RIBOMOTION: AN ATOMISTIC MOLECULAR DYNAMICS DATASET FOR RNA MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding RNA dynamics is essential for connecting structure to function, yet existing RNA resources primarily capture sequences and static structures. We introduce RiboMotion, an explicit-solvent, all-atom molecular dynamics (MD) dataset comprising nearly 100 μs of trajectories across 1,039 systems, including isolated RNAs and RNA–ligand complexes. We derive six benchmark tasks covering flexibility, geometry, and stability through nucleotide and pocket fluctuations, distance distributions, and persistent disruption of stacking and hydrogen-bond interactions. Benchmarking shows that trajectory-derived observables contain learnable information beyond initial geometry. As a demonstration of broader utility, multi-frame MD supervision improves a pretrained RNA conformation generator, reducing centered distance-distribution error by 25.4% and flexibility-amplitude error by 35.9% relative to the original model. These results suggest that MD-supervised generators could offer an alternative to direct molecular dynamics simulations for sampling RNA conformational ensembles. All data and benchmarks will be freely available to accelerate the development of AI models that capture RNA dynamics.

1 INTRODUCTION

RNA molecules mediate gene expression, molecular recognition, and catalysis, making them central to cellular regulation (Ganser et al., 2019). A defining feature of RNA is its conformational flexibility. Rearrangements of base pairing and stacking alter inter-residue distances and the accessibility of functional sites. These changes enable transitions that control ligand recognition, regulatory switching, and catalytic activity (Dethoff et al., 2012; Ganser et al., 2019). Understanding RNA function therefore requires describing how its structure fluctuates and reorganizes. Recent AI models have made substantial progress using sequence and static structural information, including secondary- and tertiary-structure prediction, protein–RNA binding preference prediction, and expression-regulation modeling (Chen et al., 2022; Shen et al., 2024). Predicting dynamic properties remains comparatively underexplored. Emerging methods address RNA flexibility and conformational generation (Feng et al., 2026; Li et al., 2025), but learning how fluctuations, geometric distributions, and interaction stability vary across RNA systems remains an open challenge.

Progress in AI-driven RNA dynamics is limited by two resource gaps. First, large-scale, consistently curated collections of RNA trajectories remain scarce. Existing RNA resources primarily support static structure prediction and sequence-based functional modeling, offering limited trajectory-level supervision for molecular motion (Szikszai et al., 2024; Ren et al., 2024; Arora et al., 2025). Second, shared benchmarks spanning complementary dynamic observables remain limited. Comparing models requires clearly defined prediction targets and metrics, together with evaluation partitions that account for molecular relatedness and RNA length. Protein MD datasets have already shown the value of trajectory supervision for flexibility prediction and conformational generation, motivating comparable resources for RNA (Siebenmorgen et al., 2024; Jing et al., 2024).

To address these gaps, we introduce RiboMotion, a curated collection of nearly 100 μs of atomistic MD trajectories spanning 730 isolated RNAs and 309 RNA–ligand complexes. From these trajectories, we derive six benchmark tasks covering flexibility, geometry, and interaction stability, with shared target definitions and evaluation metrics. We keep related molecular constructs within the same partition and balance RNA length distributions across training, validation, and test sets to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

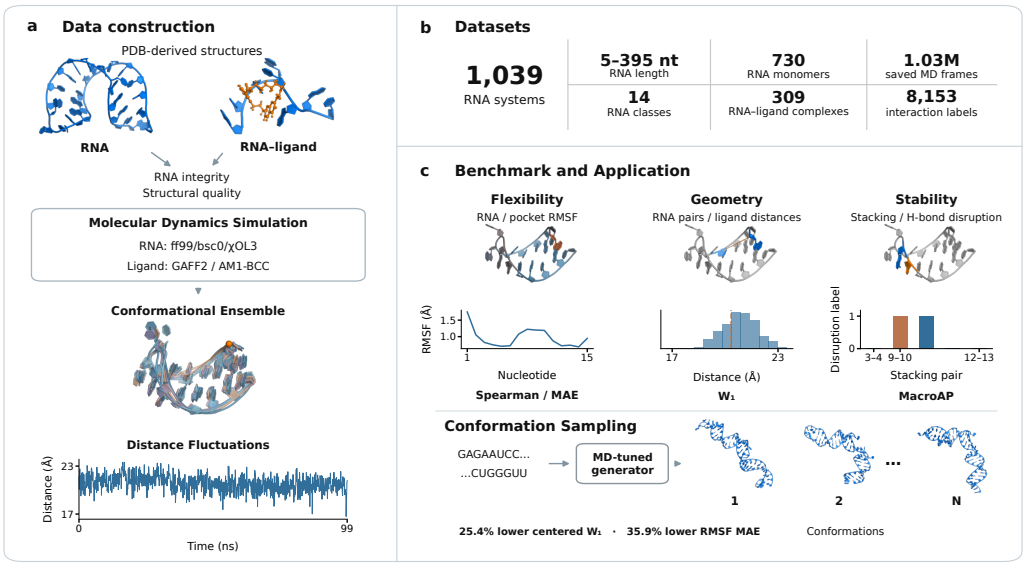

Figure 1: **RiboMotion: data construction, coverage, and applications.** (a) PDB-derived structures undergo RNA-integrity and structural-quality screening before MD. PyMOL overlays six uniformly spaced 1A4T frames; orange C4′ atoms (residues 1 and 5) define the distance trace. (b) The 1,039 systems (730 RNAs, 309 complexes; 5–395 nt) provide 1,032,430 conformations. The 14 annotation classes include a structural/design category. Across all partitions, the two interaction tasks contain 8,153 initial candidates: 7,172 stacks and 981 hydrogen bonds. (c) Six prediction tasks span flexibility, geometry, and interaction stability. Sequence-only conformation sampling is illustrated by three PyMOL-rendered samples of 68-nt test RNA 8SCH from MD-tuned Protenix. Gains over original Protenix summarize all 97 test RNAs (Table 4).

assess generalization across systems. Beyond observable prediction, the trajectories provide multi-frame supervision for adapting and evaluating RNA conformation generators.

The benchmark reveals different strengths of geometric features and neural models across tasks. Explicit geometry is competitive for flexibility rankings and interaction events; neural models improve RNA distance distributions; and controlled experiments show that ligand-distance prediction benefits from representing changes around the initial structure. These differences make strong structural references central to interpreting progress. The trajectories also provide multiple conformations for adapting a pretrained generator. On 97 test RNAs, this adaptation improves both the shape of distance distributions and the amplitude of conformational fluctuations (Figure 1).

We make four contributions:

- **Dataset.** Atomistic MD trajectories spanning 1,039 RNA systems and over one million frames from isolated RNAs and RNA–ligand complexes.
- **Benchmark.** Six tasks in three categories: flexibility, geometry, and stability.
- **Evaluation.** Grouped splits and structural baselines, complemented by temporal and replica analyses to assess dynamic information and sampling consistency.
- **Utility.** Multi-frame trajectory supervision that improves a pretrained RNA conformation generator beyond matched static-structure and single-frame adaptation.

2 Related Work

RNA data resources. RNA3DB provides structurally dissimilar partitions for structure prediction, while BEACON and RNAGym organize structural and functional prediction tasks (Szikszai

et al., 2024; Ren et al., 2024; Arora et al., 2025). CoDNaS-RNA collects experimentally observed conformational diversity (González Buitrón et al., 2022). NAFlex supports nucleic-acid simulation and analysis, and BIGNASim organizes simulation data (Hospital et al., 2013; 2016). Table 1 compares molecular coverage, structure counts, force fields, and annotations. RiboMotion provides over one million saved conformations across 730 isolated RNAs and 309 RNA–ligand complexes, with shared dynamic prediction targets.

Table 1: **RNA structure and dynamics resources.** K = 1,000 structures.

Resource	**RiboMotion**	BIGNASim	DeepRMSF	RNA3DB	CoDNaS-RNA
Data type	MD trajectories			Static structures	
Systems / groups	**1,039 systems**	37 systems	371 systems	1,645 clusters	1,643 clusters
Structures	**1,032K**	≈843K[‡]	≈742K[†]	11.2K	15.1K
Force field	ff99bsc0 + χOL3; GAFF2	AMBER variants	ff99bsc0 + χOL3	—	—
Coverage	RNA; RNA–ligand	RNA; Protein–RNA	RNA	RNA	RNA; complexes
Annotations	Flexibility; geometry; stability	Geometry; mechanics; NMR	Atomic RMSF	Structures; RNA families	Conformational diversity; interactions

Structures are experimental models or saved MD conformations; counts do not imply independence. Snapshots: RNA3DB, January 2024; CoDNaS-RNA, January 2026. Clusters and systems use different counting units.
[†]Conditional estimate: 371 × 200 ns at an assumed 100-ps saving interval; neither this interval nor the MD conformation count is reported by DeepRMSF.
[‡]Estimated from metadata and export filenames for 37 RNA-related BIGNASim systems, including hybrids (September 2026); trajectory files were not enumerated.

Predicting RNA dynamics. Learning methods address several aspects of RNA motion and molecular recognition. DeepRMSF predicts MD-derived flexibility (Feng et al., 2026), providing a direct precedent for learning dynamic observables. RNAmigos2 targets virtual screening, and SHAMAN identifies small-molecule binding sites in RNA conformational ensembles (Carvajal-Patiño et al., 2025; Panei et al., 2024). RiboMotion evaluates complementary observables: fluctuation amplitudes, distance distributions, and persistent interaction disruption. Its complex tasks characterize motion and interaction stability within a given RNA–ligand system.

RNA conformational generation. DynaRNA and RNAbpFlow generate RNA conformations, expanding the modeling objective beyond an individual structure (Li et al., 2025; Tarafder & Bhattacharya, 2026). Evaluating such ensembles requires characterizing variation across conformations as well as the accuracy of individual structures. RiboMotion provides trajectory-derived distance and fluctuation distributions for evaluating generated ensembles, as well as multiple conformations for model adaptation.

3 Dataset, Tasks, and Evaluation

3.1 Construction and Coverage

RiboMotion contains 730 isolated RNA constructs and 309 RNA–ligand complexes derived from deposited structures (Figure 1). Core RNA starts from a PDB-derived collection selected for size, structural quality, or functional interest (Berman et al., 2000). We retain canonical-RNA chains or fragments whose sequence and intact backbone can be matched to the source structure. Non-RNA partners are removed, so an isolated construct extracted from a complex is not necessarily an experimentally established apo state. The complex collection uses RCSB PDB structures represented in the nucleic-acid docking benchmark of Jiang et al. (2023). We retain continuous RNA receptors

Table 2: **RiboMotion cohorts and prediction tasks.** Counts are systems; parentheses in the first panel give components. Eligible counts follow the stated task criteria. Pockets use initial heavy-atom distance < 6 Å; ligand shells use ≤ 10 Å. Core pairs are same-chain $i < j$ with $j - i \geq 4$.

Cohort	Train / validation / test	Saved window	Input notation
Core (730)	551 / 82 / 97 (396 / 55 / 77)	991 frames, 100 ps 0–99 ns	S: prepared source P: first production frame
Ligand (309)	210 / 50 / 49 (51 / 29 / 31)	1,000 frames, nominal 80 ps 20.08–100 ns[a]	M: first exported MD complex; C: chemistry
Task	**Input and scored target**	**Primary metric**	**Eligible train / val / test**
Nucleotide flexibility	S; all mapped C4′ RMSF	Spearman ↑	551 / 82 / 97
Pair distances	S; nonlocal C4′ pairs	W_1 (Å) ↓	441 / 66 / 78[b]
Stacking disruption	P; initial internal stacks	MacroAP ↑	530 / 79 / 93
Pocket flexibility	M; pocket C4′ RMSF	MAE (Å) ↓	210 / 50 / 49
Ligand distances	M; shell heavy-atom minima	W_1 (Å) ↓	210 / 50 / 49
Hydrogen-bond disruption	M+C; initial typed contacts	MacroAP ↑	167 / 40 / 39

[a]The nominal 79.92-ns span follows from 1,000 saved observations at 80-ps intervals. [b]All-eligible uses trajectory-consistency screening; Verified contains 257/32/46 systems. All 49 ligand test systems contribute continuous targets (531 pocket and 971 shell residues). For hydrogen bonds, 49 registered systems yield 43 chemistry-valid and 39 candidate-bearing systems; chemistry failures are not negative examples.

with identifiable ligands and analyzable MD trajectories. Appendix A gives the inclusion and quality criteria; these cohorts are not an exhaustive survey of PDB RNA.

Both simulation protocols use ff99/bsc0/χOL3 RNA parameters, TIP3P water, 300 K, and a 2 fs time step (Zgarbová et al., 2011; Jorgensen et al., 1983). Core uses GROMACS NPT dynamics at 1 bar with neutralizing ions and no added bulk salt (Abraham et al., 2015); complexes use Amber PMEMD CUDA NVT dynamics with approximately 0.15 M NaCl, GAFF2 ligand parameters, and AM1-BCC charges (Salomon-Ferrer et al., 2013; Jakalian et al., 2002). These distinct conditions define two evaluation cohorts rather than an apo/holo comparison.

All 1,039 trajectories pass checks for correspondence to the intended molecular construct, finite coordinates, expected saved-frame counts, and valid continuous targets. Production logs and topology-based covalent connectivity are also checked for all 730 Core systems. Task-specific requirements, including ligand chemistry for hydrogen bonds, determine the eligible populations in Table 2.

3.2 Dynamic targets and grouped partitions

We partition related molecular constructs together to test generalization beyond near-duplicate RNA structures. A relation graph links identical, sequence-similar, or structurally similar constructs; Core also uses Rfam and alternative-conformation relationships, and complexes sharing a source PDB are linked. Complete components enter only one split. The allocation balances RNA length, source metadata, and task eligibility: median train/validation/test lengths are 28/28/26 nucleotides for Core and 67/66/64 for complexes. Appendix B gives the relation thresholds and allocation criteria. These groups do not enforce ligand-scaffold separation or independence between the two tracks.

Table 2 defines the task inputs and lists eligible systems. Flexibility is the root-mean-square fluctuation (RMSF) of aligned C4′ coordinates about their temporal mean. Geometry targets are empirical distance distributions over the recorded window, scored by exact one-dimensional Wasserstein distance W_1. Evaluators average nucleotides or pairs within each system and then weight systems equally. The continuous Core tasks use prepared source structures; stacking uses the first production frame. Ligand tasks use the first exported MD complex, with mapped chemistry additionally required for hydrogen bonds. Core distance eligibility also uses predefined trajectory-consistency criteria (Appendix C).

Interaction-disruption tasks start from clearly present stacking or hydrogen-bond candidates. Frame-wise geometry then classifies each candidate as interacting, disrupted, or ambiguous. A positive event requires a sustained run of clearly disrupted observations; interacting or ambiguous states break the run. Appendix C gives the geometric thresholds, boundary rule, and formal event definition in Equation 3. Event MacroAP assigns an AP of zero to candidate-bearing systems without positives, so its fixed-population maximum is the fraction of systems with positive events. We also report pooled AP and Brier error to characterize event ranking and probability quality.

3.3 Baselines and evaluation

The benchmark compares physical and initial-geometry references, histogram gradient boosting (HGB) on explicit geometric features, and neural models. Hyperparameters and checkpoints use training and validation data. The primary Core distance models fit the 257-RNA Verified training stratum and evaluate on all 78 eligible test RNAs. Continuous ligand models use a joint selection rule over flexibility, distance, and an auxiliary occupancy objective; occupancy is not one of the six benchmark endpoints. The main table reports this joint-objective benchmark. The parameterization controls select checkpoints by validation distance error to isolate the effect of the distance head, whereas the primary ligand benchmark uses joint-objective selection.

Neural scores average three separately evaluated optimization seeds; reported standard deviations describe those scores. Paired intervals resample whole components 2,000 times after within-system seed averaging. All task-specific models fit the released training partition, select checkpoints on validation data, and are evaluated on its fixed test partition. Appendix B specifies the component allocation.

Dynamic information and sampling. Figure 2 connects dataset coverage with two properties of the targets. Permuting hydrogen-bond states preserves occupancy but changes persistent-disruption rates, demonstrating the contribution of time ordering. Independent simulations on six distance-eligible Core RNAs provide a sampling reference for model errors. Half-window flexibility profiles remain consistent, with median within-system Spearman correlations of 0.852 for Core validation and 0.760 for ligand validation. Full window, event, and replica analyses appear in Appendices D and E.1; the targets describe the stated simulation windows rather than established equilibrium distributions (Grossfield et al., 2018).

4 Experiments and Results

4.1 Predicting RNA flexibility

Table 3 groups the six endpoints by their prediction target. For nucleotide flexibility, geometry HGB achieves Spearman correlation 0.636, compared with 0.603 for a fixed-coordinate EGNN variant (Satorras et al., 2021), 0.550 for DeepRMSF-Open-Preprocess, and 0.294 for a Gaussian network model (GNM) (Yang et al., 2006). DeepRMSF covers all 97 test RNAs without RiboMotion retraining; its normalized scores support a rank comparison but not an Å-scale error comparison. HGB's paired advantage over EGNN is 0.033 [0.003, 0.062]. Initial geometric descriptors therefore provide a strong basis for predicting where an RNA will fluctuate most.

Pocket flexibility asks a different question: the absolute fluctuation amplitude near a bound ligand. The full-complex network reaches an MAE of 0.729 ± 0.008 Å, compared with 0.780 for full-context HGB and 0.815 for reduced-context HGB. Its mean gain over the latter is 0.086 Å, with interval $[-0.001, 0.208]$. HGB predicts nucleotide flexibility rankings better than EGNN. For pocket flexibility, the full-complex network has the lowest mean MAE, although the paired interval for its gain over reduced-context HGB includes zero.

4.2 Predicting geometric distributions

Predicting a distance distribution requires recovering variation around the initial geometry. On Core, EGNN reduces W_1 to 3.791 ± 0.027 Å, compared with 3.982 for HGB, 4.140 for a calibrated Gaussian, and 4.378 for the initial point reference. The paired gain over HGB is 0.192 Å [0.086,

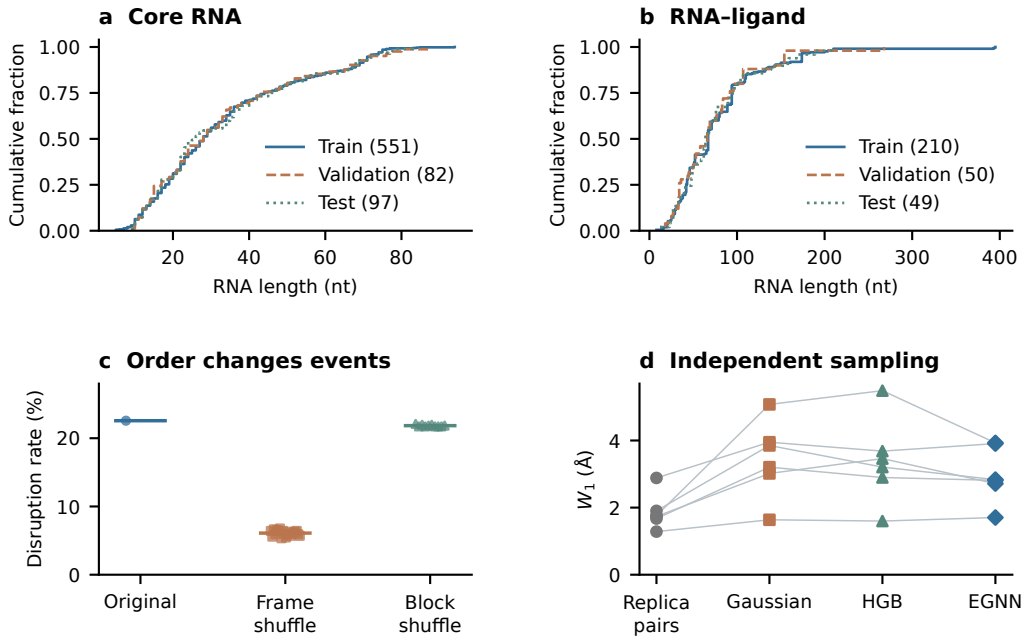

Figure 2: **Partition balance, temporal labels, and sampling variability.** (a,b) RNA lengths by split. (c) Development-cohort hydrogen-bond disruptions: 842 initial candidates; 20 frame shuffles and ten 64-frame block shuffles. Marks show means. (d) Six RNAs/five components, four independent replicas each. Lines compare mean replica-pair discrepancies with model errors averaged over replicas and three seeds. Shape/color identifies methods cross-fitted by component. Replica differences reflect finite-window variability, not an irreducible error bound. See Appendices D and E.

0.293]. The same predictors also improve distance prediction across all 97 test RNAs that meet the coordinate and target-validity requirements: EGNN/HGB errors are 4.149/4.305 Å, a gain of 0.156 Å [0.055, 0.282] (Appendix F.5).

For RNA–ligand distances, the full-complex network improves over full-context HGB by 0.238 Å [0.103, 0.387], but its 0.794-Å error remains above the calibrated Gaussian's 0.736 Å. The initial structure thus supplies a particularly strong prior in this setting. Section 4.6 examines how anchoring learned predictions to that prior changes the result. Independent Core replicas provide a complementary check: EGNN improves over the Gaussian by 0.472 Å [0.050, 0.815] on six eligible RNAs, with the protocol and sampling comparison shown in Figure 2d and Appendix E.1.

4.3 Predicting interaction disruption

Persistent disruption asks whether a clearly present interaction enters a sustained off state. For stacking, geometry MLP reaches 0.587 ± 0.010 MacroAP, compared with 0.573 for logistic regression and 0.578 ± 0.007 for scalar–vector residual prediction. A base graph network trails logistic regression by 0.046 [0.011, 0.079]. Rich initial geometry remains an effective reference for this event task; adding architectural complexity alone does not ensure better ranking.

Hydrogen-bond disruption evaluates 39 candidate-bearing complexes. Chemistry + HB GNN and geometry HGB both reach approximately 0.437 MacroAP, but the GNN improves Brier error from 0.218 to 0.196 ± 0.004. Full-context HGB gives 0.195. The fixed-population MacroAP ceiling is $20/39 \approx 0.513$, because only 20 systems contain positives. Table 3 therefore includes Brier error to assess probability quality across the full candidate population; additional pooled metrics appear in Appendix F.6.

Table 3: **RNA dynamics prediction across six tasks.** Metrics: Core flexibility, Spearman ↑; pocket flexibility, MAE (Å) ↓; distances, W_1 (Å) ↓; disruption, MacroAP ↑. RiboMotion-trained neural scores are mean ± sample SD over three separately scored seeds. Bold: best unrounded mean among displayed methods, not significance. Appendix F gives complete results and paired intervals.

Endpoint (test systems)	Method	Primary score
Nucleotide flexibility (97)	GNM	0.294
	DeepRMSF-Open-Preprocess[c]	0.550
	Geometry HGB	**0.636**
	Core EGNN	0.603 ± 0.018
Pocket flexibility (49)	Reduced-context HGB	0.815
	Full-context HGB	0.780
	Full-complex neural	**0.729 ± 0.008**
Pair distances (78)	Initial point	4.378
	Calibrated Gaussian	4.140
	Geometry HGB	3.982
	Core EGNN	**3.791 ± 0.027**
Ligand distances (49)	Initial point	0.817
	Calibrated Gaussian	**0.736**
	Full-context HGB	1.032
	Full-complex neural	0.794 ± 0.020
Stacking disruption (93)	Prevalence prior	0.265
	Logistic regression	0.573
	Geometry HGB	0.559
	Geometry MLP	**0.587 ± 0.010**
	SV residual	0.578 ± 0.007

Hydrogen-bond disruption (39 test systems) Method	MacroAP ↑	Brier ↓
Training prior	0.313	0.219
Geometry HGB	0.437	0.218
Full-context HGB	0.420	**0.195**
Chemistry + HB GNN	**0.437 ± 0.014**	0.196 ± 0.004
Heterogeneous SV	0.436 ± 0.014	0.225 ± 0.024

GNM correlation is defined for 96 of 97 RNAs. Continuous complex models use joint selection with auxiliary occupancy supervision; learning curves and distance controls use their stated selection criteria. Hydrogen-bond MacroAP has a fixed-population maximum of $20/39 \approx 0.513$. [c]The externally pretrained DeepRMSF-Open-Preprocess fold-5 checkpoint was evaluated without RiboMotion retraining on all 97 Core test systems; its output is a normalized flexibility score, so rank correlation is comparable but RMSF MAE is not.

These event labels add information beyond average proximity. In the development cohort, 134 of 190 hydrogen-bond disruptions retain ordinary distance contact for at least 90% of frames. Individual-frame shuffling lowers the development-cohort event rate from 22.6% to 6.1%, whereas 64-frame block shuffling gives 21.8% (Figure 2c). Persistent hydrogen-bond disruption therefore captures interaction geometry and temporal ordering beyond average contact occupancy.

4.4 MD SUPERVISION IMPROVES CONFORMATIONAL ENSEMBLES

The trajectories also provide supervision for conformation generation. From the same Protenix-v1 checkpoint (Zhang et al., 2026), we adapt only the diffusion module on 550 training RNAs using either their static PDB structure, one MD frame, or 32 MD frames. All three arms use the same 27,550-update budget and fixed final checkpoint. Each model receives sequence-only input and generates 32 conformations for each of 97 test RNAs.

Multi-frame adaptation improves both ensemble shape and fluctuation amplitude (Table 4). Relative to the original model, centered distance W_1 falls by 25.4% and RMSF MAE by 35.9%. The gains also hold against matched single-frame adaptation: centered W_1 improves by 0.313 Å [0.232, 0.387] and RMSF MAE by 0.473 Å [0.229, 0.674]. Static-structure adaptation improves RMSF MAE

Table 4: **Multi-frame MD improves Protenix ensembles.** Adaptation shares 550 training RNAs, the diffusion-module budget, final-step selection, and one optimization seed. Targets: one PDB structure, one MD frame, or 32 MD frames per RNA. Evaluation uses 97 test RNAs (77 components), with 32 generated conformations each. Lower is better except for Spearman.

Metric	Original	Static PDB	One MD frame	32 MD frames
Total distance W_1 (Å)	3.629	4.050	3.581	**3.207**
Centered distance W_1 (Å)	1.296	1.359	1.281	**0.967**
Width error (Å)	1.523	1.516	1.428	**0.997**
RMSF MAE (Å)	2.062	1.872	1.794	**1.321**
RMSF Spearman	0.557	0.591	0.616	**0.661**
Geometry flag rate (%)	11.02	11.89	27.19	11.86

but worsens distance-distribution errors. These contrasts show the value of multiple conformations beyond adaptation to one structure.

The single-frame arm has a higher geometric flag rate (27.19%) than the other arms (11.02–11.89%). The paired intervals above resample test components; all adapted arms share one optimization seed. Appendix H.1 gives the screening criteria, training settings, and full paired comparisons.

We also evaluate pretrained DynaRNA (Li et al., 2025) and RNAbpFlow (Tarafder & Bhattacharya, 2026) on the same 97 RNAs and common C4′ metrics. Their centered W_1 values are 3.750 and 1.460 Å, respectively, versus 0.967 Å after multi-frame Protenix adaptation. These external references use different inputs and pretraining; Appendix H.1 specifies their configurations, overlap checks and full results.

4.5 Sampling variability and benchmark discrimination

Independent simulations make the scale of prediction errors interpretable. Across nine RNAs, replica-pair flexibility MAE is 1.329 Å versus 1.544 Å for EGNN; across six distance-eligible RNAs, replica-pair W_1 is 1.874 Å versus 2.984 Å for EGNN. The distance model therefore improves on structural references while remaining above the observed between-replica discrepancy. Replica differences contextualize finite-window variability but do not quantify how much of the generation gains would persist against an independent trajectory. Appendix E.1 reports the per-RNA comparisons and protocol matching.

We also compare the UUCG ensemble with experimental eNOE distances. Averaging its 991 MD frames reduces eNOE MAE from 0.633 Å for the accompanying MD snapshot to 0.552 Å; the source NMR structures remain closer to experiment (Appendix E.3).

4.6 Effects of data coverage and distance parameterization

Training coverage. We vary complete training components in three nested random orders while fixing optimizer-update budgets: 4,400 updates for Core distances and 8,400 for pocket prediction. Three optimization seeds are used; full-data fits are shared across subset orders. Increasing Core coverage from 25% to all eligible components lowers distance error from 3.923 to 3.826 Å, with paired gain 0.097 Å [0.001, 0.177] when uncertainty includes training and subset variation. Pocket MAE changes from 0.728 to 0.709 Å, with a paired gain of 0.018 Å [−0.033, 0.070]; this interval includes zero. Broader coverage gives a modest Core distance gain under this budget. Verified-only and matched-population controls are reported in Appendix G.1.

Distance parameterization. We compare two Gaussian-center parameterizations: absolute distance and a residual around the initial distance. Crossing these with joint or distance-only supervision gives four matched conditions. Every condition receives the initial distance and uses the same update budget and validation-distance selection rule. Under joint supervision, residual prediction lowers W_1 from 0.774 to 0.706 Å, a gain of 0.068 Å [0.041, 0.095] including training variation. Distance-only training gives a similar gain of 0.057 Å [0.034, 0.080]. Joint residual prediction also

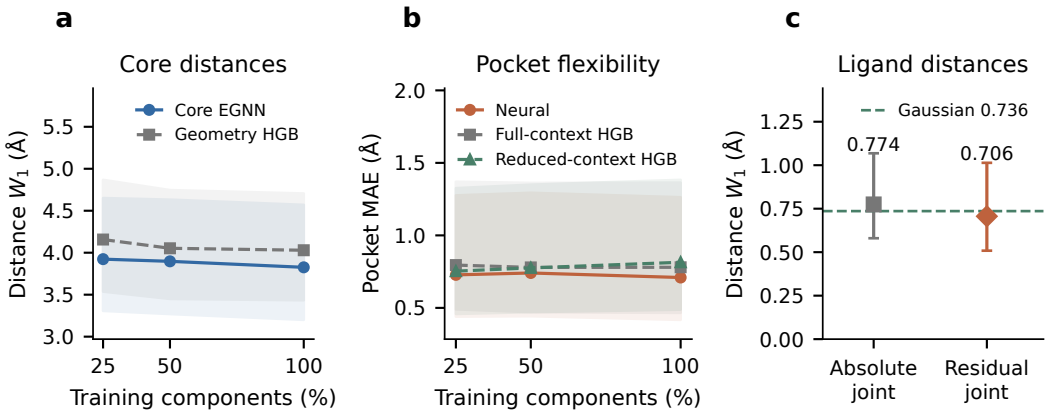

Figure 3: **Coverage and residual prediction improve distance estimates.** (a,b) Fixed-update learning curves on 78 Core RNAs/63 components and 49 complexes/31 components. Lines average optimization seeds and subset orders; three full-data fits are shared. Shading: 95% component intervals. Both ligand HGB feature sets are shown; pocket checkpoints use pocket selection. (c) Distance-selected controls on the same 49 complexes: means and component intervals, with the calibrated Gaussian dashed. Appendix G.1 reports paired intervals including training variation and the full 2×2 comparison.

improves over the calibrated Gaussian by 0.030 Å [0.020, 0.039] under component resampling, without a resolved change in pocket MAE. These controls explain how the initial structure can remain a strong prior while learned deviations add predictive value. Their selection rule differs from the primary ligand benchmark; the 0.794-Å main-table score is not the absolute-center control. Direct target-fitted Gaussian diagnostics appear only in Appendix G.4.

5 Discussion and Conclusion

We introduced RiboMotion, an atomistic MD dataset spanning 1,039 RNA systems, together with six benchmark tasks covering flexibility, geometry, and stability. Across isolated RNAs and RNA–ligand complexes, the experiments identify learnable dynamic properties beyond initial geometry. Neural models improve RNA distance-distribution prediction, and controlled comparisons show that learning residuals from initial distances improves ligand-distance prediction. For conformation sampling, multi-frame supervision improves both ensemble shape and fluctuation amplitude beyond adaptation to a single structure. RiboMotion thus supports both prediction of dynamic observables and adaptation of conformational generators. The improved ensemble fidelity motivates future work on faster conformational sampling.

Limitations and future work. Although RiboMotion provides dynamic supervision and shared benchmarks for advancing AI-driven RNA modeling, several limitations remain. The dataset spans diverse RNA types but does not yet cover many long RNAs and large assemblies. Our initial emphasis on compact functional RNAs, particularly isolated constructs shorter than 100 nucleotides, balances opportunities to sample local conformational changes with the substantial computational cost of all-atom simulations. The dynamics of larger RNAs are equally important, and we plan to expand both the size and molecular coverage of RiboMotion. The current simulation windows also provide limited access to slow transitions and rare conformational states, motivating longer trajectories and additional independent simulations. Finally, simulated ensembles depend on the chosen force fields and conditions, while experimental validation currently focuses on a UUCG hairpin. Broader comparisons with experimental measurements and alternative force fields will help assess the physical accuracy of the ensembles and strengthen their value for learning RNA dynamics.

Reproducibility Statement

The supplementary material specifies molecular preparation and simulation settings (Appendix A), grouped splits (Appendix B), target definitions and evaluators (Appendix C), model selection and benchmark results (Appendix F), and controlled experiments (Appendix G.1). The anonymous reproduction package includes labels, candidate identities and masks, frozen predictions, and a CPU evaluator for all six test endpoints. Its scores reproduce Table 3; Appendix I describes the package and generation procedures.

Data availability. The versioned MD coordinates and supporting metadata for both trajectory cohorts and a worked example are provided at https://huggingface.co/datasets/7ac23az/RiboMotion-MD. The accompanying anonymous reproduction package supports grouped-split verification and recomputation of all six test-task scores from frozen labels and predictions. It also provides generation-score records, adaptation protocols, and project-authored implementation references. Complete model retraining and raw conformation regeneration require additional prepared inputs and model assets, as specified in Appendix I.

AI Use Statement

AI tools assisted with literature discovery, research ideation and execution, and manuscript drafting and language polishing. The authors take full responsibility for the content and conclusions of this work.

References

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.

Amber Developers. The Amber force fields, n.d. URL https://ambermd.org/AmberModels.php. Official force-field documentation; accessed September 12, 2026.

Rohit Arora, Murphy Angelo, Christian Andrew Choe, Courtney A. Shearer, Aaron W. Kollasch, et al. RNAGym: Large-scale benchmarks for RNA fitness and structure prediction. bioRxiv preprint, 2025.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Sandro Bottaro, Giovanni Bussi, Giovanni Pinamonti, Sabine Reißer, Wouter Boomsma, and Kresten Lindorff-Larsen. Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA*, 25(2):219–231, 2019. doi: 10.1261/rna.067678.118.

Sandro Bottaro, Parker J. Nichols, Beat Vögeli, Michele Parrinello, and Kresten Lindorff-Larsen. Integrating NMR and simulations reveals motions in the UUCG tetraloop. *Nucleic Acids Research*, 48(11):5839–5848, 2020. doi: 10.1093/nar/gkaa399.

Juan G. Carvajal-Patiño, Vincent Mallet, David Becerra, Luis Fernando Niño Vasquez, Carlos Oliver, and Jérôme Waldispühl. RNAmigos2: accelerated structure-based RNA virtual screening with deep graph learning. *Nature Communications*, 16(1):2799, 2025. doi: 10.1038/s41467-025-57852-0.

David A. Case, Hasan Metin Aktulga, Kellon Belfon, David S. Cerutti, et al. AmberTools. *Journal of Chemical Information and Modeling*, 63(20):6183–6191, 2023. doi: 10.1021/acs.jcim.3c01153.

Andrea Cesari, Sandro Bottaro, Kresten Lindorff-Larsen, Pavel Banáš, Jiří Šponer, and Giovanni Bussi. Fitting corrections to an RNA force field using experimental data. *Journal of Chemical Theory and Computation*, 15(6):3425–3431, 2019. doi: 10.1021/acs.jctc.9b00206.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022. doi: 10.48550/arXiv.2204.00300. URL https://arxiv.org/abs/2204.00300.

Elizabeth A. Dethoff, Jeetender Chugh, Anthony M. Mustoe, and Hashim M. Al-Hashimi. Functional complexity and regulation through RNA dynamics. *Nature*, 482(7385):322–330, 2012. doi: 10.1038/nature10885.

Chenjie Feng, Xiaowen Sun, Xintao Song, Lei Bao, Weikang Gong, and Renmin Han. DeepRMSF: a deep learning-based automated approach for predicting atomic-level flexibility in RNA structure. *Briefings in Bioinformatics*, 27(1):bbaf720, 2026. doi: 10.1093/bib/bbaf720.

Laura R. Ganser, Megan L. Kelly, Daniel Herschlag, and Hashim M. Al-Hashimi. The roles of structural dynamics in the cellular functions of RNAs. *Nature Reviews Molecular Cell Biology*, 20(8):474–489, 2019. doi: 10.1038/s41580-019-0136-0.

Martín González Buitrón, Ronaldo Romario Tunque Cahui, Emilio García Ríos, Layla Hirsh, Gustavo Parisi, María Silvina Fornasari, and Nicolas Palopoli. CoDNaS-RNA: a database of conformational diversity in the native state of RNA. *Bioinformatics*, 38(6):1745–1748, 2022. doi: 10.1093/bioinformatics/btab858.

Alan Grossfield, Paul N. Patrone, Daniel R. Roe, Andrew J. Schultz, Daniel Siderius, and Daniel M. Zuckerman. Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0]. *Living Journal of Computational Molecular Science*, 1(1):5067, 2018. doi: 10.33011/livecoms.1.1.5067.

Adam Hospital, Ignacio Faustino, Rosana Collepardo-Guevara, Carlos González, Josep L. Gelpí, and Modesto Orozco. NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Research*, 41(W1):W47–W55, 2013. doi: 10.1093/nar/gkt378.

Adam Hospital, Pau Andrio, Cesare Cugnasco, Laia Codo, Yolanda Becerra, Pablo D. Dans, Federica Battistini, Jordi Torres, Ramón Goñi, Modesto Orozco, and Josep L. Gelpí. BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Research*, 44(D1):D272–D278, 2016. doi: 10.1093/nar/gkv1301.

Araz Jakalian, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, 2002. doi: 10.1002/jcc.10128.

Dejun Jiang, Huifeng Zhao, Hongyan Du, Yafeng Deng, Zhenxing Wu, Jike Wang, Yundian Zeng, Haotian Zhang, Xiaorui Wang, Jian Wu, Chang-Yu Hsieh, and Tingjun Hou. How good are current docking programs at nucleic acid–ligand docking? a comprehensive evaluation. *Journal of Chemical Theory and Computation*, 19(16):5633–5647, 2023. doi: 10.1021/acs.jctc.3c00507.

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*, pp. 22277–22303. PMLR, 2024. URL https://proceedings.mlr.press/v235/jing24a.html.

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(2):926–935, 1983. doi: 10.1063/1.445869.

In Suk Joung and Thomas E. Cheatham, III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry B*, 112(30):9020–9041, 2008. doi: 10.1021/jp8001614.

Zhengxin Li, Junjie Zhu, Xiaokun Hong, Junxi Mu, Zhuoqi Zheng, Taeyoung Cui, Yutong Sun, Ting Wei, and Hai-Feng Chen. DynaRNA: accurate dynamic RNA conformation ensemble generation with diffusion model. *Communications Biology*, 8:1472, 2025. doi: 10.1038/s42003-025-08875-2.

F. P. Panei, P. Gkeka, and M. Bonomi. Identifying small-molecules binding sites in RNA conformational ensembles with SHAMAN. *Nature Communications*, 15(1):5725, 2024. doi: 10.1038/s41467-024-49638-7.

Alberto Pérez, Iván Marchán, Daniel Svozil, Jiří Šponer, Thomas E. Cheatham, III, Charles A. Laughton, and Modesto Orozco. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophysical Journal*, 92(11):3817–3829, 2007. doi: 10.1529/biophysj.106.097782.

Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, Xinzhu Ma, Siqi Sun, Hongliang Yan, Dong Yuan, Wanli Ouyang, and Xihui Liu. BEACON: Benchmark for comprehensive RNA tasks and language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2949. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a8ea503d91320fcfe12cba61c8a6d285-Abstract-Datasets_and_Benchmarks_Track.html.

Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, 2013. doi: 10.1021/ct400314y.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, et al. Accurate RNA 3d structure prediction using a language model-based deep learning approach. *Nature Methods*, 21:2287–2298, 2024. doi: 10.1038/s41592-024-02487-0.

Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, et al. MISATO: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature Computational Science*, 4: 367–378, 2024. doi: 10.1038/s43588-024-00627-2.

Marcell Szikszai, Marcin Magnus, Siddhant Sanghi, Sachin Kadyan, Nazim Bouatta, and Elena Rivas. RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction. *Journal of Molecular Biology*, 436(17):168552, 2024. doi: 10.1016/j.jmb.2024.168552.

Sumit Tarafder and Debswapna Bhattacharya. RNAbpFlow: base pair-augmented SE(3) flow matching for conditional RNA 3d structure generation. *Nature Methods*, 23:1349–1358, 2026. doi: 10.1038/s41592-026-03128-4.

Jiří Šponer, Giovanni Bussi, Miroslav Krepl, Pavel Banáš, Sandro Bottaro, et al. RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chemical Reviews*, 118(8):4177–4338, 2018. doi: 10.1021/acs.chemrev.7b00427.

Lee-Wei Yang, A. J. Rader, Xiong Liu, Cristopher Jon Jursa, Shann Ching Chen, Hassan A. Karimi, and Ivet Bahar. oGNM: Online computation of structural dynamics using the gaussian network model. *Nucleic Acids Research*, 34(Web Server issue):W24–W31, 2006. doi: 10.1093/nar/gkl084.

Marie Zgarbová, Michal Otyepka, Jiří Šponer, Arnošt Mládek, Pavel Banáš, Thomas E. Cheatham, III, and Petr Jurečka. Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of Chemical Theory and Computation*, 7(9):2886–2902, 2011. doi: 10.1021/ct200162x.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, 19(9): 1109–1115, 2022. doi: 10.1038/s41592-022-01585-1.

Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, Yu Xia, and Wenzhi Xiao. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, 2026. doi: 10.64898/2026.02.05.703733. URL https://www.biorxiv.org/content/early/2026/02/22/2026.02.05.703733.1.

SUPPLEMENTARY MATERIAL

Appendices A–C describe construction, partitions, and targets; D–E provide detailed dynamic-target and independent-simulation evidence; F–G report models, benchmarks, and controls; H details ensemble generation; I documents data and reproduction resources.

A DATA CONSTRUCTION AND SIMULATION PROTOCOLS

A.1 STRUCTURAL SOURCES AND CONSTRUCT SELECTION

The Core cohort comprises 730 isolated constructs from a PDB-derived RNA collection (Berman et al., 2000) focused on RNAs shorter than 100 nucleotides with structural-quality or functional-priority annotations. We use the first deposited model and retain a canonical-RNA chain or fragment when its sequence and covalent backbone can be matched to the source. We remove non-RNA partners. These constructs may originate from complexes and are not necessarily experimentally established apo states. Table S1 summarizes the final cohort and quality checks.

The 309 complexes originate from RCSB PDB structures represented in the nucleic-acid docking collection of Jiang et al. (2023). We retain systems with a continuous RNA receptor, identifiable ligand, and analyzable MD trajectory. Ligands receive system-specific parameters from their molecular structures or Chemical Component Dictionary-assisted structures. Analysis chemistry is a separate requirement for the hydrogen-bond task, not a condition for membership in the 309-system cohort.

Table S1: Final RiboMotion cohorts and quality-audit coverage. Counts refer to retained systems. Ligand analysis chemistry determines hydrogen-bond eligibility without changing MD cohort membership.

Property	Isolated RNA	RNA–ligand complexes
Structural source	PDB-derived RNA collection; canonical, source-matched RNA construct	RCSB PDB structures represented in the nucleic-acid docking collection; RNA receptor and identifiable ligand
Retained MD systems	730	309
Source/coordinate/target checks	730/730 pass correspondence, atom-count, finite-coordinate, frame-count, and continuous-target checks	309/309 pass the same export-level checks
Additional physical-record audit	730/730 have production-log and topology-bond checks	Complete raw-log, topology, and original Amber time-record coverage not established
Chemistry for HB analysis	Not applicable	261/309 mapped; unmapped systems remain in the cohort and continuous tasks

The cohorts are selected resources, not an exhaustive census of eligible PDB RNA. Task-specific candidate and train/validation/test counts appear in Table S3.

A.2 SIMULATION SETTINGS AND AUDIT COVERAGE

Table S2 gives the two protocols. Both use ff99 with bsc0/χOL3 RNA refinements and TIP3P water (Pérez et al., 2007; Zgarbová et al., 2011; Jorgensen et al., 1983). The ff99/bsc0/χOL3 combination is a documented RNA refinement, but RNA force fields and ions can change sampled populations

(Šponer et al., 2018; Cesari et al., 2019). We therefore treat the two protocols as separate data-generating settings and do not attribute cross-track differences to ligand binding. Core uses GROMACS (Abraham et al., 2015); complexes use AmberTools and Amber CUDA MD (Case et al., 2023; Salomon-Ferrer et al., 2013), GAFF2 ligand parameters (Amber Developers, n.d.), AM1-BCC charges (Jakalian et al., 2002), and TIP3P-matched monovalent ions (Joung & Cheatham, 2008).

Core exports contain 991 frames at 100-ps intervals over 0–99 ns. Complex exports contain 1,000 frames nominally spaced by 80 ps across 20.08–100 ns. Event labels use the saved-frame rule in Appendix C. These frames define the observation windows and are not independent molecular samples.

All 1,039 retained exports pass checks of source-construct correspondence, RNA identity, atom counts, finite coordinates, frame counts, and numerical validity of continuous targets. Production-log and topology-based bond checks additionally cover every Core system, including finite sampled energies and temperatures and matching covalent connectivity. Ligand chemistry can be mapped for hydrogen-bond analysis in 261 of the 309 complexes; the remaining 48 retain valid dataset membership and continuous-task targets but are ineligible for the chemistry-dependent event task. These checks establish export integrity; sampling consistency is assessed through window and replica comparisons.

Table S2: Molecular-dynamics protocols and audit coverage. Per-system records link the 730 Core systems and 309 complexes to their structural and parameter inputs. Complex export times follow the nominal protocol; event labels are defined by consecutive saved frames.

Setting	Core RNA	RNA–ligand
Engine	GROMACS 2024.5 (conda-forge build).	Initial 0–10 ns: Amber24 `pmemd.cuda`. Extensions: Amber24 for 20 systems; Amber26 for 289 systems.
RNA parameters	ff99+bsc0+χOL3. The directory name `amber14sb_OL15.ff` does not imply DNA OL15 parameters for RNA.	ff99bsc0/χOL3.
Ligand parameters	Not applicable.	GAFF2 atom types, AM1-BCC charges; `antechamber`/`parmchk2` with recorded system-specific formal charge.
Water and box	TIP3P, dodecahedral box, 1.2-nm padding.	TIP3P, truncated octahedral box, 12-Å padding.
Ions	K^+/Cl^- neutralization, with no additional bulk salt specified.	Na^+/Cl^- neutralization plus approximately 0.15 M NaCl; additional ion pairs $\mathrm{round}(N_{\mathrm{water}}\,0.15/55.5)$. TIP3P-matched Joung–Cheatham parameters.
Minimization	Steepest descent, up to 50,000 steps; force tolerance 500 kJ mol^{-1} nm^{-1}; step 0.01 nm; position restraints enabled.	Two 5,000-cycle stages, each 2,500 steepest-descent and 2,500 conjugate-gradient cycles: solute-heavy restraint 10 kcal mol^{-1} Å^{-2}, then unrestrained.
Equilibration	100 ps restrained NVT followed by one executed 1-ns restrained NPT stage.	125 ps NVT heating from 0 to 300 K with restraint 5 kcal mol^{-1} Å^{-2}; NPT 0.5 ns with restraint 2, 0.5 ns with 0.5, then 1 ns unrestrained.
Thermostat	V-rescale at 300 K; $\tau_T = 0.1$ ps.	Langevin at 300 K; friction 2 ps^{-1}.
Pressure and ensemble	Isotropic C-rescale during NPT equilibration and production; 1 bar, $\tau_P = 2$ ps, compressibility 4.5×10^{-5} bar^{-1}.	Isotropic Monte Carlo barostat during NPT equilibration; 1 atm, `taup`=2 ps. Production is NVT.
Integration	2 fs; LINCS constraints on bonds involving hydrogen, iteration 1/order 4.	2 fs; SHAKE constraints on bonds involving hydrogen.
Nonbonded terms	PME, order 4, Fourier spacing 0.12 nm; Coulomb/van der Waals cutoffs 1.0 nm; Verlet lists and energy/pressure dispersion correction.	PME and 10-Å real-space/nonbonded cutoff.

Continued on next page

Setting	Core RNA	RNA–ligand
Production	1-ns unrestrained sentinel, then a 99-ns continuation from its checkpoint without new velocities.	10-ns initial production, then nine serial 10-ns extensions from predecessor restarts: 100 ns nominal total.
Seeds	Per-system NVT velocity seeds are recorded; continuation does not redraw velocities.	Initial base seed 2023 across distinct systems; extension base seeds are system-specific, with segment-specific derived seeds. Serial extensions are not replicas.
Coordinate export	Original coordinates every 10 ps; RNA-selected, made whole, centered, fitted and subsampled. 991 exported frames every 100 ps, 0–99 ns.	Eight segments covering 20–100 ns; 125 saved frames each at nominal 80-ps intervals. Source processing autoimages, fits receptor heavy atoms and removes solvent/ions. 1,000 frames; nominal 20.08–100 ns.
Quality checks	All 730 retained systems have finite production energies/temperatures, no LINCS warnings, and matching topology-defined covalent bonds.	Export-level identity, coordinate, and continuous-target checks cover all 309 retained systems. Event labels use the consecutive-observation definition in Appendix C.
Independent simulations	Four separately prepared runs for each of nine RNA constructs at 300 K with charge-neutralizing ions and no added bulk salt; excluded from the main cohort count.	No independent-replica panel is claimed for the 309-system cohort.

A.3 Replica identity and ligand distance corrections

The independent-simulation panel contains four separately prepared runs for each of nine Core RNA constructs. Production records establish a matched 300-K, charge-neutralized protocol without added bulk salt. Comparisons use 991 frames per run over 1–100 ns; repeated identities or trajectory extensions are not counted as additional independent RNAs.

The ligand-distance labels include an inherited correction for abrupt whole-ligand translations: a median heavy-atom step above 15 Å triggers subtraction of the accumulated displacement. We apply it to a small subset of complexes, where it can affect shell-residue distances; the complete exposure inventory accompanies the label records. Available exports lack valid periodic boxes, so physical validation of these corrections still requires the original boxed Amber trajectories and compatible topologies. This scope audit made no changes to labels or reported scores.

B Grouped Partitions and Task Coverage

B.1 Molecular relations and assignment

The split unit is a connected component of related molecular constructs. Exact constructs or sequences are linked; sequence-similarity edges require identity and bidirectional coverage of at least 0.8, and US-align structure edges require $\min(\mathrm{TM}_1, \mathrm{TM}_2) \geq 0.55$ (Zhang et al., 2022). Core also links Rfam families and documented alternative conformations, while complexes from the same source PDB stay together. Transitive relations remain intact. Ligand scaffolds are not used to define components, and biological independence between tracks is not assumed.

Joint allocation. We assign complete components to train/validation/test, producing 551/82/97 Core systems and 210/50/49 complexes. Allocation balances length, source composition, and per-task eligibility while preventing any single large component from dominating a held-out partition. Event-positive counts and model scores do not enter allocation.

The balance criteria use RNA-length bins and cumulative thresholds, source metadata, and task-eligible systems and components. Core source metadata include experimental method and Rfam availability; complex source categories describe preparation/export batches, not different structural databases. The Core distance criterion includes its predefined trajectory-consistency screen, whereas

event eligibility depends on an initial interaction candidate. Exact allocation settings and the final per-component assignments accompany the split manifest.

The released manifest defines the exact partition; its grouping relations are not recomputed for individual tasks. This design preserves molecular relatedness while giving each endpoint a useful held-out population.

The resulting component counts are 396/55/77 for Core and 51/29/31 for complexes. Core median train/validation/test lengths are 28/28/26 nt, and complex medians are 67/66/64 nt. The corresponding train–test empirical Kolmogorov–Smirnov distances are 0.063 and 0.112, indicating similar length distributions across the assigned cohorts.

Evaluation population. All task-specific benchmark models are fitted on the grouped training systems, selected on validation systems, and frozen before scoring the test systems. The test partition assesses grouped generalization within the curated cohort; no prospectively collected external cohort is claimed. The ensemble-adaptation experiment uses a fixed final checkpoint and its pretrained reference (Appendix H.1).

B.2 Task coverage

Table S3 distinguishes registered systems, eligible task domains and prediction instances. Core distance fitting in the primary benchmark uses 257 Verified training systems and all 66 eligible validation systems; the coverage controls use different training pools. All 49 test complexes enter both continuous tasks. For HB loss, chemical mapping and the presence of an initial candidate are separate requirements; chemistry failures are not negatives.

Table S3: Task coverage across the grouped partitions. Every triplet is train/validation/test. Instances are nucleotides, adjacent pairs or residue–ligand candidates, as indicated. Dashes mean not inventoried or not applicable, not zero.

Endpoint / scope	Systems	Instances	Positive events
Core registered	551 / 82 / 97	—	—
Nucleotide flexibility	551 / 82 / 97	All mapped nucleotides	—
Pair-distance distribution (All)	441 / 66 / 78	Eligible nonlocal pairs	—
Verified subset	257 / 32 / 46	Verified nonlocal pairs	—
Challenge subset	184 / 34 / 32	Remaining eligible pairs	—
Stacking loss	530 / 79 / 93	5,357 / 836 / 979	1,133 / 172 / 211
Ligand registered	210 / 50 / 49	—	—
Pocket flexibility	210 / 50 / 49	2,563 / 615 / 531	—
Residue–ligand distance distribution	210 / 50 / 49	4,549 / 1,092 / 971	—
Analysis chemistry pass	174 / 44 / 43	—	—
Analysis chemistry fail	36 / 6 / 6	—	—
Hydrogen-bond loss	167 / 40 / 39	698 / 144 / 139	156 / 34 / 44
Chemistry pass, no HB	7 / 4 / 4	0 / 0 / 0	—

Task-domain coverage does not imply that every secondary statistic is defined. Constant profiles or one-class systems can have undefined correlations/AUROC. Conversely, candidate systems without positive events are retained in primary event MacroAP with AP $= 0$. The hydrogen-bond-loss test has 39 candidate systems from 26 components, including 20 positive-containing systems; its maximum possible all-system MacroAP is therefore $20/39 \approx 0.513$. Pocket flexibility and residue–ligand distance distributions score all 49 test complexes from 31 components. Their neural baselines use joint training with auxiliary occupancy supervision.

C Prediction Targets and Evaluators

C.1 Input structures and observation windows

Core continuous predictions use the prepared source PDB and sequence, while Core stacking predictions use the first saved production frame and sequence. Ligand predictors use the released complex PDB, which represents its first exported MD frame; HB prediction additionally requires the accepted analysis chemistry and atom mapping. These input origins are not interchangeable. Candidate event masks use only the corresponding static input. In contrast, the frozen Core distance-distribution sampling strata described below depend on trajectory diagnostics and are not prospective static filters.

Core labels use all 991 exported frames; the continuous-label halves contain 495/496 frames and five contiguous blocks contain 199/198/198/198/198 frames. Canonical residue ordinals follow the prepared construct rather than gaps in deposited numbering. Periodic tasks require a finite, valid triclinic box. Cartesian distances are not substituted when a valid box is unavailable. Complex labels use all 1,000 exported frames, with two 500-frame halves. Future event confirmation excludes frame zero in both tracks. Continuous targets summarize their entire defined window, including its initial saved frame.

C.2 Flexibility and distance-distribution endpoints

Let $\widetilde{\mathbf{x}}_{it}$ be the aligned C4′ position of nucleotide i at frame t. Core chains are first made whole under the periodic box; all eligible C4′ sites then provide a global Kabsch fit to the first saved frame. Complex RNA likewise uses an all-RNA C4′ fit, excluding the ligand from the fit. Flexibility about the aligned temporal mean is

$$f_i(W) = \left[\frac{1}{|W|} \sum_{t \in W} \|\widetilde{\mathbf{x}}_{it} - \overline{\mathbf{x}}_{i,W}\|^2 \right]^{1/2}. \tag{1}$$

Nucleotide flexibility uses the RMSF in Equation 1, in angstroms, over all mapped nucleotides. Its primary score is within-system Spearman correlation, averaged equally across systems. Pocket flexibility uses the same physical statistic and scores pocket MAE: the pocket is fixed by the minimum RNA-residue/ligand heavy-atom distance in the released PDB, strictly below 6 Å. Constant or fewer-than-three-residue pocket profiles have undefined Spearman but can still contribute MAE. The pocket-flexibility evaluator scores finite signed regression outputs without clipping; physical RMSF targets remain nonnegative. Each half/block RMSF has its own temporal mean under the stored alignment convention.

For Core pair-distance distributions, same-chain pairs satisfy $i < j$ and $j - i \geq 4$, including terminal residues. Their variable $d_{ij,t}$ is the periodic C4′ distance without structural alignment. The target empirical measure and scoring rule are

$$\widehat{P}_{ij} = \frac{1}{|W|} \sum_{t \in W} \delta_{d_{ij,t}}, \qquad W_1(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)|\, dx. \tag{2}$$

Equation 2 compares predicted empirical samples with the full observed distance trace in angstroms. Equal-size empirical distributions reduce to the mean absolute difference of sorted samples; otherwise the evaluator integrates the CDF difference on the union of supports. Errors are averaged over pairs within each system and then equally over systems. Histogram plots or training quantiles do not replace the full trace in final scoring.

The Core distance-distribution entry screen requires half-window RMSF Pearson ≥ 0.70, median relative half-window RMSF difference ≤ 0.50, and mean absolute half-window contact-occupancy difference ≤ 0.15 when that contact statistic is defined. The relative difference divides by full-window RMSF floored at 10^{-6} nm. Contact occupancy is a sampling-consistency criterion, not a prediction endpoint. These criteria produce 585 All-eligible systems. Verified further requires median pair half-window $W_1 \leq 1.0$ Å, its 90th percentile ≤ 2.5 Å, and median pair blockwise $W_1 \leq 1.5$ Å; the last statistic covers the ten comparisons among five contiguous blocks. The remaining eligible systems form Challenge. These names describe finite-window criteria, not verified equilibrium ensembles.

For residue–ligand distance distributions, $D_{it} = \min_{a\in i, b\in L} \|\mathbf{x}_{at} - \mathbf{x}_{bt}\|$ uses RNA-residue and ligand heavy atoms. Its shell is fixed at released-PDB distance ≤ 10 Å. The continuous-label construction aligns RNA coordinates and subtracts accumulated whole-ligand translations when the median heavy-atom step exceeds 15 Å; it does not apply a new minimum-image calculation. Physical validation of this correction requires the original boxed trajectories (Appendix A.3). The distance task scores exact W_1 over 1,000 target distances, averaged over shell residues and then systems. The candidate inventory and task-matched half-window analysis cover all 309 complexes, including the 49 test systems (Appendix D.2).

C.3 Persistent stacking and hydrogen-bond loss

Each event candidate has a static, clearly on initial state and a future ternary sequence $s_t \in \{1, 0, -1\}$ for on, off and ambiguous geometry. With the saved-frame persistence threshold K, the target is

$$y = \mathbb{1}[\exists\, 1 \leq a \leq b < T : s_a = \cdots = s_b = 0 \ \wedge\ b - a + 1 \geq K]\,. \tag{3}$$

An on or ambiguous observation breaks the off run. No future persistence of the initial on state is required for admission. We use $K = 11$ for Core and $K = 14$ for complexes. The first-to-last observation spans 1.00 ns at Core's 100-ps interval and nominally 1.04 ns at the complex protocol's 80-ps interval. An off run reaching the right boundary before confirmation remains a negative finite-window label, with a separate incomplete-run flag; it is not discarded. The target means that a sustained loss was observed in this window, not that a negative candidate never loses contact or that the model estimates an unbiased lifetime.

Core candidates are adjacent bases $(i, i+1)$ that involve neither of the first two nor the last two nucleotides and are clearly stacked in production frame zero. We use Barnaba local base frames (Bottaro et al., 2019) with the following explicit margins. In angstroms, let $\mathbf{r}_1, \mathbf{r}_2$ express the inter-base-center displacement in the two frames, $z_k = |r_{k,z}|$, $\rho = \min(\|\mathbf{r}_{1,xy}\|, \|\mathbf{r}_{2,xy}\|)$, and $c = |\mathbf{n}_1 \cdot \mathbf{n}_2|$. Define

$$E = \max_{k\in\{1,2\}} \left[(r_{k,x}/5)^2 + (r_{k,y}/5)^2 + (r_{k,z}/3)^2\right]. \tag{4}$$

Using E from Equation 4, the on core requires $E < 1.6^2$, both $z_k > 2.2$ Å, $\rho < 2.2$ Å and $c > \cos 30°$. The off core holds if $E \geq 1.8^2$, either $z_k \leq 1.8$ Å, $\rho \geq 2.8$ Å or $c \leq \cos 50°$. Other configurations are ambiguous. Periodic base geometry uses the original box. These margins define the benchmark's clear-state events; they are not presented as a universal kinetic definition of stacking.

Ligand HB candidates are residue–ligand instances with at least one chemically permitted donor–hydrogen–acceptor triplet clearly on in the released PDB. An instance is on if any triplet has donor–acceptor distance ≤ 3.5 Å and D–H–A angle $\geq 130°$. It is ambiguous if no triplet is on but at least one satisfies distance ≤ 3.8 Å and angle $\geq 110°$; it is off otherwise. Permitted triplets use mapped ligand chemistry, observed hydrogen identities and canonical RNA donor/acceptor assignments. Formation of a different permitted HB in the same residue–ligand instance prevents an off state: the task does not track the lifetime of one specific atom pair. Future states for the primary HB endpoint use raw exported coordinates; translation-corrected states are evaluated separately as a sensitivity analysis.

C.4 Aggregation, missingness and interpretation

Both event endpoints use RNA MacroAP over all systems with nonempty static candidate sets. AP is zero for a system containing no positive candidate, irrespective of false positives; tied scores are handled as tied thresholds. Systems with no candidates remain in coverage counts but have no invented prediction rows. Positive-system AP, pooled AP and AUROC on two-class systems are secondary. If N_+ of N scored systems contain positives, the maximum all-system MacroAP is N_+/N; for HB test this is $20/39 \approx 0.513$. A constant profile's undefined correlation or a one-class system's undefined AUROC is omitted only from that metric, with its denominator reported.

Scores are first computed within each system and training seed. Seed means and sample standard deviations describe training variation. Component uncertainty averages scores across seeds within

a system and resamples complete components, preserving all their member systems; repeated seeds are not independent molecular samples. Half-window diagnostics measure sensitivity within one trajectory and do not substitute for independent MD replicas (Grossfield et al., 2018). Model-specific training settings and full results appear in Appendix F.

D Structural and Sampling Properties

D.1 Structural observables and temporal controls

Barnaba 0.1.9 (Bottaro et al., 2019) annotates pairing and stacking in the initial structure. A nucleotide is structured if paired or stacked, and unpaired-internal otherwise, excluding chain termini. Within each RNA, RMSF is converted to normalized ranks; the contrast subtracts the median structured rank from the median unpaired-internal rank and requires at least three sites per class. It is defined for 283 training and 35 validation RNAs. The median effect is 0.324, with a 95% component interval of [0.276, 0.348].

A secondary global descriptor, the 90th-percentile base-frame structural distance (eRMSD) (Bottaro et al., 2019), correlates with initial-contact retention at -0.613 $[-0.678, -0.542]$ across 549 development RNAs in 404 components. Retention averages occupancies of initial heavy-atom contacts at a strict 4.5-Å cutoff with minimum-image geometry. Medians and rank associations use 2,000 whole-component bootstrap draws (seed 20260903).

For the ligand interaction comparison, 190 of 842 initial HB candidates across 207 development systems undergo persistent loss. The 134 positives retaining ordinary contact for at least 90% of frames and 132 without persistent ordinary-contact loss are overlapping sets. Temporal controls preserve candidates and state occupancy, shuffling future frames individually (20 repetitions) or in 64-frame blocks (ten repetitions; approximately 5 ns per block). The aggregate comparison is shown in the main-text coverage figure; repetition-level values remain in the source records. Their spread describes permutation variability rather than uncertainty across independent RNA systems.

D.2 Window comparisons and prediction error

Core continuous windows contain frames 0–494 and 495–990; ligand windows contain 0–499 and 500–999. RMSF is recomputed about each half's temporal mean. Event comparisons keep initial candidates fixed and use future halves 1–495/496–990 for Core and 1–499/500–999 for complexes. A loss crossing the boundary can qualify over the full window without qualifying in either half; it is not reassigned to force agreement.

Table S4 uses task-eligible masks and equal-system mean absolute flexibility or distance differences. These differ from the median rank summaries reported in the main text. The full validation distributions appear in Figure S1. Core geometric diagnostics are available for 97 test RNAs, but the primary distance benchmark and its task-matched comparison use 78. Validation and test values are retained because their variability differs, especially for ligand distances.

Each system's window discrepancy is associated with its balanced-split prediction error after averaging model seeds. Partial rank correlations remove log RNA length from both ranked variables. Neural distance errors remain associated with window differences, and structural references show similar associations (Table S5). The association with pocket-flexibility MAE is unresolved for the full-complex neural model, while full-context HGB shows a positive association. These shared-trajectory quantities characterize difficult targets without estimating a causal sampling-error contribution or a theoretical noise floor (Grossfield et al., 2018).

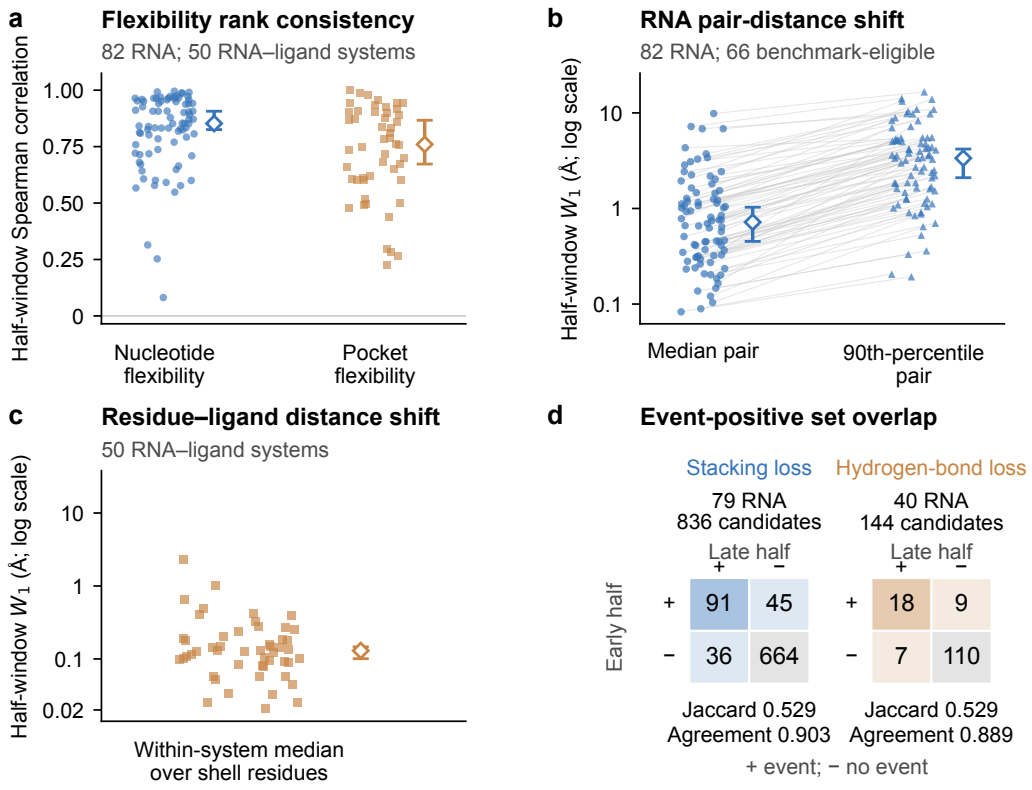

Figure S1: **Flexibility ranks persist across windows, while distance distributions and interaction losses vary.** All panels compare the first and second trajectory halves in validation. (a) Within-RNA Spearman correlations between RMSF profiles for 82 Core RNAs (55 components) and 50 ligand complexes (29 components). (b) Median and 90th-percentile half-window distance-distribution W_1 across nucleotide pairs within each Core RNA. Lines connect the two summaries for the same RNA; 66 of these 82 RNAs meet the distance-prediction eligibility criteria. (c) Median half-window residue–ligand W_1 over residues selected from the initial structure of each of 50 complexes. In (a–c), small points represent individual systems; diamonds and bars show medians across systems and 95% component-bootstrap intervals. The distance axes use a logarithmic scale and retain all observations. (d) Counts of initial interactions with or without persistent loss in each trajectory half. Jaccard measures overlap among interactions with loss; overall agreement also includes interactions without loss in either half.

Table S4: Task-matched half-window discrepancies. Entries are equal-system means and 95% component intervals; flexibility rank correlation is dimensionless and errors are in Å. The main text reports the corresponding median rank summaries, while this table retains task-matched absolute discrepancies.

Split	Observable	RNA	Groups	Mean [95% CI]
Core				
Val.	Flexibility MAE	82	55	0.620 [0.470, 0.790]
Val.	Flexibility rank correlation	82	55	0.818 [0.775, 0.857]
Val.	Distance W_1	66	42	1.550 [1.154, 1.978]
Test	Flexibility MAE	97	77	0.639 [0.497, 0.784]
Test	Flexibility rank correlation	97	77	0.802 [0.756, 0.847]
Test	Distance W_1	78	63	1.271 [1.014, 1.571]
Ligand				
Val.	Flexibility MAE	50	29	0.436 [0.232, 0.733]
Val.	Flexibility rank correlation	50	29	0.725 [0.660, 0.781]
Val.	Distance W_1	50	29	0.379 [0.261, 0.522]
Test	Flexibility MAE	49	31	0.384 [0.228, 0.661]
Test	Flexibility rank correlation	49	31	0.698 [0.624, 0.764]
Test	Distance W_1	49	31	0.484 [0.290, 0.773]

Table S5: Association of half-window discrepancy with balanced-split test error. Partial rank correlations adjust for log RNA length; intervals resample complete components.

Model	RNA	Spearman [95% CI]	Partial rank [95% CI]
Nucleotide flexibility			
EGNN	97	0.670 [0.537, 0.767]	0.659 [0.498, 0.765]
HGB	97	0.681 [0.561, 0.767]	0.673 [0.532, 0.768]
Pair distances			
EGNN	78	0.749 [0.620, 0.836]	0.723 [0.563, 0.823]
HGB	78	0.734 [0.587, 0.829]	0.709 [0.532, 0.814]
Initial point	78	0.733 [0.588, 0.827]	0.704 [0.522, 0.811]
Gaussian	78	0.721 [0.572, 0.815]	0.690 [0.506, 0.802]
Pocket flexibility			
Full-context HGB	49	0.342 [0.011, 0.610]	0.372 [0.061, 0.622]
Full-complex neural	49	0.226 [-0.155, 0.562]	0.197 [-0.185, 0.548]
Ligand distances			
Gaussian	49	0.683 [0.453, 0.832]	0.684 [0.443, 0.832]
Initial point	49	0.759 [0.541, 0.882]	0.762 [0.534, 0.880]
Full-complex neural	49	0.672 [0.428, 0.818]	0.676 [0.426, 0.818]

E Independent Simulations and Experimental Comparison

E.1 Independent simulations and model comparisons

The replica panel contains nine Core RNAs in seven components, each with four independently prepared simulations under the protocol in Appendix A.3. All 36 production records show completion, finite sampled energies/temperatures and no LINCS warnings; selected production settings and force-field parameters match Core. Different preparations and velocities contribute to replica variability. Flexibility uses all nine RNAs; the distance criteria retain six in five components.

For training members, whole-component cross-fitting excludes the target component from both fitting and checkpoint selection. Validation and test members use models fitted on the grouped training partition. Within each RNA, sampling differences average the six replica pairs; model errors average four targets and three optimization seeds. Component-bootstrap intervals preserve parent groups. Thirty task/seed cross-fits cover five training components; three distance fits have no eligible replica parent and contribute no distance score. These model fits follow the primary benchmark procedure, separately from the fixed-update coverage experiments.

Table S6: Prediction errors and independent-simulation differences under the balanced split. Entries are mean MAE or W_1 (Å), with 95% component intervals.

Comparison	Mean [95% CI]
Flexibility: nine RNAs / seven components	
Replica pairs	1.329 [0.275, 2.795]
EGNN	1.544 [0.535, 3.018]
GNM	2.355 [0.693, 4.649]
HGB	1.587 [0.452, 3.292]
Distances: six RNAs / five components	
Replica pairs	1.874 [1.532, 2.332]
EGNN	2.984 [2.347, 3.673]
HGB	3.389 [2.431, 4.374]
Initial point	3.752 [2.809, 4.680]
Calibrated Gaussian	3.456 [2.497, 4.372]

Four runs yield six replica pairs per RNA. Scores average runs and optimization seeds within RNA; whole training components are excluded from fitting and checkpoint selection.

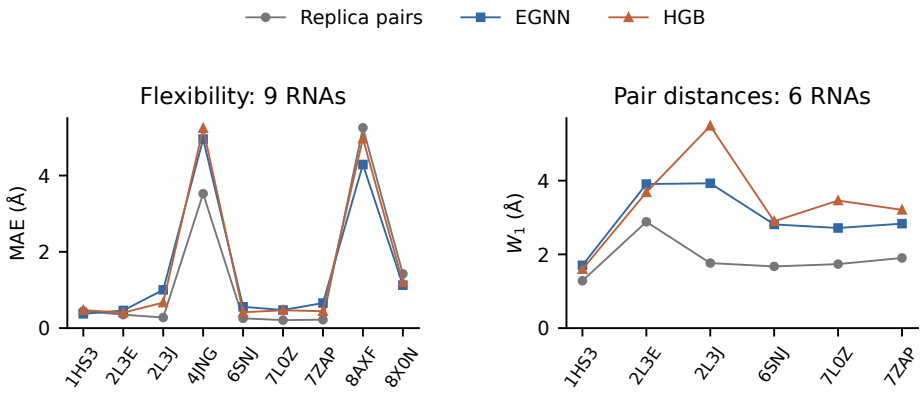

Figure S2: **Model errors and independent-simulation differences by RNA.** Grey values average six pairs of four replicas; neural and HGB values average four targets and three optimization seeds. Flexibility includes nine RNAs in seven components; distances include six RNAs in five components. Lines connect RNA identities for readability, not a temporal sequence.

EGNN's distance gain over the Gaussian is 0.472 Å [0.050, 0.815]; its 0.405-Å gain over HGB is unresolved [−0.014, 1.038]. Figure S2 shows the construct-level variation. EGNN beats HGB on all four distance replicas of 2L3J, 6SNJ, 7L0Z and 7ZAP, but on one of four for 1HS3 and two of four for 2L3E. HGB also retains higher average flexibility rank correlation (0.533 versus 0.487). This small panel establishes neither a universal model ordering nor a theoretical noise floor.

E.2 Stacking candidates and event reproducibility

The same 36 trajectories support a label-only stacking analysis under the primary candidate, geometry and 1-ns persistence rules. Initial-candidate sets have mean pairwise Jaccard 0.426 [0.250, 0.618] across nine RNAs. Their retrospective four-run intersection retains 41 candidates in six RNAs/five components, or 164/384 natural run-candidate instances (42.7%); three RNAs have empty intersections (Table S7).

Six pairwise label comparisons are averaged within each RNA before averaging RNAs. Mean agreement is 0.878 [0.817, 0.942], partly reflecting negatives; the positive fraction is 0.119 [0.066, 0.164]. Candidate Jaccard omits pairs with an empty candidate union. Positive-event Jaccard omits empty positive unions, so all-negative 6SNJ contributes to agreement but not that metric. Positive-event Jaccard averages 0.360 [0.080, 0.720] over five RNAs/five components. Empty-set rules were fixed before analysis; intervals use 2,000 component draws (seed 20260914). The shared intersection is a retrospective comparison with reduced coverage, not a prospective task-entry rule or a model generalization test.

Table S7: Persistent stacking loss across independent Core replicas. Natural candidates are the range across four independently prepared runs; common candidates are their four-way intersection. Event statistics use this retrospective intersection. "–" denotes an undefined quantity, not zero. Jaccard omits pair comparisons with no positive event in either run.

RNA	Natural candidates	Common candidates	Initial-set Jaccard	Loss fraction	Event Jaccard	Defined event pairs	Agreement
1HS3	2–3	2	0.833	0.125	0.000	3	0.750
2L3E	8–13	5	0.534	0.100	0.200	5	0.867
2L3J	26–31	19	0.669	0.053	1.000	6	1.000
4JNG	6–10	0	0.136	–	–	0	–
6SNJ	5–9	1	0.383	0.000	–	0	1.000
7L0Z	19–22	12	0.608	0.062	0.100	5	0.903
7ZAP	5–9	2	0.479	0.375	0.500	6	0.750
8AXF	8–11	0	0.191	–	–	0	–
8X0N	0–1	0	0.000	–	–	0	–

E.3 UUCG and structural references

The 14-nt UUCG construct `GGCACUUCGGUGCC` is 2KOC in Core training. Dataset A of Bottaro et al. (2020) supplies 63 active bidirectional eNOE observations. For each observation, equal-weight structures give $d_k = \langle r_k^{-6} \rangle^{-1/6}$. We report distance MAE and $63^{-1} \sum_k [(d_k - d_k^{\mathrm{exp}})/(0.1 d_k^{\mathrm{exp}})]^2$; the latter is a mean scaled residual, not a degrees-of-freedom-adjusted chi-square.

Trajectory averaging improves on the accompanying MD snapshot, while both source NMR references agree better with the observations. Hydrogen preparation matters: the actual hydrogenated starting structure scores 5.523 despite a heavy-atom RMSD of only 0.0051 Å from source model 1 (same-name hydrogen RMSD 0.738 Å). The heavy-atom-only PDB cannot directly support these proton distances. The MD result uses all 991 frames over 0–99 ns. All rows concern one training RNA and do not establish independent-system or universal experimental agreement.

Table S8: UUCG comparison against the same 63 eNOE observations. Ensemble distances use equal-weight r^{-6} averaging. Lower error is better.

Structural reference	Structures	Normalized MSE	MAE (Å)
Source NMR model 1	1	1.571	0.395
Source NMR ensemble	20	1.130	0.347
Hydrogenated prepared structure	1	5.523	0.655
MD snapshot	1	5.590	0.633
MD ensemble, 0–99 ns	991	4.674	0.552

The normalized score is a mean squared scaled residual, not a degrees-of-freedom-adjusted chi-square. All references represent one RNA construct. The MD snapshot is the accompanying GRO coordinate reference, not a verified production frame zero.

F Models and Benchmark Results

F.1 Model inputs and training

Core continuous tasks. The Core encoder uses 13 nucleotide features, four message-passing layers of width 128, 16 radial basis features, and dropout 0.1. This fixed-coordinate EGNN variant (Satorras et al., 2021) uses invariant distance messages; the heterogeneous EGNN used for hydrogen-bond disruption updates coordinates. The flexibility head predicts a nonnegative residue value. The pair-distance head combines symmetric residue representations with 17 pair features and predicts a displacement from the source distance and a Gaussian width. Its output consists of 128 deterministic Gaussian quantiles, clipped at zero. Training uses 128 target quantiles; evaluation compares the prediction with the complete empirical distance trace using exact W_1.

Both continuous Core models use AdamW with learning rate 3×10^{-4}, gradient clipping at 5, at most 40 epochs, a minimum checkpoint epoch of 12, and patience 8. Flexibility uses Huber loss with threshold 1 Å, batches of eight systems, and weight decay 10^{-5}; checkpoint selection maximizes validation Spearman correlation. Pair distances use quantile L_1 loss, batches of four, and weight decay 10^{-4}; selection minimizes validation W_1. Losses weight systems equally. The distance learner and its calibrated references use 257 training systems in Verified; the validation population is all 66 eligible systems. Test results include all 78 eligible RNAs, with 46 in Verified and 32 in Challenge.

Geometry HGB uses learning rate 0.06, 160 iterations, at most 31 leaves, a minimum of 25 samples per leaf, L_2 regularization 0.1, and no early stopping. Sample weights give each RNA equal total weight. For pair distances, HGB predicts a mean residual and log width. The fixed-width reference retains the source distance as its center and selects a common width from a training-only grid spanning 0–20 Å. GNM (Yang et al., 2006) uses a 10 Å contact cutoff and pseudoinverse tolerance 10^{-8}, with a nonnegative affine calibration fitted on training data. Its correlations are undefined for one test RNA, leaving 96 systems from 76 components for that metric.

Stacking-loss models. The 18 methods cover four simple references, logistic regression, HGB, a pairwise MLP on 62 geometric features, invariant graph encoders, an RNA-FM probe, and scalar–vector variants. The base GNN combines a four-layer, width-128 encoder with a symmetric pair readout; the rich-feature GNN adds the 62 geometric features to that readout. The RNA-FM model trains a pairwise probe on frozen sequence representations (Chen et al., 2022). Scalar–vector models use three message-passing layers of width 24, and the pair-readout variants omit those layers. Residual variants learn an additive correction to a training-fitted logistic score. Multi-task variants also learn stack formation; variants marked auxiliary learn the future nonstacked fraction. These targets supply additional training supervision and are not separate benchmark endpoints.

Stacking models retain the fixed training settings: learning rate 3×10^{-4}, weight decay 10^{-5}, at most 80 epochs, a minimum checkpoint epoch of 12, patience 8, and minimum validation improvement 0.0005. Validation MacroAP selects the checkpoint. Feature scaling, logistic offsets, and all learned weights are fitted on the training cohort. The event population contains 530 training and 79 validation systems.

Hydrogen-bond-loss models. The 15 methods range from a prevalence prior and initial geometry to chemical feature learners, separate RNA-only and ligand-only encoders, and joint-complex networks. The chemistry GNN uses ligand chemical features and typed interface messages. Chemistry + HB GNN additionally supplies 46 initial-complex features to its prediction head. The heterogeneous EGNN and SV models represent RNA, ligand, and interface relations in a joint graph with three layers; hidden widths are 64 for EGNN and 32 for SV. A pairwise MLP uses the same 46 features. These models learn from 167 training systems with 698 candidates and 156 positive events; validation contains 40 systems, 144 candidates, and 34 positives.

Neural training uses AdamW, learning rate 3×10^{-4}, weight decay 10^{-5}, gradient clipping at 5, batches of at most eight systems, and system-equal binary cross-entropy without class upweighting. The maximum is 60 epochs, the minimum checkpoint epoch is 12, and early-stopping patience is 12. Validation MacroAP selects the checkpoint. Feature scaling uses training systems, and the classical methods retain their fixed development settings. Runs include CUDA and CPU restarts under the same model/optimizer settings; seed SD therefore also reflects the recorded execution conditions.

Continuous RNA–ligand models. The full-complex model combines the four-layer, width-128 RNA encoder with a ligand graph encoder, messages from ligand geometry to RNA residues, and a pooled ligand representation. The RNA-only model omits ligand inputs. Both use AdamW with learning rate 3×10^{-4}, weight decay 10^{-5}, gradient clipping at 5, at most 40 epochs, a minimum checkpoint epoch of 12, and patience 8. System-equal losses jointly supervise pocket flexibility, contact occupancy, and residue–ligand distances, with target scales of 1 Å, 0.05, and 2 Å, respectively. Checkpoint selection uses the mean fractional validation improvement over full-context HGB across those three training objectives. The auxiliary occupancy objective is retained when these checkpoints are evaluated on the two continuous benchmark tasks. The primary continuous checkpoints retain this joint selection; the separately refitted controls use the selection criteria stated in Appendix G.1.

Core-to-complex transfer. Pretraining uses 492 retained Core training systems in 375 components after excluding entire groups related to any of the 99 ligand validation/test complexes by source PDB, exact sequence, or local sequence identity and bidirectional coverage of at least 0.8. Relation checks retain transitive links through the original component registry. Pretraining runs for 40 epochs with 19 optimizer updates per epoch, using a fixed final checkpoint; each system contributes once per epoch. The shared RNA encoder is then fine-tuned on the same 210 training complexes and joint validation rule as the from-scratch full-complex model. Each of the three seeds covers both source pretraining and task adaptation.

F.2 Aggregation and uncertainty

For a defined system metric s_{ik} from system i and training seed k, we first compute $\bar{s}_i = K^{-1} \sum_{k=1}^{K} s_{ik}$ and report the equal-system mean. The standard deviation summarizes the $K = 3$ separately evaluated seed means, using seeds 20260904, 20260905, and 20260906. Predictions are not ensembled before scoring. Deterministic references and the feature-based fits return identical scores across seed slots; tables report their scores once without assigning stochastic uncertainty to duplicate fits.

Confidence intervals resample complete RNA components with replacement 2,000 times, using random seed 20260903. Each selected component contributes all its systems, including their multiplicity when the component is sampled more than once. Paired intervals apply the same procedure to within-system score differences. Higher Spearman correlation or MacroAP and lower MAE or W_1 indicate improvement. The predefined comparisons are reported in Table S11; the additional continuous ligand comparisons appear in Table S13. These are individual 95% intervals without a familywise adjustment.

For an event task, MacroAP averages AP over all candidate-bearing systems, assigning zero to systems with no positive event. This convention does not penalize false positives within a no-positive system. Its maximum on a fixed test set is therefore N_+/N, where N_+ counts systems with positives. Positive-system AP averages only over those N_+ systems; pooled AP ranks all candidates together, and system-weighted AP ranks them together with each system given equal total weight. These answer different questions and remain separate metrics. AUROC is undefined within single-

class systems and is averaged only where defined. A bootstrap sample can change N_+/N, so the event confidence intervals are not clipped to the original test set's maximum.

All reported fitted methods use the grouped training and validation partitions defined in Appendix B. Their fitted parameters and selected checkpoints correspond to those partitions; test systems do not enter model fitting or checkpoint selection.

External RNA flexibility baseline. DeepRMSF-Open-Preprocess uses its published fold-5 checkpoint and the open preprocessing implementation on the balanced Core test manifest. Predictions were frozen before test-label evaluation; 97/97 systems have native-coverage scores. The mean within-RNA Spearman correlation is 0.550 [0.491, 0.605] and top-20% AUPRC is 0.702 [0.663, 0.745], with 2,000 component-bootstrap draws over 77 components. Validation correlation is 0.553 [0.508, 0.599] on 82 systems. This is an external transfer reference, not a three-seed model trained on RiboMotion; pretraining/source overlap has not been independently audited. Its normalized output has no directly comparable RMSF MAE in angstroms.

F.3 Core and hydrogen-bond benchmark results

Tables S9 and S10 give every method's primary score, seed SD and component interval. The Core distance strata contain 46 Verified and 32 Challenge RNAs: EGNN/HGB errors are 2.541/2.777 Å and 5.587/5.715 Å, respectively. HGB improves over the calibrated Gaussian on all 78 eligible RNAs by 0.158 Å [0.052, 0.274]. The 39 HB systems include 20 with positives and 16 with both classes. Table S11 reports the predefined paired comparisons. Seedwise and secondary metrics remain in the accompanying numerical records.

Table S9: Complete Core results. Flexibility covers 97 RNAs; distances 78; stacking 93. Undefined correlations are excluded only from correlation summaries. Verified and Challenge contain 46 and 32 RNAs, respectively; their components can overlap.

Method	Mean ± SD	Component 95% CI	RNA / comp.
Nucleotide flexibility: Spearman			
GNM	0.294	[0.210, 0.377]	96 / 76
Geometry HGB	0.636	[0.594, 0.676]	97 / 77
Core EGNN	0.603 ± 0.018	[0.549, 0.654]	97 / 77
Pair distances: All			
Initial point	4.378	[3.750, 5.123]	78 / 63
Calibrated Gaussian	4.140	[3.522, 4.869]	78 / 63
Geometry HGB	3.982	[3.347, 4.727]	78 / 63
Core EGNN	3.791 ± 0.027	[3.180, 4.533]	78 / 63
Pair distances: Verified			
Initial point	3.097	[2.666, 3.599]	46 / 39
Calibrated Gaussian	2.895	[2.470, 3.391]	46 / 39
Geometry HGB	2.777	[2.364, 3.245]	46 / 39
Core EGNN	2.541 ± 0.020	[2.141, 3.004]	46 / 39
Pair distances: Challenge			
Initial point	6.218	[5.163, 7.329]	32 / 29
Calibrated Gaussian	5.929	[4.887, 7.020]	32 / 29
Geometry HGB	5.715	[4.673, 6.881]	32 / 29
Core EGNN	5.587 ± 0.046	[4.555, 6.708]	32 / 29
Stacking loss: MacroAP			
Prevalence	0.265	[0.210, 0.329]	93 / 75
Initial distance	0.420	[0.352, 0.487]	93 / 75
Stacking margin	0.449	[0.378, 0.526]	93 / 75
Plane angle	0.468	[0.402, 0.537]	93 / 75
Logistic regression	0.573	[0.501, 0.645]	93 / 75
Geometry HGB	0.559	[0.488, 0.633]	93 / 75
Geometry MLP	0.587 ± 0.010	[0.516, 0.657]	93 / 75

Method	Mean ± SD	Component 95% CI	RNA / comp.
Rich-feature GNN	0.581 ± 0.007	$[0.512, 0.652]$	93 / 75
Base GNN	0.527 ± 0.011	$[0.449, 0.608]$	93 / 75
RNA-FM GNN	0.452 ± 0.003	$[0.381, 0.530]$	93 / 75
Pair-readout GNN	0.558 ± 0.009	$[0.491, 0.628]$	93 / 75
Scalar–vector	0.578 ± 0.005	$[0.506, 0.653]$	93 / 75
Residual pair-readout GNN	0.573 ± 0.001	$[0.501, 0.645]$	93 / 75
SV residual	0.578 ± 0.007	$[0.506, 0.646]$	93 / 75
Base GNN + formation	0.516 ± 0.007	$[0.440, 0.598]$	93 / 75
Static-feature GNN + formation	0.579 ± 0.015	$[0.509, 0.654]$	93 / 75
SV + continuous auxiliary	0.581 ± 0.008	$[0.511, 0.659]$	93 / 75
SV residual + auxiliary	0.574 ± 0.001	$[0.503, 0.647]$	93 / 75

Table S10: All hydrogen-bond-loss methods on the same 39 test complexes in 26 components. Twenty systems contain positive events. Means average separately evaluated seeds.

Method	Mean ± SD	Component 95% CI	RNA / comp.
Hydrogen-bond loss: MacroAP			
Chemistry + HB GNN	0.437 ± 0.014	$[0.299, 0.596]$	39 / 26
Geometry HGB	0.437	$[0.295, 0.598]$	39 / 26
Heterogeneous SV	0.436 ± 0.014	$[0.297, 0.592]$	39 / 26
Geometry MLP	0.431 ± 0.003	$[0.293, 0.587]$	39 / 26
Heterogeneous EGNN	0.429 ± 0.001	$[0.289, 0.585]$	39 / 26
Chemistry GNN	0.429 ± 0.006	$[0.292, 0.588]$	39 / 26
Full-context logistic	0.425	$[0.284, 0.575]$	39 / 26
Full-context HGB	0.420	$[0.283, 0.572]$	39 / 26
Geometry logistic	0.420	$[0.280, 0.572]$	39 / 26
Initial HB margin	0.420	$[0.285, 0.575]$	39 / 26
Base complex GNN	0.416 ± 0.004	$[0.285, 0.570]$	39 / 26
RNA-only GNN	0.409 ± 0.006	$[0.270, 0.567]$	39 / 26
Initial distance	0.397	$[0.266, 0.540]$	39 / 26
Training prior	0.313	$[0.202, 0.445]$	39 / 26
Ligand-only GNN	0.313	$[0.202, 0.445]$	39 / 26

Table S11: Predefined Core and HB comparisons. Positive gains favor the first method; intervals resample paired component-level clusters after within-system seed averaging.

Comparison	Gain	95% CI	RNA / comp.
Core EGNN vs. Geometry HGB	-0.033	$[-0.062, -0.003]$	97 / 77
Core EGNN vs. Geometry HGB	0.192	$[0.086, 0.293]$	78 / 63
Geometry HGB vs. Calibrated Gaussian	0.158	$[0.052, 0.274]$	78 / 63
Base GNN vs. Logistic regression	-0.046	$[-0.079, -0.011]$	93 / 75
SV residual vs. Logistic regression	0.004	$[-0.008, 0.017]$	93 / 75
Chemistry + HB vs. full HGB	0.017	$[-0.017, 0.057]$	39 / 26

F.4 RNA–LIGAND CONTINUOUS PREDICTIONS

Both continuous tasks evaluate all 49 test complexes in 31 components, with 531 pocket and 971 shell residues selected from the first exported complex. Distance error compares 1,000 nonnegative predicted samples with the full 1,000-frame target by exact W_1, averaging shell residues and then systems. The candidate masks are fixed by the input complex. Table S12 includes the from-scratch, pretrained and structural references; Table S13 gives the predefined static-reference comparisons.

Core pretraining followed by complex fine-tuning gives pocket MAE 0.726 Å, versus 0.729 from scratch; its paired gain is 0.003 Å $[-0.018, 0.031]$. Pocket rank correlation increases by 0.049 $[-0.008, 0.104]$, while distance error increases by 0.031 Å [0.017, 0.046]. Pretraining therefore does not yield a uniform transfer benefit under this architecture and joint objective. All three source fits use the relation-filtered 492-RNA training population. Source-batch constant profiles have undefined within-pocket correlation.

Table S12: RNA–ligand continuous predictions on 49 test complexes in 31 components. Core transfer uses pretraining on 492 relation-filtered Core training RNAs followed by fine-tuning on the same complexes. Errors are in Å; seed SDs refer to separately scored models.

Method	Pocket MAE	Pocket ρ	Distance W_1
Prior / Gaussian	0.900	0.039	0.736
Initial point	—	—	0.817
Reduced-context HGB	0.815	0.317	1.437
Full-context HGB	0.780	0.358	1.032
RNA-only neural	0.775 ± 0.008	0.334 ± 0.006	2.351 ± 0.060
Full-complex neural	0.729 ± 0.008	0.313 ± 0.022	0.794 ± 0.020
Source-batch prior	0.879	—	2.306
Core transfer	0.726 ± 0.017	0.362 ± 0.044	0.825 ± 0.034

Table S13: Full-complex neural gains relative to static references on the same 49 test complexes. Positive values indicate lower error (Å).

Target	Reference	Gain	95% CI
Pocket MAE	Reduced HGB	0.086	$[-0.001, 0.208]$
Distance W_1	Gaussian	-0.058	$[-0.081, -0.036]$
Distance W_1	Full HGB	0.238	$[0.103, 0.387]$

F.5 Distance evaluation without consistency screening

Technical eligibility requires mapped same-chain pairs, finite features and targets, and audited 991-frame coordinates with valid boxes. Omitting the trajectory-consistency entry screen admits all 97 registered test RNAs in 77 components: 78 primary RNAs in 63 components and 19 additional RNAs in 18 components. Component counts overlap between strata. The same frozen training, calibration and validation-selected models are used throughout. Primary mean scores replay within 10^{-7} Å; additional inference makes no parameter changes.

Table S14: Core distance evaluation with frozen models under the primary and expanded test populations. Values are exact W_1 in Å. Components can appear in both primary and additional strata.

Population	Method	W_1	95% CI
Primary (78)	Initial point	4.378	$[3.691, 5.074]$
Primary (78)	Calibrated Gaussian	4.140	$[3.464, 4.820]$
Primary (78)	Geometry HGB	3.982	$[3.314, 4.685]$
Primary (78)	Core EGNN	3.791 ± 0.027	$[3.133, 4.478]$
Additional (19)	Initial point	6.402	$[4.686, 8.165]$
Additional (19)	Calibrated Gaussian	6.105	$[4.378, 7.890]$
Additional (19)	Geometry HGB	5.631	$[4.106, 7.242]$
Additional (19)	Core EGNN	5.620 ± 0.143	$[4.081, 7.287]$
All technical (97)	Initial point	4.774	$[4.136, 5.486]$
All technical (97)	Calibrated Gaussian	4.525	$[3.897, 5.235]$
All technical (97)	Geometry HGB	4.305	$[3.697, 4.987]$
All technical (97)	Core EGNN	4.149 ± 0.042	$[3.533, 4.845]$

Across all 97 systems, EGNN improves over HGB by 0.156 Å [0.055, 0.282] and over the Gaussian by 0.376 Å [0.281, 0.472]. On the additional 19, its HGB gain is 0.011 Å [−0.164, 0.255], whereas its Gaussian gain is 0.485 Å [0.133, 0.846]. The additional-minus-primary gain contrasts are −0.181 Å [−0.373, 0.049] for HGB and 0.136 Å [−0.232, 0.530] for the Gaussian, resampling shared components jointly. Broader coverage preserves the overall neural advantage, without establishing a distinct effect of screening on model gains.

F.6 Complementary hydrogen-bond metrics

Four fixed methods are compared in all six pairs. System-weighted AP pools candidates with weight $1/n_i$ per candidate in system i; weighted Brier averages squared probability errors with the same weights. AP groups tied thresholds. Scores are computed for each seed before averaging, and each of 2,000 bootstrap draws applies identical component multiplicities across methods (seed 20260914).

Table S15: Complementary HB metrics on 39 systems and 139 candidates. Higher system-weighted AP and lower Brier are better.

Method	Weighted AP	Weighted Brier
Geometry HGB	0.473	0.218
Full-context HGB	0.575	0.195
Chemistry + HB GNN	0.528 ± 0.012	0.196 ± 0.004
Heterogeneous SV	0.446 ± 0.010	0.225 ± 0.024

Full-context HGB and Chemistry + HB GNN outperform heterogeneous SV on both secondary metrics; the corresponding four individual paired intervals exclude zero. Their differences from geometry HGB remain unresolved. The primary MacroAP ranking and the pooled probability metrics thus capture different aspects of event prediction. All twelve comparisons are retained in Table S16; no model fitting, probability ensembling or recalibration is performed here.

Table S16: All twelve paired HB secondary comparisons. Positive gains favor the first displayed method. Intervals are individual 95% component intervals, without familywise adjustment.

Comparison	Metric	Gain	95% CI
Full HGB vs. Geo. HGB	AP	0.102	$[-0.036, 0.239]$
Full HGB vs. Geo. HGB	Brier	0.023	$[-0.015, 0.063]$
Chem. HB vs. Geo. HGB	AP	0.055	$[-0.103, 0.247]$
Chem. HB vs. Geo. HGB	Brier	0.022	$[-0.016, 0.063]$
SV vs. Geo. HGB	AP	-0.026	$[-0.151, 0.121]$
SV vs. Geo. HGB	Brier	-0.007	$[-0.045, 0.034]$
Chem. HB vs. Full HGB	AP	-0.047	$[-0.172, 0.116]$
Chem. HB vs. Full HGB	Brier	-0.001	$[-0.031, 0.027]$
SV vs. Full HGB	AP	-0.129	$[-0.233, -0.007]$
SV vs. Full HGB	Brier	-0.030	$[-0.060, -0.001]$
SV vs. Chem. HB	AP	-0.082	$[-0.178, -0.016]$
SV vs. Chem. HB	Brier	-0.029	$[-0.049, -0.010]$

G Training Coverage and Distance Parameterization

G.1 Training populations, budgets, and uncertainty

The controlled experiments use the same grouped partitions, input structures, labels, and task masks. They comprise 75 fixed-budget neural fits: 33 monomer distance fits for coverage and data-size comparisons, nine length-matched fits, 21 ligand-pocket data-size fits, and 12 ligand-distance parameterization fits. These are controlled retrainings of the existing architectures, distinct from the primary benchmark checkpoints in Table 3. All comparisons use the released trajectories.

Monomer distance fits use 4,400 optimizer updates with four systems per update (17,600 system visits). Smaller populations are sampled repeatedly so that data reduction does not reduce the update budget. Each ligand fit uses 8,400 updates with one system per update. Validation is evaluated at 40 intervals; the best checkpoint is selected from intervals 12–40 without early stopping. All monomer distance fits select on the same 66 eligible validation RNAs. The ligand data-size experiment selects by pocket MAE, whereas all four distance-parameterization conditions select by distance W_1 on the same 50 validation complexes. Optimization seeds are 20260904–20260906 and subset-order seeds are 20260914–20260916. Optimizer and encoder settings otherwise follow Appendix F.1.

Data-size subsets retain complete components in three nested random orderings. For monomers, 25% retains 82 components and 141/132/115 RNAs; 50% retains 163 components and 242/224/239 RNAs; all eligible data comprise 326 components and 441 RNAs. For complexes, the respective counts are 13 components and 51/58/67 systems, 26 components and 92/99/141 systems, and 51 components and 210 systems. The full-data condition is shared by the subset orderings and has three distinct seed fits, not nine independent fits.

Scores first average optimization seeds within each held-out RNA, then subset orderings, and finally RNAs with equal weight. Component-only intervals use 2,000 bootstrap draws of whole held-out components. The training-randomness sensitivity analysis also resamples the global optimization seeds and, where applicable, nested subset orderings. The same resampled seeds and orderings are used on both sides of each paired comparison, and a shared full-data fit remains shared. Intervals describe uncertainty in these comparisons; residues, frames, or repeated uses of a full-data fit do not become additional independent test systems. Tables S17–S20 summarize these comparisons; all 33 sensitivity contrasts remain in the numerical records.

G.2 Sampling-consistency strata and length matching

Verified denotes the predefined distance-sampling criteria in Appendix C. Its 257 training RNAs span 199 components. Three approximately equal-count samples from All-eligible contain 257/261/257 RNAs, with 176/185/179 components. Three further All-eligible samples each contain exactly 257 RNAs and match Verified's counts in the prespecified length bins < 25, 25–49, and ≥ 50 nucleotides: 101, 103, and 53. These retain whole components (176/194/211 per sample), controlling system count and coarse length distribution rather than every aspect of molecular composition.

Verified-only EGNN reaches 3.766 Å, compared with 3.871 for approximately equal-count All-eligible, 3.847 for length-matched All-eligible, and 3.826 for the complete population. The Verified advantage over equal-count data is resolved under component-only resampling, but not when the three training seeds and subset orders are also resampled. All other reported population contrasts are likewise unresolved under the latter analysis. HGB does not show a resolved population benefit. Length matching controls a coarse composition difference but does not establish a causal source of difficulty.

The Core neural 25%-to-full gain is 0.097 Å [0.001, 0.177] with training variation included, whereas the corresponding HGB gain of 0.128 Å has interval $[-0.002, 0.270]$. The pocket-neural gain is 0.018 Å $[-0.033, 0.070]$. Distance error from the pocket-selected neural checkpoints improves by 0.143 Å [0.083, 0.235]; this secondary output is retained in the numerical records rather than substituted for the prespecified pocket endpoint.

G.3 Initial-distance residuals and auxiliary supervision

The ligand experiment crosses two Gaussian-center parameterizations (absolute or initial-distance residual) with two objectives (the original joint objective or distance-only supervision). Both centers receive identical initial-distance information, are nonnegative, and use a predicted width greater than 0.02 Å. All conditions output the same 1,000 clipped Gaussian quantile samples. Distance-only training removes flexibility and occupancy auxiliary losses while retaining the distance coefficient of $1/6$, so the comparison does not also change the scale of the primary loss. There are three optimization seeds in each condition.

Table S17: Monomer training-population controls. Test distance W_1 (Å) with 95% component intervals on 78 RNAs in 63 components.

Training population	n_{train}	EGNN	HGB
Verified	257	3.766 [3.128, 4.497]	3.982 [3.353, 4.694]
All, approximate equal count	257–261	3.871 [3.243, 4.603]	4.054 [3.446, 4.743]
All, length matched	257	3.847 [3.221, 4.589]	4.061 [3.462, 4.751]
All eligible	441	3.826 [3.200, 4.570]	4.029 [3.434, 4.705]

Equal-count and length-matched arms average three complete-component subsets. Neural fits use three optimization seeds. The initial-point reference is unchanged; Gaussian calibration and HGB are refitted within each training subset.

Table S18: Monomer test distance errors by length and sampling-consistency stratum. Entries are equal-system mean W_1 (Å); columns identify training populations.

Test stratum	n	Verified	Equal count	Length matched	All
EGNN / <25 nt	35	2.812	3.002	2.958	2.968
EGNN / 25–49 nt	25	3.428	3.493	3.475	3.446
EGNN / 50–99 nt	18	6.091	6.086	6.092	6.022
EGNN / Verified	46	2.507	2.634	2.595	2.561
EGNN / Challenge	32	5.576	5.649	5.647	5.643
HGB / <25 nt	35	2.972	3.019	3.035	2.973
HGB / 25–49 nt	25	3.607	3.772	3.734	3.726
HGB / 50–99 nt	18	6.467	6.458	6.510	6.505
HGB / Verified	46	2.777	2.906	2.918	2.911
HGB / Challenge	32	5.715	5.704	5.704	5.637

Length and consistency strata overlap and are not additional independent evaluations.

Table S19: Fixed-budget learning curves. Entries are test mean errors (Å) with component intervals: distance W_1 for 78 monomers and pocket MAE for 49 complexes.

Model	25%	50%	100%
Core EGNN	3.923 [3.308, 4.650]	3.897 [3.270, 4.632]	3.826 [3.200, 4.570]
Core HGB	4.157 [3.537, 4.867]	4.053 [3.450, 4.745]	4.029 [3.434, 4.705]
Pocket neural	0.728 [0.440, 1.275]	0.741 [0.443, 1.293]	0.709 [0.418, 1.263]
Full-context HGB	0.795 [0.490, 1.370]	0.779 [0.471, 1.362]	0.780 [0.466, 1.363]
Reduced-context HGB	0.753 [0.458, 1.325]	0.776 [0.471, 1.349]	0.815 [0.488, 1.383]
Pocket prior	0.864 [0.569, 1.387]	0.835 [0.543, 1.363]	0.890 [0.599, 1.384]

Fractions refer to complete training components. Static models are fitted to each subset. Pocket neural checkpoints select pocket MAE. Full-data fits are shared across subset orderings.

Table S20: Coverage, full-range learning and parameterization contrasts. Positive gains favor the second condition. Intervals resample paired test components, optimization seeds and subset orders; component-only intervals and all contrasts remain in the numerical records.

Model / target	Change	Gain	95% sensitivity CI
Core Neural / distance	Verified → Equal count	-0.105	[-0.251, 0.020]
Core Neural / distance	Verified → Length matched	-0.081	[-0.251, 0.072]
Core Neural / distance	Verified → All	-0.060	[-0.230, 0.069]
Core Neural / distance	Equal count → All	0.045	[-0.049, 0.127]
Core Neural / distance	25% → All	0.097	[0.001, 0.177]
Core HGB / distance	Verified → Equal count	-0.072	[-0.211, 0.066]
Core HGB / distance	Verified → Length matched	-0.079	[-0.184, 0.040]
Core HGB / distance	Verified → All	-0.047	[-0.150, 0.076]
Core HGB / distance	Equal count → All	0.025	[-0.055, 0.119]
Core HGB / distance	25% → All	0.128	[-0.002, 0.270]
Ligand Neural / distance	Absolute joint → Residual joint	0.068	[0.041, 0.095]
Ligand Neural / pocket	Absolute joint → Residual joint	-0.000	[-0.056, 0.056]
Ligand Neural / distance	Absolute distance → Residual distance	0.057	[0.034, 0.080]
Ligand Neural / distance	Absolute joint → Absolute distance	0.006	[-0.012, 0.025]
Ligand Neural / distance	Residual joint → Residual distance	-0.005	[-0.013, 0.003]
Ligand Neural / distance	25% → 100%	0.143	[0.083, 0.235]
Ligand Neural / pocket	25% → 100%	0.018	[-0.033, 0.070]
Ligand Full HGB / pocket	25% → 100%	0.015	[-0.046, 0.089]

Residual centers improve distance prediction under both joint and distance-only supervision, by 0.068 Å [0.041, 0.095] and 0.057 Å [0.034, 0.080], respectively, when training variation is included. Removing auxiliary supervision does not resolve a further distance gain within either parameterization. The two jointly trained models have virtually identical pocket MAE (0.726 Å); their paired pocket gain is -0.0002 Å [-0.056, 0.056]. Pocket outputs from distance-only models are unsupervised and have much higher errors (Table S21).

Table S21: Ligand parameterization controls, selected by validation distance error. Means average three seeds and the same 49 test complexes.

Center, supervision	Distance W_1 (Å)	Pocket MAE (Å)	Pocket ρ
Absolute, joint	0.774	0.726	0.371
Absolute, distance only	0.769	1.328	-0.125
Residual, joint	0.706	0.726	0.361
Residual, distance only	0.712	1.378	0.097

Distance-only models receive no pocket supervision. These distance-selected checkpoints differ from the pocket-selected learning curves.

G.4 Distance-distribution parameterization

To assess the expressiveness of the output family, we fit Gaussian widths with either a fixed initial-distance center or a free center and width directly to each trajectory-derived distribution. The monomer and ligand fits use their respective model clipping rules and quantile grids (128 and 1,000 samples). Two initializations and 500 optimization steps are used for each target fit, retaining the lower exact W_1. Independent scalar refinements on three predefined objects per system check the numerical results; the largest additional reductions are less than 5×10^{-7} Å for monomers and 6×10^{-8} Å for complexes. These are numerical approximations, not certified global optima.

On the test sets, freeing the Gaussian center and width reduces mean target-fit error to 0.600 Å for monomers and 0.122 Å for complexes (Table S22). Fitting only the width at the initial center leaves substantially larger errors, 3.824 and 0.616 Å, respectively. The family can therefore represent the target distributions more closely than the current predictors achieve. Inferring these parameters from

a single structure remains the predictive challenge: target-fitted scores use the answer trajectory and are neither prediction baselines nor estimates of irreducible sampling error.

Table S22: Prediction references and target-fitted Gaussian approximations on the balanced split. Entries are mean test distance W_1 (Å) with 95% component-bootstrap intervals.

Distribution	Monomers	Complexes
Initial-distance point	4.378 [3.712, 5.107]	0.817 [0.611, 1.131]
Training-calibrated Gaussian	4.140 [3.489, 4.853]	0.736 [0.534, 1.046]
Target-fit width, initial center	3.824 [3.227, 4.491]	0.616 [0.466, 0.833]
Target-fit center and width	0.600 [0.478, 0.743]	0.122 [0.072, 0.194]

Target fits use the answer trajectory and describe approximation capacity, not predictive performance. Test populations are 78/63 and 49/31 systems/components.

H MD-Supervised Conformational Ensemble Generation

H.1 MD adaptation for conformation generation

This experiment starts from the external `protenix_base_default_v1.0.0` checkpoint (Zhang et al., 2026). RNA multiple-sequence-alignment (MSA) and template inputs are disabled in all four arms, and the same sequence-derived representations are used. We freeze the representation network and adapt only the diffusion module; the multi-frame arm uses 32 uniformly spaced MD frames per training RNA. Of the 551 assigned Core training RNAs, 550 pass the full-atom adaptation checks; one fails the covalent-connectivity criterion. This task-specific exclusion leaves the dataset assignment and the six-task benchmarks unchanged.

AdamW uses learning rate 10^{-4}, weight decay 0.01, $\beta = (0.9, 0.999)$, gradient clipping at 1.0, and 1,102 warmup updates. Each of 27,550 updates samples one RNA uniformly with replacement and uses four diffusion samples. Optimization uses one seed (20260916), bfloat16, and the inherited diffusion, bond and smooth local distance difference test (lDDT) loss weights (4, 1 and 1). The final checkpoint is fixed in advance; validation and test scores do not select it.

The static-PDB control uses the unchanged source coordinates for each RNA, including 11 documented chain-geometry exceptions. The single-frame control uses MD frame 495 for each RNA. The multi-frame arm draws among 32 uniformly spaced MD frames. All three controls share the pretrained weights, 550 training systems, frozen representation, optimizer and update count, checkpoint rule, and test sampling protocol; only the target-coordinate policy changes. The single-frame and static controls do not receive 32 distinct target structures. The corresponding training and evaluation records preserve the target policy, frozen checkpoint hash, and per-system test results. The 11 static-source exceptions and the high single-frame geometry flag rate prevent a blanket physical-quality claim for either control.

The original and three adapted Protenix models each generate 32 conformations per test RNA using 200 diffusion steps and inference seed 101. All 12,416 conformations across the four arms and all 97 test RNAs are retained. For C4′ pairs separated by at least four sequence positions, empirical distance W_1 compares generated distances with all 991 MD frames, using minimum-image MD distances. Centered W_1 subtracts each pair distribution's own mean before comparison; width error is the absolute difference in its standard deviation. Pair scores are averaged within RNA. C4′ RMSF is computed after rigid alignment to the initial MD C4′ structure and compared by within-RNA MAE and Spearman correlation. These atom and pair definitions differ from the primary benchmark, so their scores are reported separately.

A generated conformation is flagged if any checked bond lies outside 0.7–2.4 Å, any atom pair more than two covalent bonds apart overlaps within 1.0 Å, or any base ring has plane-fit RMS deviation above 0.2 Å. This is a geometric screen, not an energy or kinetic assessment. Paired intervals resample the 77 complete test components 2,000 times (seed 20260916); no training-seed variation or multiplicity adjustment is included. The grouped split controls MD adaptation, while overlap with the original model's pretraining structures has not been independently audited. The evaluation measures finite-window ensemble fidelity, not transition rates or exhaustive conformational coverage.

External generators and common evaluation. We ran the published DynaRNA (Li et al., 2025) and RNAbpFlow (Tarafder & Bhattacharya, 2026) checkpoints on the same frozen 97-system test manifest, producing 32 samples per system without MD coordinates in either generation input. DynaRNA receives the source PDB C4′ coordinates and uses the official `DynaRNA.pkl` checkpoint with 800 forward/reverse diffusion steps. Its shared initial noisy structure had to be explicitly repeated across the 32-sample batch to avoid an in-place broadcasting failure; the model and checkpoint were otherwise unchanged. The official input/output coordinate conversions cancel, and saved samples are in Å. DynaRNA outputs C4′ coordinates only.

RNAbpFlow uses the official CASP16 checkpoint, 50 flow steps, and the published sample seeds $420, 840, \ldots, 13440$. Inputs are the RNA sequence and predicted base-pair maps from RibonanzaNet-SS and IPknot. The three channels are [RibonanzaNet-SS, IPknot, RibonanzaNet-SS], a repository-supported two-map mode. SPOT-RNA weight downloads were unavailable in the reproduction environment, so this run is *not* an exact replication of the paper's three-predictor input configuration. Both external models used frozen official weights selected before RiboMotion test evaluation; no test-driven checkpoint selection or sample filtering was applied. Git revisions, dependency versions, checkpoint SHA-256 values, input identities, and per-system outputs are recorded in the experiment provenance.

The common comparison uses the same 991-frame MD reference, minimum-image C4′ distances for residue pairs separated by at least four positions, and aligned C4′ RMSF calculations. All 97 systems and 32 samples per method are retained. Full-atom geometry is deliberately absent from Table S23: DynaRNA has no corresponding atoms, so that screen cannot be harmonized across methods. The external models have different inputs and pretraining corpora from Protenix and were not fine-tuned on the 550 RiboMotion training RNAs. The RNAbpFlow CASP16 training list contains 15 exact PDB-chain matches among the 97 test systems; the 82-system sensitivity analysis removes those exact matches, but cannot exclude homologous overlap. DynaRNA pretraining overlap and Protenix's original pretraining overlap remain unaudited. The external-generator comparison in Table S23 reports shared ensemble metrics across methods. The matched Protenix controls isolate the effect of the supervision target.

Table S23: **Common C4′ ensemble metrics for pretrained and adapted generators.** Values are equal-system means on all 97 test RNAs and on the 82 without an exact PDB-chain match to the RNAbpFlow CASP16 training list. Every method generates 32 conformations per RNA. Lower is better except for Spearman. External methods differ in inputs and pretraining; the 82-system subset does not establish homology-level independence.

Scope	Generator	Total W_1	Centered W_1	Width	RMSF MAE	Spearman
97	Protenix original	3.629	1.296	1.523	2.062	0.557
97	Static PDB-adapted	4.050	1.359	1.516	1.872	0.591
97	Single-frame-adapted	3.581	1.281	1.428	1.794	0.616
97	Multi-frame-adapted	3.207	0.967	0.997	1.321	0.661
97	DynaRNA	6.701	3.750	4.475	6.476	0.079
97	RNAbpFlow (two-map)	3.549	1.460	1.590	2.159	0.458
82	Protenix original	3.468	1.261	1.480	1.950	0.558
82	Static PDB-adapted	4.001	1.311	1.468	1.795	0.599
82	Single-frame-adapted	3.430	1.268	1.408	1.742	0.624
82	Multi-frame-adapted	3.038	0.938	0.964	1.245	0.673
82	DynaRNA	6.587	3.667	4.378	6.312	0.079
82	RNAbpFlow (two-map)	3.434	1.420	1.533	2.008	0.452

I Data and Reproduction Resources

I.1 Data and prediction interfaces

The MD archive at https://huggingface.co/datasets/7ac23az/RiboMotion-MD provides processed trajectories and supporting metadata. The accompanying anonymous reproduction package contains the final grouped split and six-task test labels, with explicit residue or

interaction identities and scoring masks. Frozen predictions cover the methods in Table 3, including the external DeepRMSF reference. A CPU evaluator recomputes within-system scores from these arrays before averaging systems and separately scored seeds. Identity checks, tied-score AP, no-positive-system conventions, undefined correlations, and exact empirical W_1 are implemented explicitly.

The evaluator reproduces 8,645 system–method–seed records to an absolute tolerance of 10^{-6} and all 25 displayed primary-score rows of Table 3, including the reported seed standard deviations and five Brier values. The package includes analytical metric checks and a machine-readable record of the score-reproduction checks. Four small validation examples remain available for introducing the prediction interfaces; the full test evaluator covers all six tasks. This release reproduces scoring from frozen predictions, while complete model retraining additionally requires development labels, prepared training inputs and model assets.

I.2 Conformation-generation reproduction

The generation records allow recomputation of the six methods' common C4′ scores on the 97-RNA cohort and the exact-match overlap sensitivity. Appendix H.1 specifies target selection, adaptation, sampling, metrics and the RNAbpFlow two-map configuration. The package also includes project-authored Protenix adaptation and external-generator wrapper source references, with configurable workspace paths. These sources were checked for syntax but not rerun in a fresh GPU environment for this release. Raw conformer regeneration requires the pinned upstream code and official weights, prepared inputs, and adapted model checkpoints; these assets are not redistributed in the scoring package.

I.3 Evaluation records

All primary task scores use the final grouped split. Replica comparisons and fixed-budget controls use the same component assignment with the selection rules specified in their methods. Occupancy remains an auxiliary training objective in one ligand model, rather than a benchmark endpoint. Each prediction is linked to its system, method, seed, and evaluation population.