beyond Static Snapshots: A Large-Scale Dataset for Dynamics-Aware Protein–Nucleic Acid Modeling

Luiz Felipe Caparelli Piochi [1] **Omid Mokhtari** [1] **Hamed Khakzad** [1] **Yasaman Karami** [1]

Abstract

Protein–nucleic acid (NA) interactions govern essential biological processes, from transcription and translation to viral replication. Yet, AI models for protein-NA down-stream tasks are limited by the scarcity of high-quality dynamic data, as nearly all current datasets provide only static structures. To overcome this, we propose the first large-scale, open repository of molecular dynamics (MD) simulations for protein–NA complexes. It curates ∼1,000 high-resolution protein–RNA and protein–DNA complexes from PDB and generates 3 μs of dynamics per system (3×1 μs replicates), amounting to ∼3 ms of total simulation time. Each entry provides atomistic trajectories, post-processed features, and metadata, and all data will be integrated as a new MDDB node. This resource enables the development and systematic evaluation of dynamic-aware AI models, with binding site prediction as the primary target task, focusing on fast-timescale conformational variability and the associated local fluctuations, side-chain plasticity, and induced-fit rearrangements that static structures cannot represent. Crucially, while MD-trained generative conformational samplers have recently emerged for proteins, protein–NA complexes have lacked the large-scale atomistic MD training data needed to extend such approaches. This dataset fills that gap, opening the door to all-atom generative conformational samplers, protein–NA complex sampling, and nucleic-acid-targeting protein design.

1. AI Task Definition

Protein–nucleic acid (NA) interaction poses a cornerstone AI challenge in structural biology. Beyond fundamental interest, progress here would transform multiple application areas, from designing inhibitors of viral RNA (Dai et al., 2024; Leonard & Schaffer, 2005), to developing CRISPR-based tools with improved specificity (Ruffolo et al., 2025; Tycko et al., 2016). We focus on **binding site prediction** as the primary task, i.e., localizing residues and nucleotides involved in protein–NA interfaces. This task is particularly sensitive to conformational variability. Short-timescale induced-fit effects can modulate interaction patterns, even when the overall structure remains similar. Static structure-based models often fail in such settings, as they cannot represent the ensemble of conformations underlying binding. Recent methods such as AlphaFold3 (Abramson et al., 2024) and RoseTTAFoldNA (Baek et al., 2024) provide accurate static complex structures but do not explicitly model conformational dynamics. As a result, they may miss interaction-relevant states that emerge through local fluctuations or transient rearrangements.

The proposed dataset is designed to bring a shift from static to dynamic modeling. By providing large-scale MD trajectories of protein–NA complexes, it enables the development of **dynamic-aware graph-based deep learning models** that explicitly encode conformational ensembles. Beyond binding site prediction, the dataset supports three additional open challenges: **conformational ensemble generation**, **protein–NA complex sampling**, and **nucleic-acid-targeting protein design**, tasks that currently lack dedicated dynamic training data (see Appendix A for details).

2. Dataset Rationale

Conformational dynamics represent a critical bottleneck in AI for structural biology, as existing datasets overwhelmingly focus on static structures from sources like the Protein Data Bank (PDB), with over 200,000 entries but negligible dynamic information. This scarcity hampers AI's ability to model real biomolecular behavior, where flexibility drives function, for example in intrinsically disordered proteins or protein-NA interactions. While MD datasets for protein–protein and protein–ligand systems are increasingly available through community resources such as MDDB (Mokhtari et al., 2026a), protein–NA complexes remain underrepresented at this scale. This is despite RNA/DNA recognition being among the most dynamics-dependent in-

Accepted by ICML 2026 AI for Science Workshop. [1]Université de Lorraine, Inria, F-54000 Nancy, France. Correspondence to: Yasaman Karami <yasaman.karami@inria.fr>.

teraction classes, where single-stranded RNA flexibility, DNA breathing, and nucleobase stacking transitions directly modulate binding. This scarcity has a direct consequence: while MD-trained generative conformational samplers have recently emerged for proteins (Jing et al., 2024; Vander Meersche et al., 2024), protein–NA complexes lack a comparable atomistic MD training corpus, leaving this interaction class without the data needed to train or benchmark generative conformational models. Beyond enabling better AI models, MD simulations provide a physically rigorous reference for conformational dynamics that generative approaches can only approximate, thus making high-quality MD data both a training resource and an accuracy benchmark.

The proposed dataset is conceived as a natural extension of earlier single-chain and protein–protein complex MD efforts, filling the gap for protein–NA systems. We curate $\sim$1,000 high-resolution experimental structures of protein-RNA and protein-DNA complexes from PDB according to explicit structural and functional criteria, detailed in Appendix B. For each complex, we generate 3 μs of MD trajectories (3 replicates of 1 μs) using GROMACS with state-of-the-art nucleic acid force fields (AMBER99SB-ILDN with bsc0 and XOL3 corrections for protein–RNA, and ff14SB with bsc1 (Ivani et al., 2016) and CUFIX (Yoo & Aksimentiev, 2018) corrections for protein–DNA). The 1 μs timescale is deliberately chosen to capture local fluctuations, side-chain plasticity, and fast induced-fit rearrangements rather than rare global transitions. This choice is supported by prior protein–NA simulation studies. Microsecond trajectories of transcription-factor–DNA complexes show that most interface contacts fluctuate on sub-nanosecond timescales, so a 1 μs window samples each local contact mode $\sim 10^4$–10^5 times and resolves the time-dependent recognition patterns that distinguish binding sites (Etheve et al., 2016). Independently, internal B-DNA structure has been shown to converge on the 1–2 μs timescale, with intrahelical motions on the μs–ms range effectively absent in naked DNA (Galindo-Murillo et al., 2014). These two observations are complementary, as global helical structure converges within our window while the interface side-chain and contact dynamics that modulate binding-site occupancy are fully sampled within 1 μs. Per-system convergence (block-averaged RMSD/RMSF and replicate agreement) will be reported as a quality-control metric for every entry. Each entry provides atomistic trajectories, post-processed features such as RMSD, RMSF, PCA vectors, clustering, dynamical correlations, and displacement vectors, along with metadata (PDB/UniProt IDs, simulation parameters, DOIs). The dataset encompasses $\sim$3 ms total simulation time, $\sim$20 TB of raw data, atomic-level resolution at a 2 fs timestep, and labels including interaction sites and binding affinities. While force fields remain approximate, the resulting trajectories provide consistent, physically grounded representations of fast-timescale conformational dynamics, serving both as training data for AI models and as a physical reference for validating generative approaches.

3. Acceleration Potential

This dataset provides a foundation for incorporating conformational dynamics into AI model development, enabling dynamic-aware architectures that move beyond static representations. For protein–protein and protein-ligand interactions, recent studies have shown that integrating MD-derived features leads to measurable improvements in binding site prediction, affinity estimation, and mutational effect modeling (Mokhtari et al., 2026b; Guo et al., 2025; Min et al., 2024). By providing comparable large-scale dynamic data for protein–NA systems, our resource lays the foundation for similar advances. Concretely, dynamics-aware models could improve drug discovery (identifying transient binding pockets), protein design (generating binders that adapt to flexible RNA/DNA targets), and systems biology (modeling NA-binding proteins in transcriptional regulation). In medicine, the dataset will support transfer learning for understanding mutational effects in cancer-related transcription factors or for designing inhibitors against RNA viruses. A particularly high-impact application is the development of generative conformational samplers for protein–NA complexes. By making these data openly available under FAIR principles (Amaro et al., 2025) via MDDB integration (`https://mddbr.eu`), the dataset lowers barriers for the community to develop, test, and benchmark next-generation AI methods that explicitly account for flexibility as an essential but currently missing ingredient in protein–NA modeling.

4. Data-Creation Pathway

Data will be generated via high-performance computing on national resources ($\sim$270,000 GPU hours already allocated). An automated, fully reproducible pipeline has been developed and validated at scale (Mokhtari et al., 2026a), providing a concrete foundation for large-scale deployment. Starting from initial structures, the automated pipeline (Dockerized) handles preparation (solvation, ionization), simulation, and post-processing, ensuring reproducibility across systems. Trajectories will be hosted on an MDDB node, extending current simulations on single-chain protein and protein–protein complex MD datasets.

5. Cost and Scalability

We budget €150,000 total (primarily one research engineer over two years, plus storage and dissemination), which is modest relative to the scale of data produced. Scalability is

ensured by (i) integration into an existing federated MD data infrastructure, (ii) use of containerized, automated pipelines for reproducibility, and (iii) support from institutional storage and secure logging systems. These measures guarantee that the resource can grow beyond the initial release and remain sustainable in the long term.

Acknowledgements

YK was supported by the French National Research Agency (ANR) under the France 2030 grant reference number ANR-24-RRII-0002 operated by the Inria Quadrant Program. HK was supported by the ANR, under grants ANR-22-CPJ2-0075-01, and ANR-24-CE45-4243-01.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Amaro, R. E., Åqvist, J., Bahar, I., Battistini, F., Bellaiche, A., Beltran, D., Biggin, P. C., Bonomi, M., Bowman, G. R., Bryce, R. A., et al. The need to implement fair principles in biomolecular simulations. *Nature methods*, pp. 1–5, 2025.

Baek, M., McHugh, R., Anishchenko, I., Jiang, H., Baker, D., and DiMaio, F. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature methods*, 21 (1):117–121, 2024.

Dai, J., Jiang, X., da Silva-Júnior, E. F., Du, S., Liu, X., and Zhan, P. Recent advances in the molecular design and applications of viral rna-targeting antiviral modalities. *Drug Discovery Today*, 29(8):104074, 2024.

Etheve, L., Martin, J., and Lavery, R. Protein–DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors. *Nucleic Acids Research*, 44(20):9990–10002, 2016.

Galindo-Murillo, R., Roe, D. R., and Cheatham III, T. E. On the absence of intrahelical dna dynamics on the μs to ms timescale. *Nature communications*, 5(1):5152, 2014.

Guo, P., Correia, B., Vandergheynst, P., and Probst, D. Boosting protein graph representations through static-dynamic fusion. *bioRxiv*, pp. 2025–02, 2025.

Hu, S., Ma, Z., Liu, X., Hong, H., Liu, Y., Zhai, J., and Shen, H.-B. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Research*, 50(9):e51, 2022.

Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A., et al. Parmbsc1: a refined force field for dna simulations. *Nature methods*, 13(1):55–58, 2016.

Jing, B., Berger, B., and Jaakkola, T. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

Leonard, J. N. and Schaffer, D. V. Computational design of antiviral rna interference strategies that resist human immunodeficiency virus escape. *Journal of Virology*, 79 (3):1645–1654, 2005.

Min, Y., Wei, Y., Wang, P., Wang, X., Li, H., Wu, N., Bauer, S., Zheng, S., Shi, Y., Wang, Y., et al. From static to dynamic structures: Improving binding affinity prediction with graph-based deep learning. *Advanced Science*, 11 (40):2405404, 2024.

Mokhtari, O., Bignon, E., Khakzad, H., and Karami, Y. Dynarepo: the repository of macromolecular conformational dynamics. *Nucleic Acids Research*, 54(D1):D393–D401, 2026a.

Mokhtari, O., Grudinin, S., Karami, Y., and Khakzad, H. Dynamicgt: A dynamic-aware geometric transformer model to predict protein-binding interfaces in flexible and disordered regions. *Cell Systems*, 17(1):101454, January 2026b. ISSN 2405-4712.

Ruffolo, J. A., Nayfach, S., Gallagher, J., Bhatnagar, A., Beazer, J., Hussain, R., Russ, J., Yip, J., Hill, E., Pacesa, M., et al. Design of highly functional genome editors by modelling crispr–cas sequences. *Nature*, pp. 1–8, 2025.

Siebenmorgen, T., Menezes, F., Benassou, S., Merdivan, E., Didi, K., Mourão, A. S. D., Kitel, R., Liò, P., Kesselheim, S., Piraud, M., et al. Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature computational science*, 4(5):367–378, 2024.

Tycko, J., Myer, V. E., and Hsu, P. D. Methods for optimizing crispr-cas9 genome editing specificity. *Molecular cell*, 63(3):355–370, 2016.

Vander Meersche, Y., Creighton, A. T., Rihon, J., Henrard, A., and Nguyen, V.-A. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 2024.

Yoo, J. and Aksimentiev, A. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Physical Chemistry Chemical Physics*, 20(13):8432–8449, 2018.

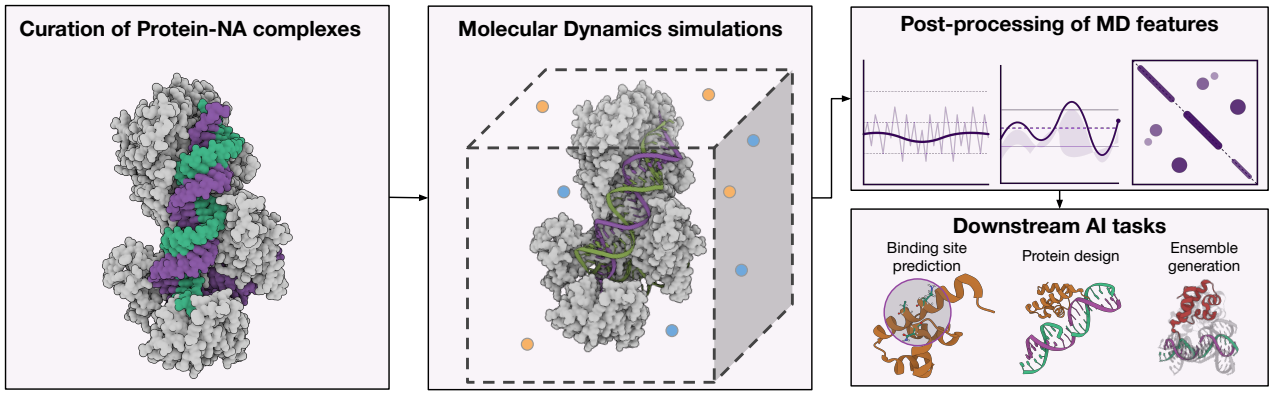

Figure 1. **Overview of the dataset generation and application pipeline.** The workflow begins with the curation of high-resolution protein–nucleic acid complexes. These structures are subjected to molecular dynamics (MD) simulations to capture fast-timescale conformational dynamics, local fluctuations, and induced-fit rearrangements. The resulting atomistic trajectories are post-processed to extract dynamic features (e.g., RMSD, RMSF, and dynamical correlations). Ultimately, this resource provides the foundation for downstream AI tasks, including dynamics-aware binding site prediction, nucleic-acid-targeting protein design, and conformational ensemble generation.

A. Extended AI Task Discussion

Binding site prediction can be rigorously evaluated against established benchmarks such as GraphBind (Hu et al., 2022).Concretely, we cast binding-site prediction as per-residue (and per-nucleotide) binary classification, where a residue is labeled positive if any heavy atom lies within 5 Å of the partner chain in the experimental complex. Models are trained on static- and MD-derived node features and evaluated with AUROC, AUPRC, MCC, and F1 against the GraphBind protein–NA test set, using its published sequence-identity-based train/test partition so that scores are directly comparable to prior work; the added value of dynamics is quantified as the score gain of MD-augmented over structure-only inputs on identical splits. Beyond this primary task, the dataset supports three additional open challenges in protein–NA modeling: **conformational ensemble generation** (sampling the distribution of bound-state configurations), **protein–NA complex sampling** (exploring interface dynamics and alternative binding modes), and **nucleic-acid-targeting protein design** (generating proteins that bind specific RNA/DNA targets with defined flexibility), all currently lacking dedicated dynamic training data. In detail, (i) *Ensemble generation*: given a single experimental structure, generate a conformational ensemble; evaluate by how well the predicted per-residue RMSF and the distribution of interface contacts match the held-out MD reference (e.g., RMSF correlation), (ii) *Complex sampling*: given unbound or perturbed components, recover near-native interface geometries, and evaluate with interface-RMSD against the native complex and against the MD-sampled basin. (iii) *NA-targeting design*: given a target RNA/DNA, generate binders and evaluate *in silico* with predicted interface recovery on held-out complexes.

Addressing these challenges requires models that explicitly encode conformational dynamics rather than relying on static structures. For protein–protein systems, incorporating MD-derived ensemble features already yields measurable improvements in binding site prediction (Mokhtari et al., 2026b), whereas for protein–ligand interactions, short simulations encoded as heterogeneous graphs with dynamic correlation edges improve both binding site and affinity prediction (Guo et al., 2025; Siebenmorgen et al., 2024). These gains are consistent with MD trajectories capturing substantially greater conformational diversity than flow-matching-based generative models. As quantified in Figure 2 and Table 1, AlphaFlow (Jing et al., 2024), trained on the ATLAS MD dataset (Vander Meersche et al., 2024), underestimates per-residue flexibility for 91% of matched proteins, confirming MD as the physically rigorous reference for conformational dynamics. Crucially, protein–NA complexes remain unserved at atomistic resolution in the fast-timescale bound-state regime, as coarse-grained approaches operate below all-atom detail, and no atomistic generative sampler trained on standard MD trajectories currently exists for these systems.

B. Extended Acceleration Potential

A particularly high-impact application is the development of **generative conformational samplers for protein–NA complexes**. AlphaFlow (Jing et al., 2024), trained on the ATLAS MD dataset (Vander Meersche et al., 2024) of single-chain

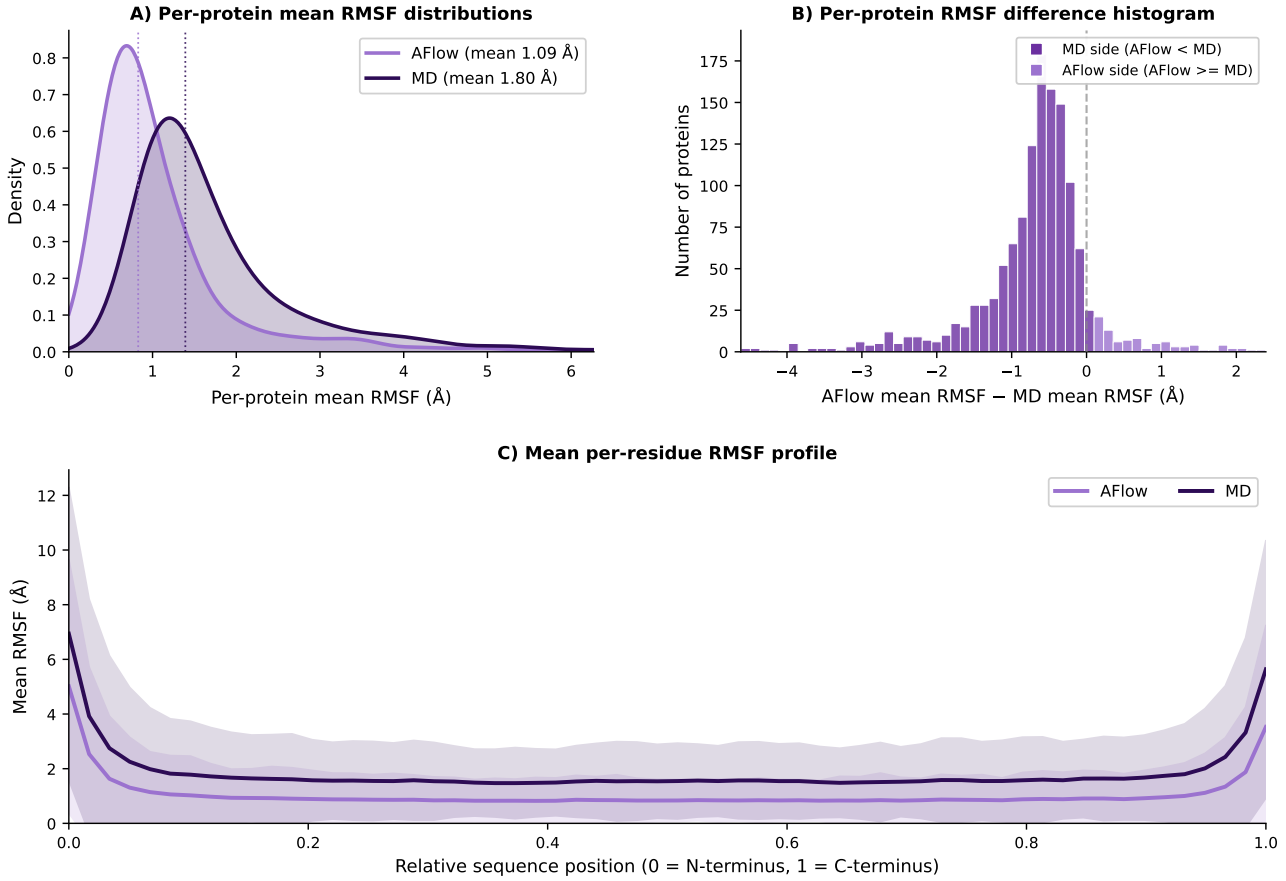

Figure 2. MD simulations systematically capture greater conformational diversity than the generative sampler AlphaFlow (Jing et al., 2024). **Left:** Per-protein mean RMSF from AlphaFlow-generated ensembles versus MD trajectories from the ATLAS, and MDDB datasets (Vander Meersche et al., 2024) for 1,329 matched proteins. Points below the identity line ($y = x$) indicate that MD reports higher RMSF; 91% of proteins fall in this regime. **Right:** Distribution of RMSF differences (AlphaFlow − MD); the median difference is -0.58 Å and mean -0.71 Å, confirming systematic underestimation of conformational flexibility by AlphaFlow. Together, these results underscore the value of MD trajectories as a physically rigorous training and validation resource, motivating the present dataset.

Table 1. Summary statistics of per-protein mean RMSF across all residues for each database category. MD aggregates MDDB, and ATLAS. IQR denotes the interquartile range [Q1–Q3].

Database	# Proteins	Median ± SD (Å)	IQR (Å)
AFlow	12,385	1.41 ± 2.08	[0.76–2.82]
MD	2,100	1.54 ± 1.71	[1.14–2.29]
NMR	1,388	1.13 ± 2.40	[0.71–2.09]

proteins, demonstrated that MD trajectories are the key ingredient for training generative models that produce physically accurate and diverse conformational ensembles, surpassing what can be learned from static structures alone. However, ATLAS and analogous resources cover only single-chain proteins, and no equivalent atomistic MD corpus exists for protein–NA complexes. Our dataset fills this gap, providing the first MD training corpus at the scale required to develop generative conformational samplers, complex sampling models, and protein design methods for nucleic-acid-binding proteins. Importantly, MD trajectories will serve a dual role as training data for generative models and as a physical ground truth against which generated ensembles can be validated.

Dataset Curation and Quality Control

Curation criteria. We select $\sim$1,000 PDB complexes based on: (i) protein–NA contact; (ii) resolution; (iii) valid biological assemblies; and (iv) non-redundancy. This size balances structural diversity with our compute.

Quality control. Trajectories pass an automated QC gate before release. Runs are excluded or repeated if they show instability complex dissociation (based on RMSD). Per-system QC diagnostics are provided as metadata.

Capturing binding-relevant conformational change. To validate that simulations capture binding-relevant dynamics, we provide interface-focused analyses on: per-residue RMSF, time-resolved contact maps, side-chain rotamer transitions, and conformational clustering. As preliminary corpus-level evidence, Figure 2 and Table 1 show MD systematically captures greater interface flexibility than a static-trained generative sampler, exposing fluctuations crucial for dynamic-aware binding predictions.

C. The Workflow in Details

This project contributes to the broader effort to generate high-quality scientific data representing the conformational dynamics and flexibility of biological macromolecules. Its key innovation lies in the large-scale simulation and analysis of macromolecular complexes, particularly protein–NA systems. The resulting data repository will serve as a critical resource for training deep learning models aiming to predict biological mechanisms by incorporating molecular flexibility in addition to static structural information.

This project highly depends on the digital resources and consists of two parts: I) Computing resources to carry out MD simulations. We plan to investigate a total of 1,000 macromolecular complexes by performing MD simulations consisting of 3 replicates of 1 μs for each system, leading to 3 μs simulations per complex. These simulations will be carried out on a national computing center. For an average system size of about 200,000 atoms, using one computing H100 node (with one GPU), we obtained a performance of 0.12 h/ns. Taking this performance into account, a total of about 360,000 GPU hours is required.

On average, each simulation requires about 62.5 hours of runtime. Therefore, we expect to carry out approximately 48 simulations per week. To ensure the continuous and efficient performance of the project, we will set up an automatic pipeline to clean the generated trajectories (e.g., by removing solvent atoms), transfer the files to our local computing node, and perform regular post-analysis. The purchase of a computing node with sufficient storage capacity is required to store and back up the data. This node can also be used for the post-processing stage of the project, where we will apply a range of classical analyses on the MD simulations, including RMSD, RMSF, PCA, radius of gyration, distance and interaction analysis, and clustering.