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,*]

[1] Université de Lorraine, Inria, F-54000 Nancy, France

* Correspondence to: *yasaman.karami@inria.fr*

1. Motivation & Gap

Understanding the conformational dynamics of proteins, RNA, and DNA is essential to decipher their function and interactions. While deep learning has transformed static structure prediction, modeling dynamics remains challenging due to limited data availability. Methods such as AlphaFold3 and RoseTTAFoldNA predict accurate complexes but cannot represent the local fluctuations, side chain plasticity, and induced fit rearrangements that shape binding.

Large molecular dynamics datasets now exist for single proteins (ATLAS) and in previous work we provided an open repository of large-scale molecular dynamics simulation of protein-protein systems (DynaRepo), but protein-nucleic acid complexes remain largely unserved at atomistic resolution. This is the most dynamics dependent interaction class in structural biology, where single stranded RNA flexibility, DNA breathing, and nucleobase stacking transitions directly modulate recognition.

Our dataset fills this gap.

Key numbers

~1,000 Protein-NA complexes

~360,000 GPU hours

Total MD simulation time of ~3 ms

2. The Dataset

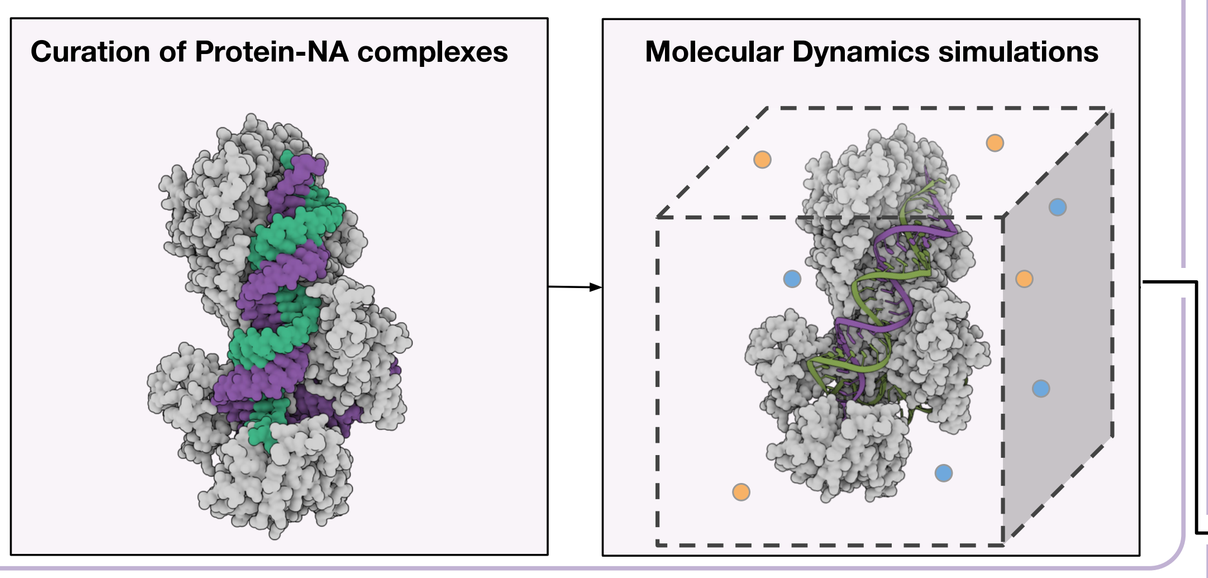
Around 1,000 high resolution protein RNA and protein DNA complexes curated from the PDB based on contact criteria, resolution, valid biological assemblies, and non redundancy.

Per system we generate 3 microseconds of atomistic molecular dynamics as 3 replicates of 1 microsecond, run with GROMACS at a 2 femtosecond timestep. Force fields are AMBER99SB ILDN with bsc0 and XOL3 corrections for protein RNA, and ff14SB with bsc1 and CUFIX corrections for protein DNA.

Criterion	Description
Protein–NA contact	Heavy atoms ≤ 5Å between protein and NA chains
Resolution	≤ 3.0Å ; no missing residues at interface
Biological assembly	No crystallographic artifacts
Non-redundancy	Proteins clustered at 40% sequence identity

The 1 microsecond window is chosen to capture fast timescale local dynamics rather than rare global transitions. Prior work shows interface contacts fluctuate on sub nanosecond timescales and B DNA structure converges within 1 to 2 microseconds, so each local contact mode is sampled 10,000 to 100,000 times per replicate.

Each entry provides atomistic trajectories, post processed features (RMSD, RMSF, PCA vectors, clustering, dynamical correlations, displacement vectors), and metadata (PDB and UniProt IDs, simulation parameters, DOIs). Total volume is around 3 milliseconds of simulation and 20 TB of data, openly released through an MDDB node under FAIR principles.



Why dynamics matters?

A static structure shows **one** conformation. Binding happens across **many**.

Protein-NA interfaces reshape on the nanosecond to microsecond timescale through side chain rotamer flips, backbone fluctuations, and induced fit rearrangements. These motions create **transient pockets**, **expose or bury interface residues**, and **modulate which contacts actually form**. Static structure models see none of this, whereas MD trajectories see all of it.

A residue that looks buried in the crystal may spend 40% of its time exposed in solution. That is the difference between a missed and a recovered binding site.

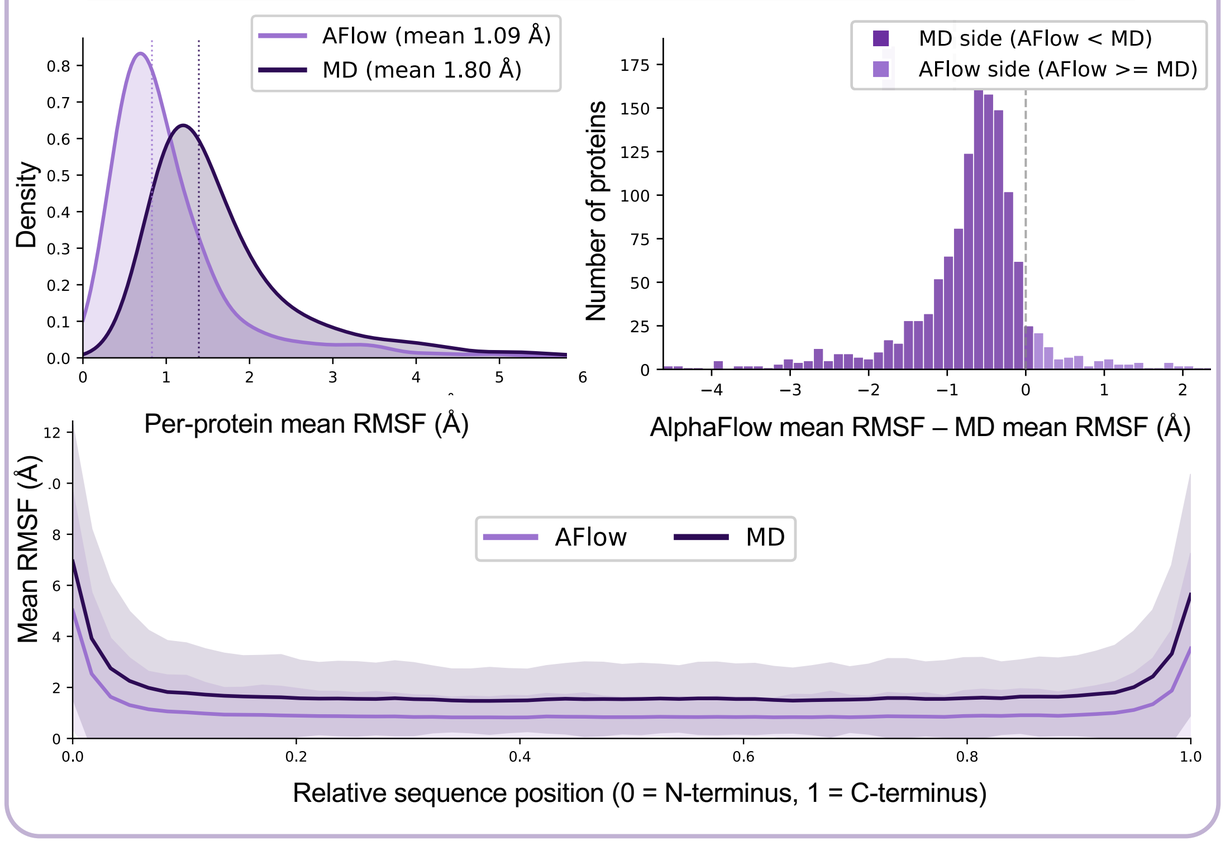
3. Pipeline & Validation

A fully automated, containerized pipeline handles curation, solvation, ionization, simulation, and post processing, ensuring reproducibility across all 1,000 systems. Around 270,000 GPU hours have already been allocated on national computing resources (GENCI).

Every trajectory passes an automated quality control gate before release. Runs are excluded or repeated if they show instability or complex dissociation, and per system diagnostics (block averaged RMSD and RMSF, replicate agreement) are released as metadata.

Preliminary evidence at corpus scale shows that molecular dynamics systematically captures greater conformational diversity than the static trained generative sampler AlphaFlow. Across 1,329 matched proteins from ATLAS, AlphaFlow underestimates per residue flexibility for 91% of cases, with a median RMSF gap of 0.58 angstroms. This confirms molecular dynamics as the physically rigorous reference for conformational dynamics and motivates the present resource.

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]



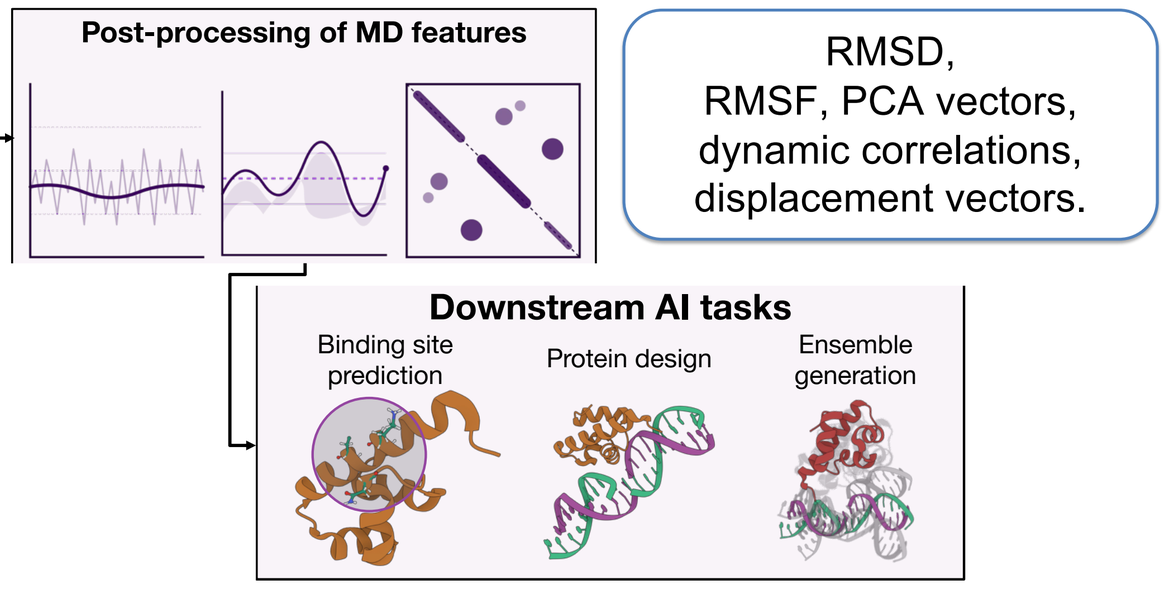
4. Downstream Tasks & Impact

Primary task. Binding site prediction cast as per residue and per nucleotide binary classification, a task that we previously reported to improve by leveraging MD data (DynamicGT), and evaluated against the GraphBind benchmark with AUROC, AUPRC, MCC, and F1 on its published sequence identity splits. The added value of dynamics is measured as the score gain of molecular dynamics augmented inputs over structure only inputs on identical splits.

Open challenges newly enabled.

- Conformational ensemble generation for bound state distributions
- Protein nucleic acid complex sampling and alternative binding modes
- Nucleic acid targeting protein design with defined flexibility

Broader impact. Identification of transient binding pockets for drug discovery, design of binders that adapt to flexible RNA and DNA targets, modeling of transcription factor dynamics in cancer, and design of inhibitors against RNA viruses and CRISPR systems with improved specificity.



Acknowledgments: This work was supported by the French Agence Nationale de la Recherche (ANR), under the France 2030 grant reference number ANR-24-RRII-0002 operated by the Inria Quadrant Program, and by the ANR under grants ANR-22-CPJ2-0075-01 and ANR-24-CE45-4243-01. This work was granted access to the HPC resources of IDRIS under the allocations 2023-A0150714660 and 2025-A0180714660 (granted to Y.K.) made by GENCI.

References:

Abramson et al. 2024 (AlphaFold3). Baek et al. 2024 (RoseTTAFoldNA). Jing et al. 2024 (AlphaFlow). Vander Meersche et al. 2024 (ATLAS). Hu et al. 2022 (GraphBind). Mokhtari et al. 2026a (DynaRepo, https://inria.mddbr.eu). Mokhtari et al. 2026b (DynamicGT).