RPDynaFlow: Generating Protein–RNA Conformational Ensembles with Flow Matching

Yuntao Li[1,*], Kexin Lu[1]

[1]Center for Interdisciplinary Studies, Westlake University, Hangzhou, China
*corresponding author

Abstract

Protein–RNA complexes function through dynamic conformational ensembles, yet most structure prediction models produce a single static structure. We present RPDynaFlow (RNA–Protein Dynamics Flow matching), which learns per-atom displacement fields over all solute heavy atoms by conditional flow matching and generates all-atom dynamic conformational ensembles directly from a single static structure, trained on only fifteen 40 ns MD trajectories. The tiered dynbench evaluation and independent DRPScore scoring show that the model captures the collective motion patterns of MD while expanding phase-space coverage without losing native-state accuracy. We also explore physics-supervised variants with force/energy matching and energy-guided sampling, yielding results that merit further attention.

1 Introduction

In protein–RNA complex structure prediction, conformational ensembles offer far richer phase-space coverage than a single static structure. Many RNA–protein interactions (such as conformational selection and induced fit [1]) in fact involve functionally relevant intermediate states, which endpoint-state prediction models do not address. Dynamic conformational ensemble sampling is therefore particularly important. Classical molecular dynamics (MD) simulation can sample a specific system, but at substantial computational and time cost; deep generative models such as flow matching and diffusion are becoming a new route to accelerate this process. We aim to use deep neural networks to extend phase-space coverage while maintaining reasonable accuracy, and to attempt sampling of possible local saddle-point states.

In recent years, deep generative models have begun moving from single static structure prediction toward conformational ensemble sampling. DiG [2] uses deep networks to transform a simple distribution toward the equilibrium distribution of molecular systems; DiffDock [3] formulates molecular docking as a diffusion generative process over the manifold of ligand poses; AlphaFlow [4] fine-tunes single-state structure prediction models under a flow-matching framework to sample protein conformational diversity; RoseTTAFoldNA [5] extends structure prediction to nucleic acids and protein–nucleic acid complexes; BioEmu [6] integrates over 200 ms of MD data to directly generate Boltzmann-distributed protein ensembles; PLACER [7] builds conformational ensembles of small molecules and protein side chains given the protein context; SimpleFold [8] shows that a flow-matching model built only from general-purpose Transformer modules can perform folding and ensemble prediction efficiently on consumer hardware. For protein–RNA complexes, the smaller dataset and the scale disparity between RNA and protein are the key bottlenecks: only about a thousand usable complex structures exist, with redundancy that can inflate scores through train–test leakage [9]; the scale disparity reduces the reference value of coarse-grained modeling, calling for all-atom modeling.

We present RPDynaFlow (RNA–Protein Dynamics Flow matching), an attempt at small-data generation of protein–RNA conformational ensembles under these constraints: an atomic representation, training on fifteen short 40 ns trajectories, the tiered dynbench evaluation framework combined with independent external scoring, and a systematic comparison of purely geometry-driven and physics-supervised variants.

2 Results

2.1 Protein–RNA complex dynamics modeling

We built RPDynaFlow, a generative model of dynamic conformational ensembles for protein–RNA complexes (Fig. 1). Conditioned on a single static experimental structure, it uses an atomic representation retaining all heavy atoms of protein, RNA, and ligands (6 element classes, 25 residue types). A SchNet-style graph neural network [10] (5 interaction blocks, ∼0.3M parameters) predicts per-atom velocity fields on the 8 Å cutoff graph of the static structure, and Euler integration directly generates all-atom conformational ensembles. The training objective is optimal-transport conditional flow matching [11, 12], with states defined as normalized displacements relative to the static structure.

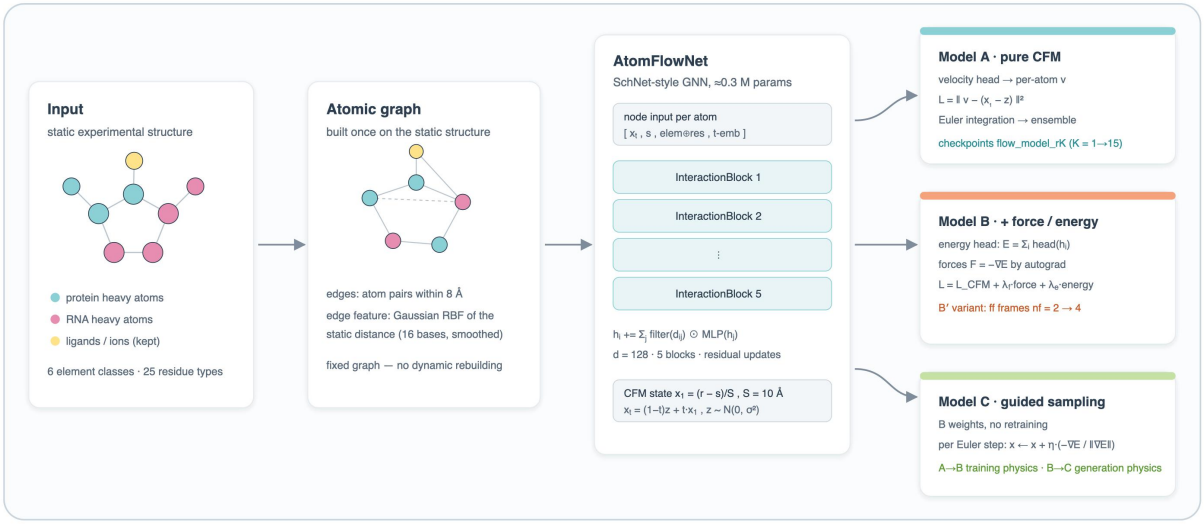

Figure 1: RPDynaFlow architecture. All solute heavy atoms of the static structure form an 8 Å cutoff atomic graph; five SchNet-style interaction blocks predict per-atom velocity fields. Model A is pure conditional flow matching; Model B adds an energy head and derives forces by automatic differentiation for joint force/energy supervision; Model C applies energy-guided sampling on top of B's weights.

We designed three mutually controlled variants: Model A is purely geometry-driven flow matching; Model B adds an energy head to A's backbone and computes forces by automatic differentiation for joint force and energy supervision; Model C applies energy-guided sampling at inference on top of B's weights. Unless otherwise stated, results below use K=15 (15 training systems).

2.2 The dynbench evaluation framework

We developed dynbench, a tiered evaluation framework for dynamic conformational ensembles, assessing consistency between generated and MD reference ensembles at three levels: (1) per-atom fluctuation (RMSF Pearson r) and dynamic cross-correlation (DCCM Pearson r); (2) collective motion modes (RMSIP of top-10 principal components) and inter-residue contact frequen-

cies; (3) phase-space coverage (5 Å Recall) and free-energy landscape error ($|\Delta G|$ MAE). Each metric is reported alongside two references: a Gaussian null model matched to MD fluctuation amplitudes, and MD trajectory half-split self-consistency (MD-self).

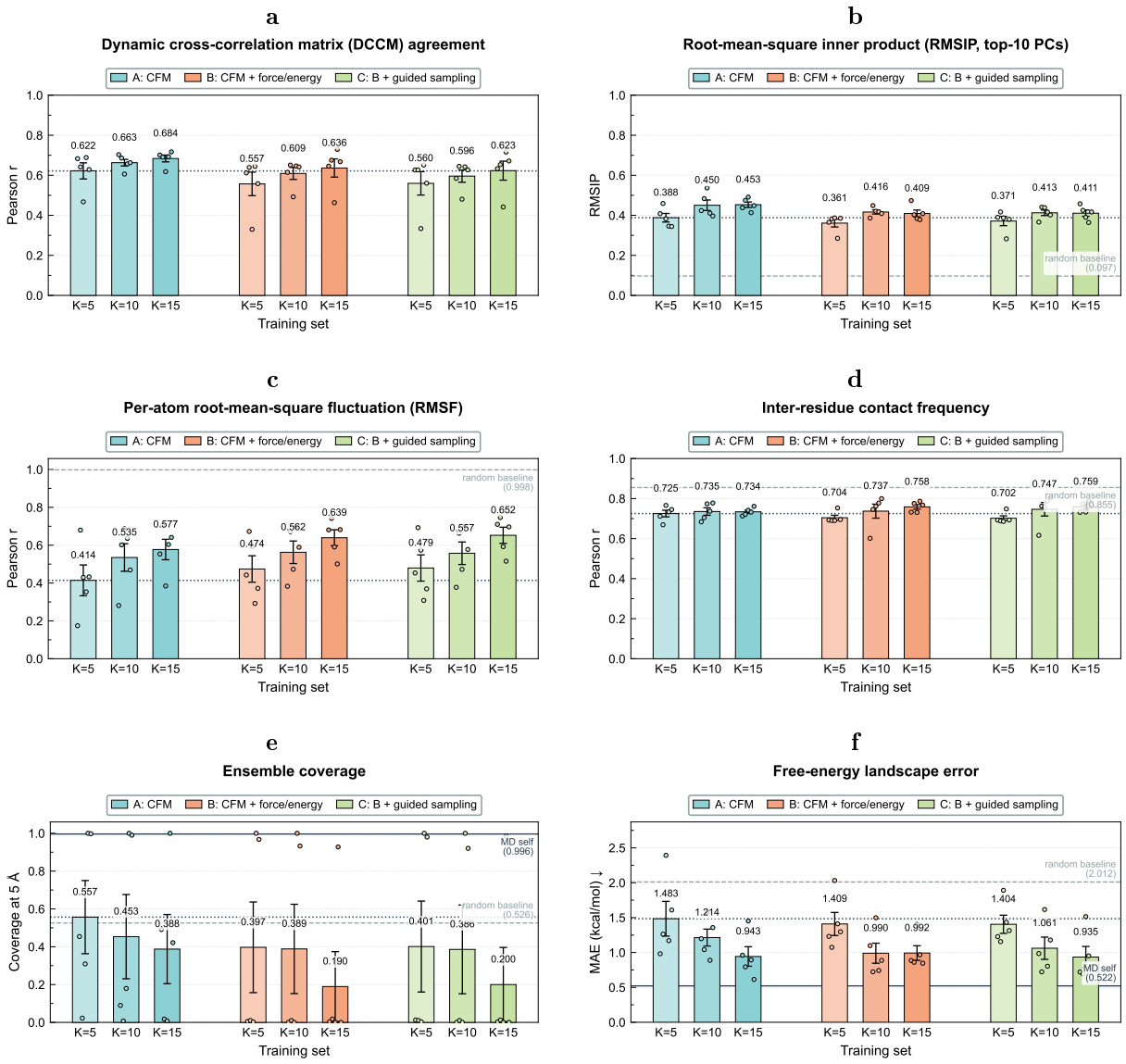

Figure 2: Six core dynbench metrics. The x-axis shows three model families (A: pure CFM; B: CFM + force/energy matching; C: B + guided sampling) with K=5/10/15 training systems within each family. Bar heights are five-system means, dots are per-system values, error bars are SEM; the grey dashed line is the random baseline (Gaussian null matched to MD fluctuation amplitudes), the dark solid line is MD-self. **a**, DCCM; **b**, RMSIP; **c**, RMSF; **d**, contact frequency; **e**, 5 Å coverage; **f**, free-energy landscape error (lower is better).

Figure 2 reports six core metrics across five held-out test systems. Model A clearly outperforms the null model on correlation-type metrics: DCCM 0.684 (null ≈0), RMSIP 0.453 (null 0.097), and free-energy landscape error 0.94 kcal/mol (null 1.98, MD-self 0.52), indicating that flow matching captures genuine collective motion patterns. Absolute gaps remain: 5 Å coverage of 0.388 falls below the MD-self value of 0.996; per-atom RMSF consistency of 0.577 does not reach the null model (which by construction matches MD fluctuation amplitudes). Reproducing fluctuation amplitudes is currently the weakest aspect.

Increasing training data from K=5 to 15 monotonically improves DCCM, RMSIP, RMSF,

and the free-energy landscape error. Models B and C differ from A but maintain comparable prediction scores; we speculate that the newly added parameters may require more training epochs to fit. A detailed comparison is given in the Discussion.

2.3 Phase-space coverage

Figure 3 projects the MD trajectories and the A, B, C (K=15) generated ensembles of the five test systems onto the first two principal components of their respective MD trajectories. Coverage differs markedly across systems: 6TPH is almost fully covered (Recall 1.0), 2HGH and 7K9D are partially covered (0.49, 0.42), while the generated ensembles of 2LBS and 4M4O remain contracted near the static structure (0.02, 0.01). B and C achieve near-zero coverage on 2HGH, 7K9D, and 4M4O; under the current setup we do not yet observe large changes in phase-space exploration from the physics-supervised variants.

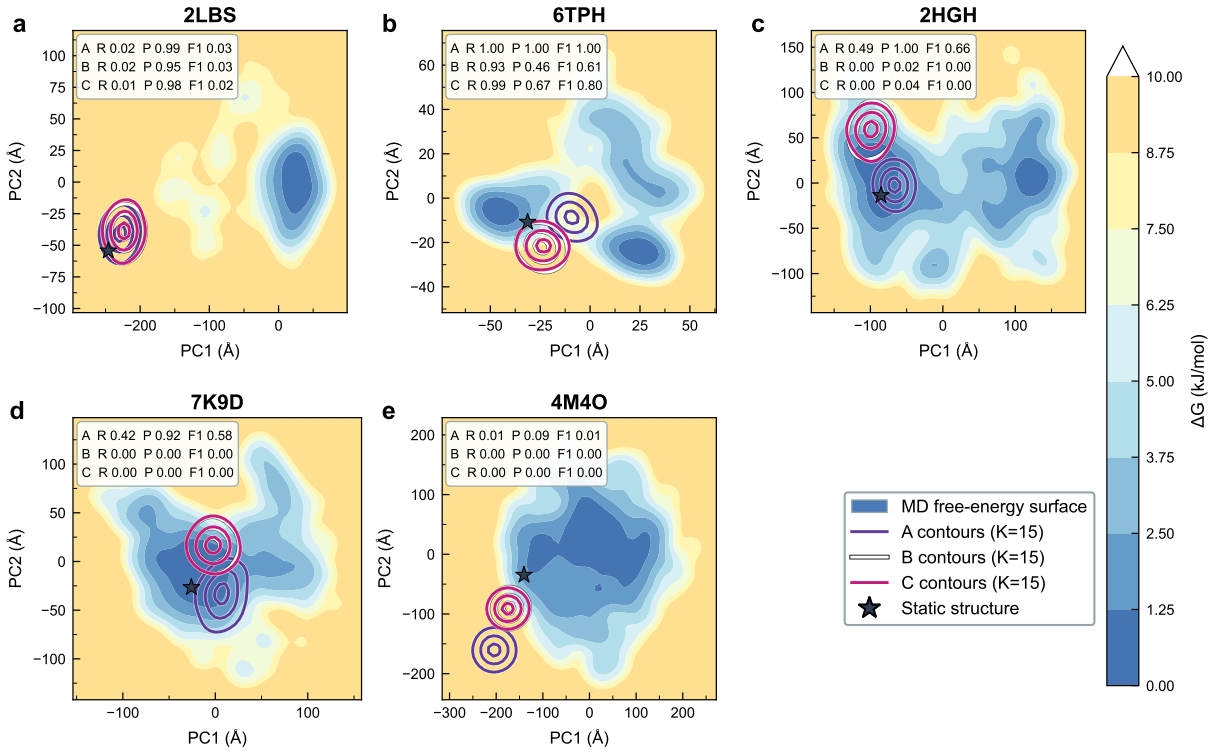

Figure 3: Principal component projections for five test systems. Background shows the MD free-energy surface ($\Delta G = -RT\ln(P/P_{\max})$, kJ/mol; blue = low-energy basins, orange = high-energy barriers). Violet, white, and magenta contours represent A, B, and C generated ensemble densities, respectively; stars mark the static conditioning structure. Upper-left annotations show 5 Å coverage Recall/Precision/F1. **a**, 2LBS; **b**, 6TPH; **c**, 2HGH; **d**, 7K9D; **e**, 4M4O.

2.4 DRPScore native-likeness scoring

Classical scoring functions assign higher scores to complexes closer to the correctly docked conformation; since our goal is a dynamic conformational ensemble, we necessarily obtain a spread of scores under such metrics. DRPScore scoring is our attempt to verify that the ensemble retains reasonable accuracy while sampling is expanded. Figure 4 shows the independent DRPScore [13] results. Mean P(native) of the generated ensembles lies between 0.83 and 0.96, comparable to or higher than the MD trajectory itself (0.829). The distribution shape shifts with training data size: at K=5, samples are tightly concentrated on the native state (mean 0.959, only 2% with P(native) < 0.5); at K=15, the mean relaxes to 0.826, almost identical to MD,

with 18% of samples below 0.5 (MD: 14%) — the distribution develops an MD-like deviation tail while preserving native-state accuracy.

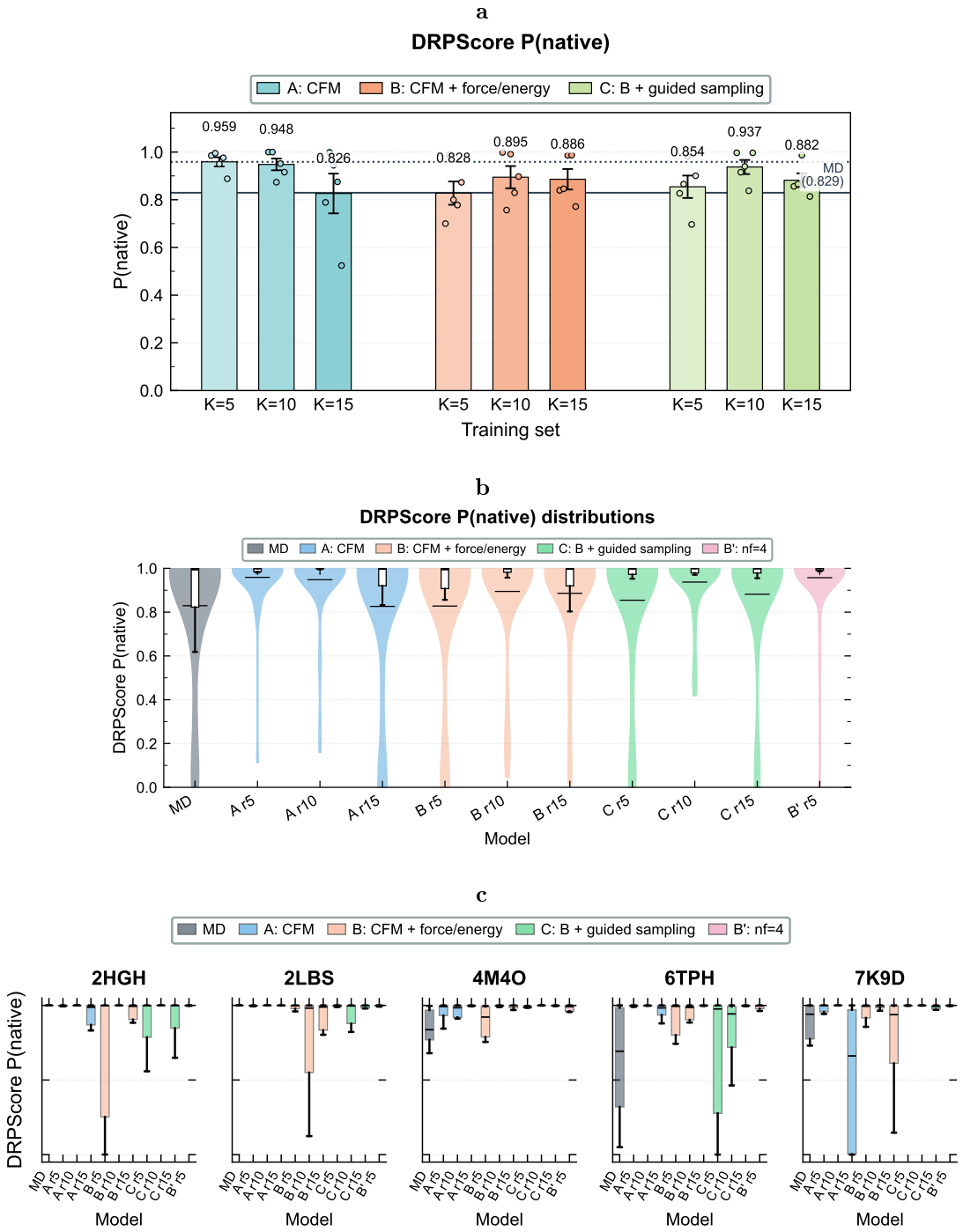

Figure 4: DRPScore external validation. **a**, Mean P(native) per model (layout as Fig. 2); the dark solid line is the MD reference (0.829). **b**, Violin distributions of all scores (including the B′ nf4 variant and MD). **c**, Score distributions per test system.

Table 1: The 20 RNA–protein complexes used in this study. Fluctuation is the NMR ensemble experimental fluctuation (Å); X-ray systems have no such value.

PDB	Method	Heavy atoms	NMR fluctuation	Role
1NYB	NMR	1144	5.55	initial training
2ESE	NMR	2100	2.14	initial training
1EKZ	NMR	2173	6.45	initial training
1A1T	NMR	1514	1.84	incremental training
4PDB	X-ray	1831	–	incremental training
1DK1	X-ray	1978	–	incremental training
2XDB	X-ray	2055	–	incremental training
2Y8W	X-ray	2098	–	incremental training
6GBM	NMR	2290	2.78	incremental training
2FY1	NMR	2343	14.67	incremental training
2N82	NMR	2385	3.40	incremental training
2L2K	NMR	2438	2.20	incremental training
2N3O	NMR	2621	3.95	incremental training
1RKJ	NMR	3403	4.89	incremental training
1FJE	NMR	3444	2.62	incremental training
7K9D	X-ray	2003	–	test
4M4O	X-ray	2273	–	test
2LBS	NMR	2434	1.66	test
6TPH	NMR	2647	1.79	test
2HGH	NMR	3163	0.99	test

3 Methods

3.1 Data selection and processing

We used 20 complexes from the Zhao-lab RNA–protein dataset [9] (pdb and fasta data of NMR and X-ray structures; Table 1). The three smallest complexes with NMR ensemble fluctuation ≥ 2 Å (1NYB, 2ESE, 1EKZ) form the initial training set; another 12 systems form the incremental training pool for the K=1→15 ablation; five systems (2LBS, 6TPH, 2HGH, 7K9D, 4M4O) are held out as the test set and never participate in training.

For each system, we ran 40 ns all-atom MD simulations with GROMACS [14] using the CHARMM36 force field [15] and the TIP3P water model [16]. After periodic-boundary correction and rigid-body superposition, all solute heavy-atom coordinates were extracted as training and evaluation data.

3.2 Leakage control

Protein–RNA complex data are scarce, and train–test similarity can easily inflate scores. We therefore applied part of the data-cleaning procedure described in the GEMS paper when preparing the structures, and computed pairwise protein and RNA sequence similarities between the training and test sets (Fig. 5): train–test similarity is fairly uniform, with no obvious leakage of similar structures into the test set. In addition, the five test systems never participate in any training round (Table 1).

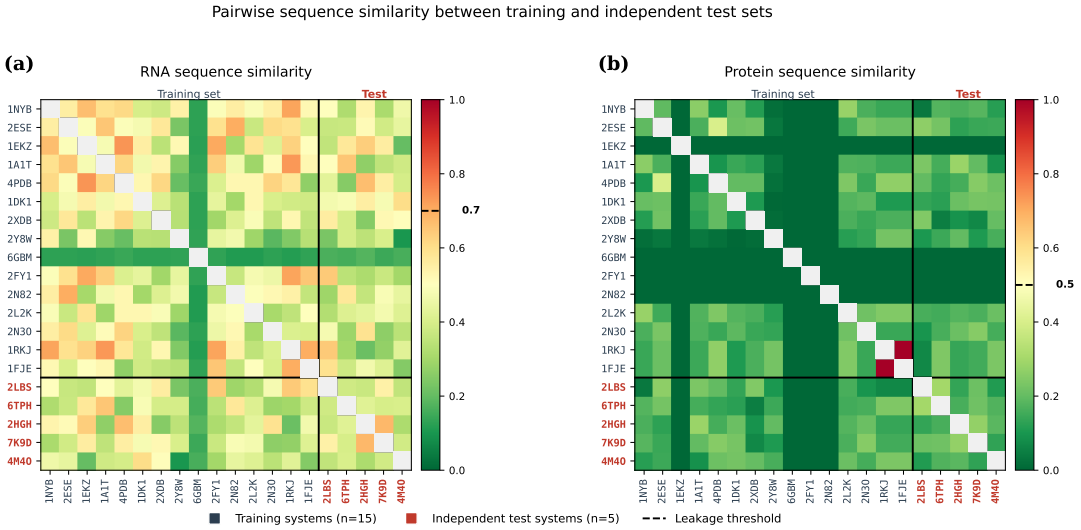

Figure 5: Pairwise sequence similarity heatmaps (protein and RNA) between the training and test sets, examining train–test leakage.

3.3 Model architecture and training

We adopt flow matching as the main framework, motivated by its light weight and fast inference demonstrated in small-scale models [8]. Given a static structure $\mathbf{s} \in \mathbb{R}^{A\times 3}$, the model learns a conditional flow from Gaussian noise

$$\mathbf{z} \sim \mathcal{N}(0, \sigma^2 I)$$

to the normalized displacement

$$\mathbf{x}_1 = (\mathbf{r} - \mathbf{s})/S$$

(S=10 Å). σ is the training-set displacement standard deviation, fixed from the first round so that rounds are comparable. The interpolation path is

$$\mathbf{x}_t = (1-t)\mathbf{z} + t\mathbf{x}_1,$$

the target velocity is $\mathbf{v}^* = \mathbf{x}_1 - \mathbf{z}$, and the loss is

$$\mathcal{L}_{\mathrm{CFM}} = \|\mathbf{v}_\theta(\mathbf{x}_t, \mathbf{s}, t) - \mathbf{v}^*\|^2.$$

As shown in Fig. 1, edges connect atom pairs within 8 Å on the static structure, with edge features given by a Gaussian RBF expansion of the static distance; the node input concatenates $[\mathbf{x}_t, \mathbf{s}, \mathrm{emb}_{\mathrm{elem}} \oplus \mathrm{emb}_{\mathrm{res}}, \mathrm{emb}_t]$; five SchNet interaction blocks (d=128) are followed by a linear head outputting per-atom velocities. Sampling proceeds by Euler integration of the velocity field. The model performs conformational ensemble inference rapidly on a single consumer GPU (e.g., RTX5090), generating candidate structures in batches; the data requirement is small — fifteen 40 ns short trajectories suffice for training.

Training samples one system per step (8 frames per step, single system) and runs 300 epochs at K=15. The incremental ablation K=5→15 appends systems one by one, continuing 150 epochs per round from the previous checkpoint.

3.4 DRPScore

DRPScore [13] is an independent deep-learning scoring function that outputs the probability of an RNA–protein complex structure being native. We scored 50 generated conformations per model per test system, and 50 evenly spaced MD frames as the reference.

4 Discussion

RPDynaFlow achieves small-data-driven prediction of concerted motions in RNA–protein complexes, with markedly faster sampling than classical MD. On the dynamics side, since the energy term remains almost unchanged, we do not yet observe large metric improvements from our current attempts with Models B and C: DCCM and coverage decrease slightly (B 0.636, C 0.623; coverage B 0.190, C 0.200), while RMSF and contact frequency increase marginally (B 0.639, C 0.652), and the free-energy landscape is unchanged; at the current training scale, force and energy supervision do not yet lead to large changes in overall ensemble quality. Increasing the number of force/energy-supervised frames per batch nf to 4 (the B′ variant) raises coverage from 0.397 to 0.470 and DRPScore from 0.828 to 0.957 (Fig. 6), but DCCM and RMSF decrease slightly and the energy loss still tends to zero, so we remain cautious about this interpretation.

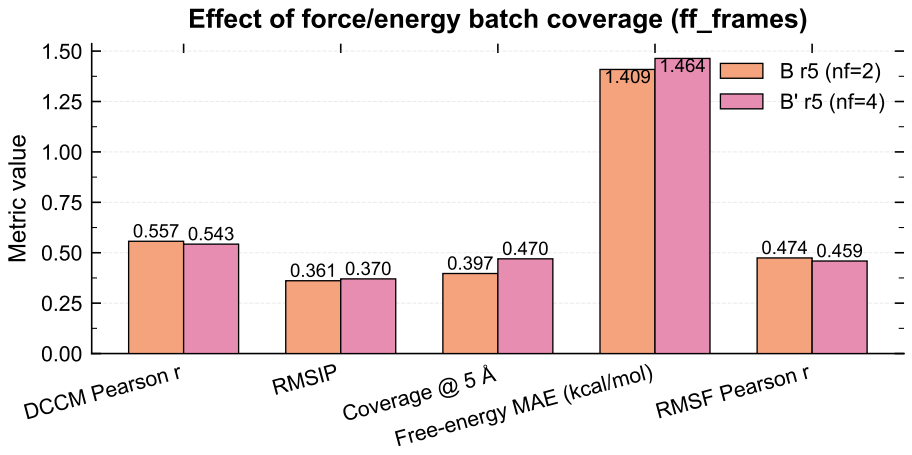

Figure 6: B′ control: effect of increasing nf from 2 to 4 (K=5). Coverage and DRPScore improve at nf = 4, while DCCM and RMSF decrease slightly.

Regarding scoring, higher classical scores indicate complexes closer to the correctly docked conformation, whereas our goal of a dynamic conformational ensemble necessarily yields a spread of such scores; the DRPScore results show that the model preserves MD-comparable native-state accuracy while expanding sampling. We currently regard phase-space coverage as a relatively novel observation metric. Several limitations remain: training trajectories are only 40 ns, so motions slower than this timescale may not be covered; the model is conditioned on a single static structure and uses no sequence or homology information; training data remain limited to RNA–protein complexes, and compatibility with other small-molecule systems awaits verification. Longer trajectories, a larger training pool, sequence conditioning, and more thorough convergence of physical constraints are the next directions.

A Supplementary material

Acknowledgements

This work is developed during the 2026 PEBBLE BioFusion Workshop at Westlake University in Hangzhou, China, from July 24th to August 4th.

References

[1] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007. doi: 10.1038/nature06522. URL https://www.nature.com/articles/

nature06522.

[2] Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, et al. Towards predicting equilibrium distributions for molecular systems with deep learning. *arXiv preprint arXiv:2306.05445*, 2023. URL https://arxiv.org/abs/2306.05445.

[3] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=kKF8_K-mBbS.

[4] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22277–22303. PMLR, 2024. URL https://proceedings.mlr.press/v235/jing24a.html.

[5] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, et al. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, 21:117–121, 2024. doi: 10.1038/s41592-023-02086-5. URL https://www.nature.com/articles/s41592-023-02086-5.

[6] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor Garcia Satorras, Osama Abdin, Dmitry Veprintsev, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761): eadv9817, 2025. doi: 10.1126/science.adv9817. URL https://www.science.org/doi/10.1126/science.adv9817.

[7] Ivan Anishchenko, Yakov Kipnis, Indrek Kalvet, Guangfeng Zhou, Rohith Krishna, Samuel J. Pellock, Anna Lauko, Gyu Rie Lee, Linna An, Justas Dauparas, Frank DiMaio, and David Baker. Modeling protein–small molecule conformational ensembles with PLACER. *Proceedings of the National Academy of Sciences*, 122(45):e2427161122, 2025. doi: 10.1073/pnas.2427161122. URL https://www.pnas.org/doi/10.1073/pnas.2427161122.

[8] Yuyang Wang, Jiarui Lu, Navdeep Jaitly, Josh Susskind, and Miguel Angel Bautista. SimpleFold: Folding proteins is simpler than you think. *arXiv preprint arXiv:2509.18480*, 2025. URL https://arxiv.org/abs/2509.18480.

[9] Rui Yang, Haoquan Liu, Liu Yang, Ting Zhou, Xinyao Li, and Yunjie Zhao. RPpocket: An RNA–protein intuitive database with RNA pocket topology resources. *International Journal of Molecular Sciences*, 23(13):6903, 2022. doi: 10.3390/ijms23136903. URL https://www.mdpi.com/1422-0067/23/13/6903.

[10] Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779. URL https://pubs.aip.org/aip/jcp/article/148/24/241722/962591.

[11] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[12] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2307.03672.

[13] Yanyang Yan and Sheng-You Huang. A deep-learning-based method for scoring protein–RNA docking models. *Nature Communications*, 14:5862, 2023. doi: 10.1038/s41467-023-41459-w. URL https://www.nature.com/articles/s41467-023-41459-w.

[14] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015. doi: 10.1016/j.softx.2015.06.001. URL https://www.sciencedirect.com/science/article/pii/S2352711015000059.

[15] Jing Huang and Alexander D. MacKerell. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34: 2135–2145, 2013. doi: 10.1002/jcc.23354. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23354.

[16] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926–935, 1983. doi: 10.1063/1.445869. URL https://pubs.aip.org/aip/jcp/article/79/2/926/776316.