MDRNACAT: A MOLECULAR DYNAMICS RNA CATALOG OF REST2 CONFORMATIONAL ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning molecular models of RNA requires simulation data that capture conformational variability and provide consistent physical training signals. Here, we introduce mdRNAcat, a curated dataset of all-atom molecular dynamics (MD) simulations of 48 RNA systems selected to promote sequence diversity. Each system is simulated using replica-exchange with solute scaling (REST2), with 32 Hamiltonian states and production lengths of 200–300 ns per state. The dataset provides atomic coordinates, instantaneous forces, and forces reevaluated under the unscaled reference Hamiltonian, together with structural annotations and replica-exchange metadata. Coordinates and forces are available at 1 ps intervals for the reference state and 100 ps intervals across all states in a standardized HDF5 format. We characterize conformational sampling, exchange behavior, base pairing, and sugar-pucker distributions to document the properties of the simulated ensembles. mdRNAcat provides a reusable resource for developing and evaluating data-driven RNA models, including coarse-grained force fields.

1 INTRODUCTION

At the sub-cellular scale, life is an intricate choreography of biomolecules in motion, whose understanding requires to know both their structure and dynamics. Despite major advances in experimental techniques, many important biomolecules remain structurally unresolved, with their conformational dynamics, at best, partially characterized. In principle, molecular dynamics (MD) simulations provide atomistic access to such dynamics, but the relevant time scales are often out of reach due to prohibitive computational expenses (Hollingsworth & Dror (2018)). Recently, machine learning methods have made substantial progress in structure prediction, such as Alphafold 3 (Abramson et al. (2024)) but also for learning molecular interactions and dynamics (Kovács et al. (2025)). However, much of this progress and the underlying training data have centered on proteins rather than RNA. This is in part due to differences in available data. Machine learning methods rely on available data to achieve the aforementioned results. For proteins, large collections of experimentally determined structures (Berman et al. (2000)), MD data sets, such as mdCATH and ATLAS (Mirarchi et al. (2024); Vander Meersche et al. (2024)), and data sets of quantum-mechanical properties (Levine et al. (2025); Eastman et al. (2024)) are available. In contrast, RNA lacks a comparably large and systematically generated MD resource. While curated RNA structure collections and nucleic-acid simulation repositories exist Upadhyay et al. (2025); Hospital et al. (2016), they do not provide the combination of scale, homogeneous protocols, large-scale conformational sampling, and consistently available force information targeted by mdRNAcat. To our knowledge, no existing RNA MD dataset combines these properties. This gap is particularly relevant given the growing interest in data-driven RNA modeling Upadhyay et al. (2026) and the need for large, consistent training datasets for coarse-grained machine-learned force fields Li & Chen (2026).

More precisely, while existing molecular simulation repositories are valuable resources, their use for systematic data-driven analyses is limited when trajectories originate from heterogeneous simulation protocols Vander Meersche et al. (2024); Kouba et al. (2023). Variations in force fields, solvent models, simulation conditions, sampling strategies, and trajectory lengths can introduce systematic differences that are difficult to separate from genuine molecular variation. Moreover, available observables are often inconsistent across repositories and datasets; in particular, atomic forces, which are essential for learning molecular interactions and fitting force fields, are not routinely retained Mirarchi et al. (2024). For studying RNA conformational ensembles and developing transferable

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.

Table 1: Statistical overview over the data set.

Total RNA systems	48
Systems with 200 ns simulation time	45
Systems with 300 ns simulation time	3
Replicas	32
State trajectories	1,536
Total sampled time	316.8 µs
Average atoms per system	1,955
Average residues per system	61
Total file size	699 GB

data-driven models, an ideal dataset should therefore combine broad coverage of relevant RNA systems with a homogeneous simulation protocol, enhanced conformational samplingWang et al. (2011); Bussi & Laio (2020a), and a consistent set of physical observables, including both coordinates and forces.

To address these limitations, we present mdRNAcat, a homogeneous, enhanced-sampling MD dataset designed for systematic analysis and data-driven modeling of RNA conformational ensembles (Figure 1 and Table 1). The dataset is intended to support tasks such as force matching, learned molecular potentials, coarse-grained model development, and representation learning of RNA conformational ensembles.

Our main contributions are:

- We **generate a systematically curated set** of 48 sequence-, structurally and functionally diverse RNA systems simulated under a single, consistent all-atom MD protocol with REST2 enhanced sampling, **spanning diverse regions of RNA sequence space while broadening conformational sampling beyond conventional MD**.
- We provide a **rich and standardized set of physical training signals**, including atomic coordinates, instantaneous forces, and reference-Hamiltonian forces for configurations sampled across REST2 states, together with structural annotations and replica-exchange metadata **in a common HDF5 representation**.
- We **quantitatively characterize and validate mdRNAcat** in terms of sequence-space coverage, enhanced conformational sampling, replica-exchange statistics, and structural properties, **providing dataset-wide quality-control measures for downstream data-driven RNA modeling**.

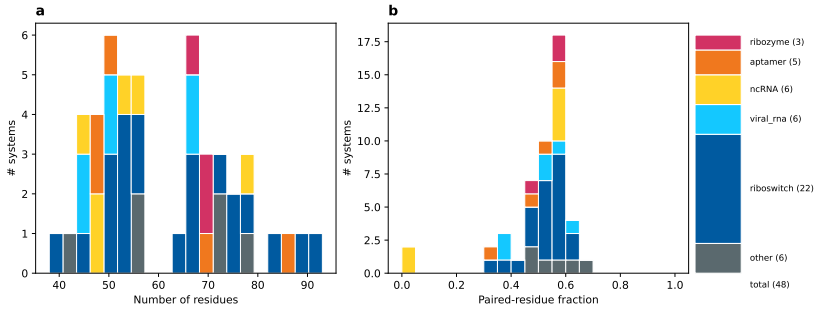

Figure 1: (**a**) Distribution of the number of residues per RNA system in the dataset. (**b**) Distribution of the percentage of paired residues in the dataset; most systems display similar behavior, apart from one RNA system in an extended state. Colors indicate different RNA domains.

2 Dataset Requirements

To achieve the goals for the dataset, we set a number of requirements that the data should adhere to.

- *Reproducibility:* The `.pdb` files and GROMACS topology files are included in the dataset; the processing and simulation protocols are described in detail in the Appendix A.4. This ensures comparable and consistent future work on this dataset.
- *Interoperability:* The dataset is saved in the HDF5 format as a widely used and supported data format. The formatting is kept as similar to mdCATH (Mirarchi et al. (2024)) as possible, to ensure reusability of code.
- *Accessibility:* The files are annotated with all relevant meta data on the trajectories as well as derived parameters such as radius of gyration or sugar pucker angle. The trajectories contain coordinates as well as the forces and for higher temperature replicas also the recalculated forces at the lowest temperature.
- *High quality data:* For accurate learning of the RNA conformation, an accurate representation in the training set is vital. Thus we used state-of-art force-field, sampling methods and simulation code to generate mdRNAcat. Critically, the simulations should be homogeneously run.
- *Sequence space coverage:* Sequence-space coverage was quantified against a filtered set of all available RNA sequences with structures in the PDB, using the analysis described in Section 5.1.
- *Conformation space coverage:* To also sample the conformation space widely, we used state-of-art a replica exchange algorithm in REST2 to find and sample conformations further from the global minimum.

3 METHODS

3.1 SEQUENCE SET GENERATION

The data set is based on experimentally resolved sequences from the RCSB Protein Data Bank (PDB) (Berman et al. (2000)). With NucleoSeeker (Upadhyay et al. (2025)), a set of sequences was searched that have a pairwise sequence identity of less than 50 % and contain the highest number of sequences in the set, thus covering as much of the known RNA structure space as possible. Further criteria were structures containing only RNA and a sequence length of at least 40 residues to ensure that short common sequences do not disallow most of the sequence space due to high sequence identity. Additionally, only X-Ray structures with a resolution below 3.6 Å were initially considered. This led to a set containing 105 RNA sequences and structures with length between 40 and 192 residues. From this set 43 structures were chosen and partly edited to be readable for GROMACS with truncation of modified regions or insertion of missing atoms. The remaining 62 structures were not usable for GROMACS, out of scope for the chosen simulation setup and available compute resources,[1] or de-prioritized in favor of other systems. Structures were not usable for GROMACS because of non-recoverable modifications that were not implemented in the used forcefield or large gaps in the structure that were too large to reconstruct. Lastly, 5 NMR-structures were manually added to broaden the sequence space, yielding a total of 48 structures. A detailed list with all systems considered and used can be found in the Appendix A.2.

3.2 SIMULATION

The trajectories contained in mdRNAcat were generated with all-atom Molecular Dynamics (MD) simulation. MD simulation are physics-based approach and common tool for computational analysis of biomolecular system (Hollingsworth & Dror (2018)). For MD simulations, the system needs to be described with physical parameters such as atom charge and bond strengths, which are called force fields (FF). These are used to calculate frame-wise the forces and thus velocities and positions for each particle in the simulated system. Because the energy barriers between different minima in the conformation space of biomolecule are commonly large, often enhanced sampling is used to improve MD simulations (Abrams & Bussi (2013)). In this work, we used replica exchange with solute scaling (REST2) algorithm (Wang et al. (2011)) implemented via PLUMED (Bonomi et al. (2009); Tribello et al. (2014); consortium (2019)). REST2 does not scale the temperature of the

[1] Scope cutoff is not solely based on residue number of the system, the system sizes were tested in short simulations to produce good replica exchange statistics, see Section 5.3

replica systems, but modifies their Hamiltonian and scales the solute-solute and solute-solution interactions. As a consequence, only the lowest temperature replica is physically correctly simulating its temperature, but since it only scales the interaction, it also allows for recalculating the forces correctly for all replicas as if they were at the lowest temperature.

In this work, GROMACS (Abraham et al. (2015)) was used for all pre-processing and simulations of the systems. For the RNA parameters the tAMBER99SB-ildn-star force field with parmbsc0 corrections (Case et al. (2025); Pérez et al. (2007); Guy et al. (2012)) with reparameterization from Tan et al. (2018) and phosphate parameters from Steinbrecher et al. (2012) was used. Solvation was done in cubic boxes with TIP4PD water(Piana et al. (2015)) and a padding of 10 to 20 Å depending on the system. The systems were neutralized and ionized with K^+ and Cl^- ions at 0.15 M concentration with CHARMM22-ions (MacKerell Jr et al. (1998)). A number of cations equal to half the number of residues in the system was manually and randomly placed closer to the negatively charged solute to achieve faster equilibration. Additionally particle-mesh Ewald (PME) with a cutoff of 12 Å for long-range electrostatic forces, hydrogen constrains and 2 fs simulation steps were used.

For equilibration, first an energy minimization run with a maximum force of 10 kJ mol^{-1} nm^{-1} was done. Then two 100 ps simulation runs of an NVT and NPT ensemble were carried out, respectively.

Production simulations were done for 200 ns (apart from three systems for 300 ns) and with 32 replicas for each system with a geometric temperature distribution between 300 K and 500 K. The full list of simulation parameters can be found in the Appendix A.4.

3.3 Post-Processing

Adding further derived quantities to the data set, MDAnalysis (Michaud-Agrawal et al. (2011); Gowers et al. (2016)) was used to calculate the root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF) and radius of gyration. For the dot bracket annotation of base pairs and the sugar pucker angle, MDtraj (McGibbon et al. (2015)) and barnaba (Bottaro et al. (2019)) were used.

The forces of all higher temperature trajectories were additionally recalculated at the lowest temperature 300 K with the `rerun` function from GROMACS.

4 Data Records

mdRNAcat is available on Hugging Face at `https://huggingface.co/datasets/anonymous8576/mdrnacat` (Anonymous (2026)), with access via direct download or the Hugging Face API. For each system in the dataset there are two files in the Hierarchical Data Format version 5 (HDF5), both labeled with the respective PDB ID of the system. The first contains the coordinates and forces of the lowest temperature replica with 1 ps time steps. The second file, labeled with `replica`, is only saved in with 100 ps time steps, but contains all replicas, replica exchange meta-data and additional derived quantities.

The organization within the HDF5 files is kept as similar as possible to the formatting used in the mdCATH dataset (Mirarchi et al. (2024)) to ensure compatibility of code. A detailed list of the saved properties and values can be found in Table 2. Each file has the PDB ID as the root group, which contains for each atom in the system `chain`, `element`, `resid`, `resname` and its atom number `z`. Additionally the number of atoms, residues and replicas, as well as the temperature minimum and maximum are saved, along with the `.pdb` file and the GROMACS topology file. The `replica` files contain the `repex` group with replica exchange parameters such as the energy per replica, exchange attempts, successful exchanges and temperature ladder. For each replica at temperature `<T>/` (so for the non-`replica` file only the lowest replica), there is a group containing the box parameters, temperature, number of frames, stride as well as the coordinates and forces. In the `replica` files, these groups also contain derived analytical values such as radius of gyration, root-mean-square deviation and fluctuation (RMSD, RMSF), base-pair dot brackets, sugar pucker angle and the forces scaled to the lowest temperature. The latter allows additional lowest-temperature-frames with physically correct forces. However, the distribution in the conformation space of these frames is not thermodynamically correct, which has to be taken account depending on the intended usage.

Table 2: mdRNAcat is made available as HDF5 files with an hierarchical organization as shown in this table. For each system in the data set, there are 2 files both labeled with the respective PDB ID. The first contains only the lowest trajectory with 1 ps time steps. The second file contains all replicas and additional derived metrics indicated with a blue background in the table with 100 ps time steps. Key: `<T>`, REST2 effective temperature (K); N, number of atoms; R, number of residues; F, trajectory length in frames (= `frameStridePs`); M, number of replicas; K, energy terms; S, energy samples; X, exchange records; P, log segments.[2]

Field	Size	Type	Unit	Description
`<pdbid>/`				
`chain`	N	string		Chain ID
`element`	N	string		Chemical element
`resid`	N	integer		Residue number
`resname`	N	string		Residue name
`z`	N	integer		Atomic number
`pdb`	1	string		RNA reference PDB
`top`	1	string		GROMACS topology
`.numProteinAtoms`	1	integer		Number of (RNA) atoms N (attribute)
`.numResidues`	1	integer		Number of residues R (attribute)
`.numReplicas`	1	integer		Number of replicas M (attribute)
`.rest2TempMin/Max`	1	float	K	REST2 ladder end points (attributes)
`<T>/`				
`0/`				
`coords`	$F \times N \times 3$	float	Å	Atom coordinates
`forces`	$F \times N \times 3$	float	kcal/mol/Å	Forces
`scaledForces`	$F \times N \times 3$	float	kcal/mol/Å	λ=1 reference forces
`gyrationRadius`	F	double	nm	Gyration radius
`rmsd`	F	float	nm	RMSD w.r.t. first frame
`rmsf`	R	float	nm	Per-residue RMSF
`secondaryStructure`	$F \times R$	string		Base-pairing dot-bracket
`sugarPucker`	$F \times R$	float	degree	Ribose pseudorotation phase
`box`	3×3	float	nm	Simulation unit cell
`.temperature`	1	float	K	Effective temperature (attribute)
`.numFrames`	1	integer		Number of frames F (attribute)
`.frameStridePs`	1	float	ps	Time between frames (attribute)
`repex/`				
`terms`	K	string		Interaction-energy term names
`energy`	$M \times S \times K$	float	kJ/mol	Per-state interaction energies
`energyTime`	$M \times S$	float	ps	Energy sample times
`energyNSamples`	M	integer		Valid samples per state
`replicaIndex`	$X \times M$	integer		Demux occupancy (replica per state)
`replicaIndexTime`	X	float	ps	Exchange-record times
`nAttempts`	P	float		Exchange attempts per log segment
`nExchanges`	$P \times (M-1)$	float		Accepted exchanges per adjacent pair
`temps`	M	float	K	Effective-temperature ladder
`<T′>/, ...` *remaining replicas*				

As of the publication date, mdRNAcat contains 48 different RNA systems with a total simulation time of 316.8 µs and a total file size of 699 GB (Table 1).

5 Analysis and Validation

5.1 Sequence and Structural-Motif Space Coverage

We assessed mdRNAcat against filtered X-ray PDB reference sets released by 25 September 2026, reflecting the experimental method used for initial dataset selection (Figure 2). All 48 mdRNAcat systems were included as representatives, irrespective of experimental method. Coverage was eval-

[2] A comment on the units used in mdRNAcat can be found in the Appendix A.1

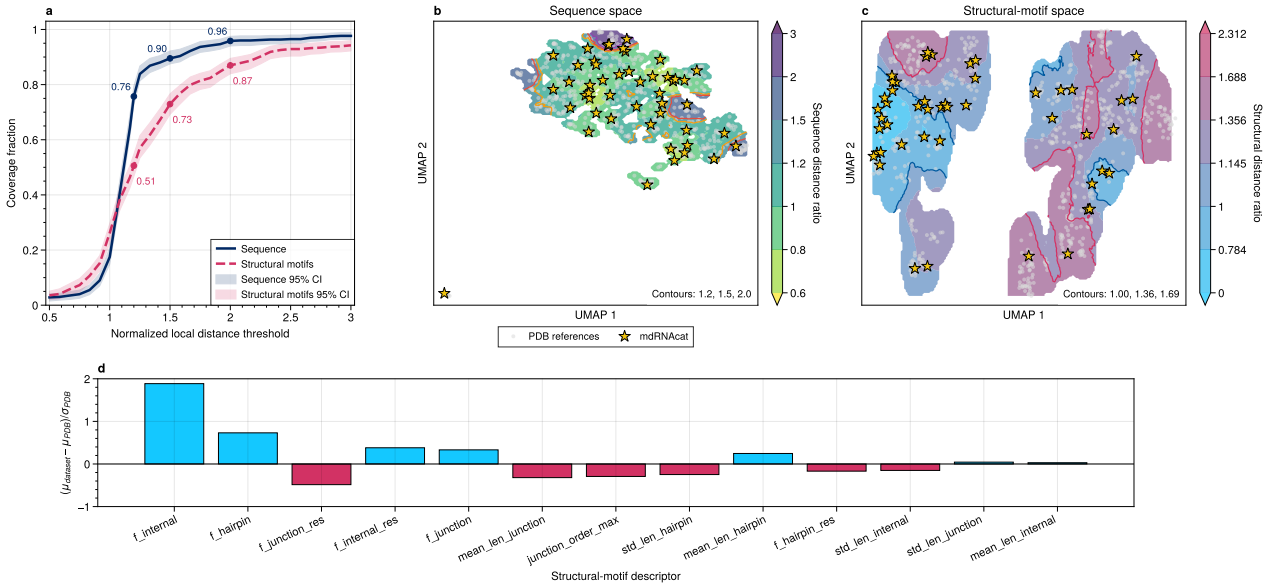

Figure 2: Sequence and structural-motif coverage of mdRNAcat against filtered X-ray PDB reference sets. (**a**) Coverage versus locally normalized nearest-representative distance. (**b**,**c**) UMAP visualizations McInnes et al. (2018a;b); stars denote mdRNAcat representatives and colors show smoothed distance ratios on separate scales, with structural values clipped for display. (**d**) Mean structural-descriptor differences in units of the reference-set standard deviation. Coverage is calculated before UMAP projection.

uated using Euclidean distances in SVD-reduced 3-mer TF–IDF space (Moussa & Măndoiu (2018)) and in standardized space of 13 RNA 3D Hub structural-motif descriptors (Petrov et al. (2013)).

Distances to the nearest representative were normalized by the median distance to ten non-self neighbors (Appendix A.3). Excluding mdRNAcat from evaluation, coverage at 1.5 local-neighbor scales reached 89.4 % for sequences and 72.9 % for structural motifs. The latter characterizes deposited structures, not the conformational ensembles sampled by REST2.

5.2 Enhanced Sampling Validation

To validate the methodological use of REST2 enhanced sampling of RNA conformational states, we compare the free energy landscapes (FELs) (Bussi & Laio (2020b); Rohrdanz et al. (2013)) obtained through sampling both plain MD and REST2 trajectories of an illustrative system, the Glutamine II Riboswitch (PDB ID: 6qn3) (Figure 3). The REST2 landscape is constructed only from the lowest temperature trajectory, which is in thermodynamic equilibrium. Compressing the $3 \cdot N_{atoms}$-dimensional conformation space down to two dimensions and enabling planar visualization requires targeted dimensionality reduction techniques. The landscapes are constructed using time-lagged independent component analysis (TICA) (Pérez-Hernández et al. (2013)), which gives the axis of slowest dynamical motion along the trajectory. The plain MD simulation is not included in the mdRNAcat dataset. It has the same pre-processing and simulation parameters as the REST2 simulations with the same simulation length of 200 ns, apart from sampling only a single trajectory at 300 K (which is equivalent to the lowest temperature of the REST2 simulations) without any replica exchange.

For the REST2 simulation, one can see that the trajectory successfully samples outside of the experimentally resolved state. There are three distinct energy minima which get sampled more often than the rest of the conformation space, one of which contains the experimentally resolved structure. The plain MD simulation only finds minima of the experimentally resolved state where it started and fails to find the other important ensemble states. In total only 47.6 % of the area sampled by REST2 is also sampled in the plain MD. This result sufficiently demonstrates the enhanced sampling offered by state-of-the-art algorithms such as REST2 which we used for this dataset.

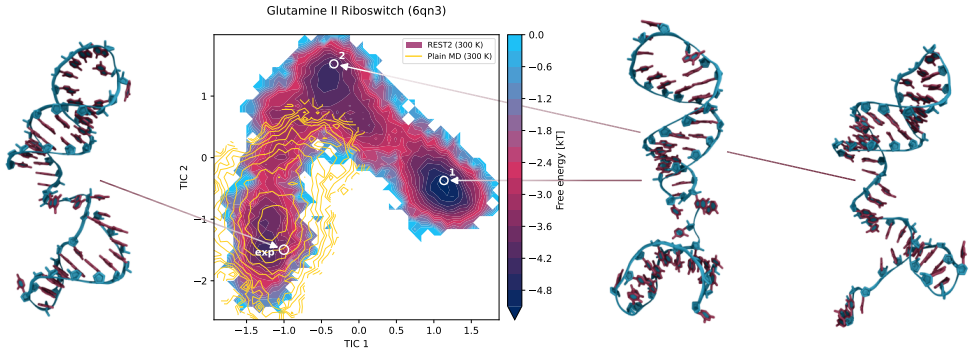

Figure 3: Free energy landscape (FEL) of the lowest temperature (300 K) trajectory of the 6qn3 system, the axes are determined with time-lagged independent component analysis (TICA) (Pérez-Hernández et al. (2013)), marked with the experimentally resolved structure (exp) and the two lowest minima (1 and 2); experimentally resolved structure is displayed to the left and structures from the two minima to the right; in yellow FEL of a plain single replica 300 K simulation of 6qn3 (not part of mdRNAcat) mapped to the same TICA coordinates.

Table 3: Replica exchange statistics averaged over all 48 systems with standard deviation between systems.

Avg. transition probability	0.183	$\pm$0.065
Avg. energy overlap	0.269	$\pm$0.070

5.3 Replica Exchange Validation

To ensure that the replica exchange algorithm performed as intended, we monitored the exchange statistics and characteristics of each system (Figure 4 as an illustrative system, all other systems can be found in the Appendix A.5). Replica exchange requires exchange rates around 20 % (Kone & Kofke (2005); Rathore et al. (2005)) are favorable, which we checked in (a). For exchanges to happen, there needs to be an adequate overlap in the energy distribution of the different thermodynamic states (which can be seen in (b)). Lastly, subplot (c) does not track the thermodynamic state, but the energy distribution of the frames of a single replica that gets exchanged to higher and lower temperatures. This demonstrates that they are not stuck in certain temperatures. A statistical summary over all systems can be found in Table 3.

In higher energy replicas, one expects a broader sampling of the conformation space. We plotted the illustrative FEL for 3 different systems at 4 different temperatures (Figure 5). In all systems, the conformation space is explored more widely with rising temperature. This confirms the validity of the temperature scaling.

5.4 Base Pairing and Backbone Torsion

In this section we wanted to validate the structures of the system. The first metric looked at is the angle of the sugar pucker in the backbone. This indicates the angle torsion and thus the secondary structure of the RNA (Altona & Sundaralingam (1972); Saenger (1984)). In Figure 6 (a), the percentage of each torsion angle is plotted in a wheel plot. A north facing angle means C3'-endo conformation and an A-form helix, which most residues and frames in the dataset represent. There is also a small fraction facing south indicating C2'-endo conformation in loops and bulges.

Additionally, we investigated the base pairings. In Figure 6 (b) and (c), the averaged probability for each sequence separation between two paired bases is plotted. As the systems are not of the same length, the tail of the distribution in (b) is dominated by a few larger systems. In (c), the separation distance is normed to 1, so that all systems in the dataset are on the axis. This reveals that the pair probability grows with the separation distance. This is due to the double helical conformation,

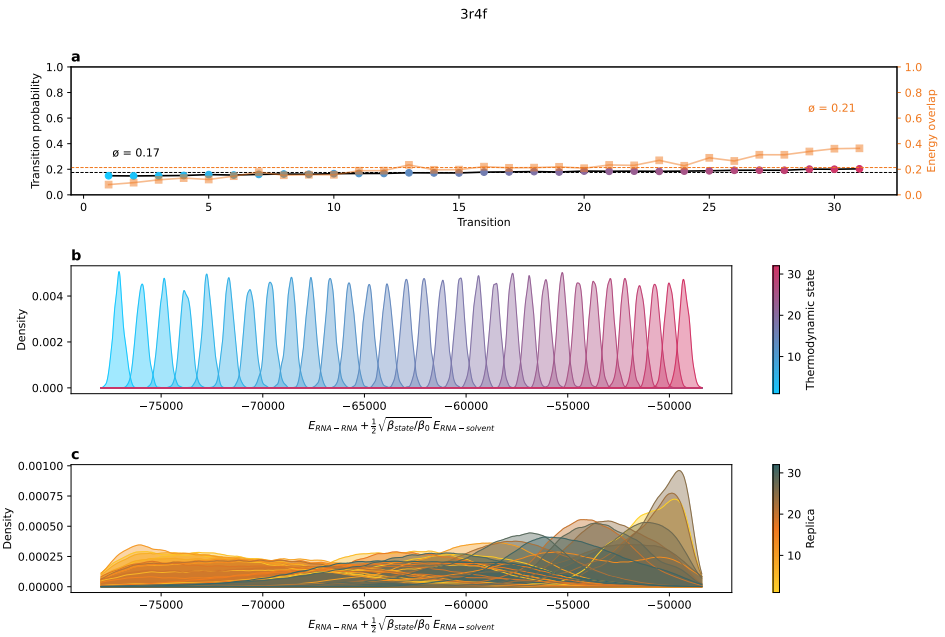

Figure 4: (**a**) Transition probability between two thermodynamic states and energy overlap between two thermodynamic states against the 31 transitions of the states for the 3r4f system, with dotted lines indicating the average over all transitions. (**b**) Density of the thermodynamic states against the energy of the same system. (**c**) Density of the replicas against the energy of the same system.

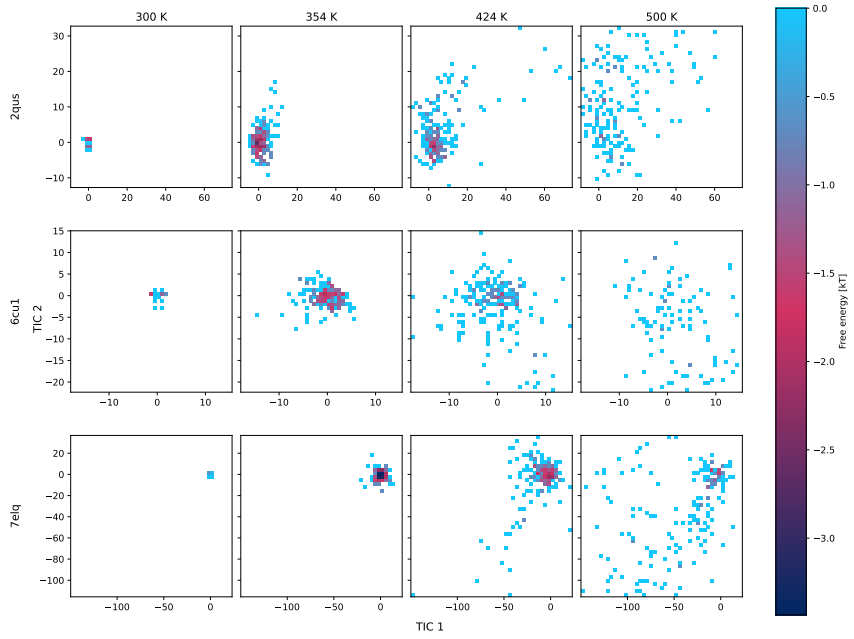

Figure 5: Free energy landscape (FEL) of the 2qus, 6cu1 and 7elq systems, the axes are determined with time-lagged independent component analysis (TICA) (Pérez-Hernández et al. (2013)) at 300 K, 354 K, 424 K and 500 K. Note that the axes are capped at the extension of the 424 K-plot, so the 500 K-plot has values outside of that range not displayed. The conformation space gets explored more with higher temperature.

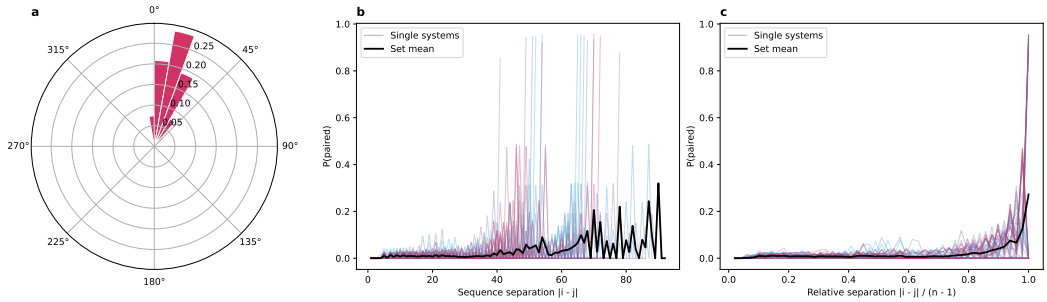

Figure 6: (**a**) Wheel histogram over the sugar-pucker angle of all residues of all systems; north (0°) indicates C3'-endo, so A-form helical pucker; south (180°) indicates C2'-endo, so loops and bulges. (**b**) Distribution of the residue pair likelihood against the sequence separation between the two paired residue for all systems in the data set; thick is the set mean. (**c**) Same as (**b**) but the x-axis normalized to 1.

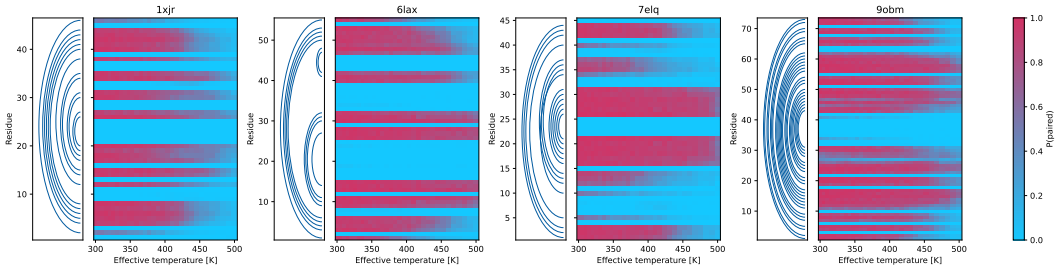

Figure 7: Probability of a paired residue in residue and temperature dimensions for the 1xjr, 6lax, 7elq and 9obm systems. On the left of each plot is a bow plot indicating the paired residues in the experimentally resolved structure.

pairing both ends of a sequence, while the middle part forms a loop, confirming the results of the backbone torsion analysis.

Lastly, we expect base pairs to break in higher energy states, for that we plotted the pair probability per residue and per temperature (Figure 7). On the left of each plot, bow plots also indicate which bases are paired in the experimentally resolved structure. Again one can observe that the ends are paired and the center builds a loop. With higher temperature the pairs are breaking, usually starting with the ends, while the center of the RNA stays stable for higher energies. This is expected behavior for RNA and further confirms the validity of our simulations.

AI USE STATEMENT

In this work, we used generative AI tools for writing part of the code used for post-processing and plotting. We have not used generative AI tools for the conception of the dataset, any pre-processing or the simulations and the generated data themselves. The rest of the required disclosure tasks such as mathematical claims or proofs are not applicable to this work. Additionally, we used generative AI tools for searching relevant literature and other works. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we included all `.pdb` and topology files in the dataset itself. Additionally detailed record of all pre-processing and simulation parameters can be found in the Appendix.

The code used to create the validation done in this paper can be found under: https://anonymous.4open.science/r/mdrnacat-C06D/

Author Contributions

This section will not be included in the anonymous version of the paper.

Acknowledgments

This section will not be included in the anonymous version of the paper.

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*, 1:19–25, 2015.

Cameron Abrams and Giovanni Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2013.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

C t Altona and M Sundaralingam. Conformational analysis of the sugar ring in nucleosides and nucleotides. new description using the concept of pseudorotation. *Journal of the American Chemical Society*, 94(23):8205–8212, 1972.

Anonymous. mdrnacat (revision 599ab8b), 2026. URL https://huggingface.co/datasets/anonymous8576/mdrnacat.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A. Broglia, and Michele Parrinello. Plumed: a portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961–1972, 2009. doi: 10.1016/j.cpc.2009.05.011. URL https://doi.org/10.1016/j.cpc.2009.05.011.

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.

Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, 2020a. doi: 10.1038/s42254-020-0153-0.

Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, 2020b.

David A Case, David S Cerutti, Vinícius Wilian D Cruzeiro, Thomas A Darden, Robert E Duke, Mahdieh Ghazimirsaeed, George M Giambasu, Timothy J Giese, Andreas W Gotz, Julie A Harris, et al. Recent developments in amber biomolecular simulations. *Journal of chemical information and modeling*, 65(15):7835–7843, 2025.

The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods*, 16(8):670–673, 2019. doi: 10.1038/s41592-019-0506-8. URL https://doi.org/10.1038/s41592-019-0506-8.

Peter Eastman, Benjamin P Pritchard, John D Chodera, and Thomas E Markland. Nutmeg and spice: models and data for biomolecular machine learning. *Journal of chemical theory and computation*, 20(19):8583, 2024.

Richard Gowers, Max Linke, Jonathan Barnoud, Tyler Reddy, Manuel Melo, Sean Seyler, Jan Domański, David Dotson, Sébastien Buchoux, Ian Kenney, and Oliver Beckstein. Mdanalysis: A python package for the rapid analysis of molecular dynamics simulations. In *Proceedings of the 15th Python in Science Conference*, pp. nil, 2016. doi: 10.25080/majora-629e541a-00e. URL https://doi.org/10.25080/majora-629e541a-00e.

Andrew T Guy, Thomas J Piggot, and Syma Khalid. Single-stranded dna within nanopores: conformational dynamics and implications for sequencing; a molecular dynamics simulation study. *Biophysical journal*, 103(5):1028–1036, 2012.

Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6): 1129–1143, 2018.

Adam Hospital, Pau Andrio, Carles Cugnasco, Laia Codo, Yolanda Becerra, Pablo D. Dans, Federica Battistini, Javier Torres, Ramon 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.

Aminata Kone and David A Kofke. Selection of temperature intervals for parallel-tempering simulations. *The Journal of chemical physics*, 122(20), 2005.

Petr Kouba, Pavel Kohout, Faraneh Haddadi, Anton Bushuiev, Raman Samusevich, Jiri Sedlar, Jiri Damborsky, Tomas Pluskal, Josef Sivic, and Stanislav Mazurenko. Machine learning-guided protein engineering. *ACS catalysis*, 13(21):13863, 2023.

Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Yixuan Pu, Venkat Kapil, William C Witt, Ioan-Bogdan Magdau, Daniel J Cole, et al. Mace-off: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025.

Daniel S Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G Taylor, Muhammad R Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, et al. The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Wenfei Li and Shi-Jie Chen. Building rna coarse-grained force fields: Design principles and training strategies. *Biophysical Journal*, 125(18):4898–4912, 2026. doi: 10.1016/j.bpj.2026.03.041. URL http://dx.doi.org/10.1016/j.bpj.2026.03.041.

Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, Houyang Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.

Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528–1532, 2015. doi: 10.1016/j.bpj.2015.08.015. URL https://doi.org/10.1016/j.bpj.2015.08.015.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *CoRR*, 2018a. URL http://arxiv.org/abs/1802.03426v3.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018b. doi: 10.21105/joss.00861. URL http://dx.doi.org/10.21105/joss.00861.

Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. Mdanalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10):2319–2327, 2011. doi: 10.1002/jcc.21787. URL https://doi.org/10.1002/jcc.21787.

Antonio Mirarchi, Toni Giorgino, and Gianni De Fabritiis. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11(1):1299, 2024.

Marmar Moussa and Ion I. Măndoiu. Single cell rna-seq data clustering using tf-idf based methods. *BMC Genomics*, 19(S6):569, 2018. doi: 10.1186/s12864-018-4922-4. URL http://dx.doi.org/10.1186/s12864-018-4922-4.

Alberto Pérez, Iván Marchán, Daniel Svozil, Jiri Sponer, Thomas E Cheatham, 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.

Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of Chemical Physics*, 139(1):015102, 2013. doi: 10.1063/1.4811489. URL http://dx.doi.org/10.1063/1.4811489.

Anton I. Petrov, Craig L. Zirbel, and Neocles B. Leontis. Automated classification of rna 3d motifs and the rna 3d motif atlas. *RNA*, 19(10):1327–1340, 2013. doi: 10.1261/rna.039438.113. URL http://dx.doi.org/10.1261/rna.039438.113.

Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *The journal of physical chemistry B*, 119(16):5113–5123, 2015.

Nitin Rathore, Manan Chopra, and Juan J de Pablo. Optimal allocation of replicas in parallel tempering simulations. *The Journal of chemical physics*, 122(2):024111, 2005.

Mary A Rohrdanz, Wenwei Zheng, and Cecilia Clementi. Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annual review of physical chemistry*, 64(1):295–316, 2013.

Wolfram Saenger. Structures and conformational properties of bases, furanose sugars, and phosphate groups. In *Principles of Nucleic Acid Structure*, Springer Advanced Texts in Chemistry, chapter 4. Springer, New York, 1984. ISBN 978-0-387-90761-1. doi: 10.1007/978-1-4612-5190-3.

T Steinbrecher, J Latzer, and DA Case. Revised amber parameters for bioorganic phosphates. *Journal of chemical theory and computation*, 8(11):4405, 2012.

Dazhi Tan, Stefano Piana, Robert M Dirks, and David E Shaw. Rna force field with accuracy comparable to state-of-the-art protein force fields. *Proceedings of the National Academy of Sciences*, 115(7):E1346–E1355, 2018.

Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014. doi: 10.1016/j.cpc.2013.09.018. URL https://doi.org/10.1016/j.cpc.2013.09.018.

Utkarsh Upadhyay, Fabrizio Pucci, Julian Herold, and Alexander Schug. Nucleoseeker-precision filtering of rna databases to curate high-quality datasets. *NAR Genomics and Bioinformatics*, 7(1):lqaf021, 2025. doi: 10.1093/nargab/lqaf021. URL http://dx.doi.org/10.1093/nargab/lqaf021.

Utkarsh Upadhyay, Anton Dorn, Christian Faber, and Alexander Schug. From sequence to structure: A comprehensive review of deep learning models for rna structure prediction. *Current Opinion in Structural Biology*, 97:103216, 2026.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

Lingle Wang, Richard A. Friesner, and B. J. Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (rest2). *The Journal of Physical Chemistry B*, 115(30):9431–9438, 2011. doi: 10.1021/jp204407d. URL https://doi.org/10.1021/jp204407d.

A Appendix

A.1 Units used in mdRNAcat

We note that mdRNAcat intentionally uses different unit conventions across data fields. This choice primarily reflects compatibility with mdCATH Mirarchi et al. (2024) and with the native output conventions of the simulation software. Coordinates and forces follow the mdCATH convention and are reported in Å and kcal mol^{-1} $\mathrm{Å}^{-1}$, respectively, whereas replica-exchange quantities retain the corresponding GROMACS conventions, such as kJ mol^{-1} for energies. The unit of every field is explicitly specified in the dataset schema. Quantities representing the same physical observable can be converted between these conventions by standard unit conversions, which are readily supported by common molecular-simulation libraries.

A.2 RNA System in mdRNAcat

In this section, we list all systems that were considered for the mdRNAcat as a result of the initial Nucleoseeker-query in Table 4. Additionally the 5 NMR-resolved systems were added. The table contains information on the simulation length of the systems used in mdRNAcat and information on the exclusion reason if not.

Table 4: RNA systems considered or used in mdRNAcat.

PDB ID	Residues	Simulation length	Status
1duh	44	200 ns	Part of mdRNAcat
1ehz	76		Modified residues not parameterized
1hr2	158		Too large
1kh6	53		Modified residues not parameterized
1kxk	70	300 ns	Part of mdRNAcat
1nbs	155		Missing residues in the pdb
1u9s	161		Missing residues in the pdb
1vtq	75		Modified residues not parameterized
1xjr	47	300 ns	Part of mdRNAcat
1z43	101		System size out of scope
2h0s	125		Missing residues in the pdb
2il9	142		Missing residues in the pdb
2m8k	48	200 ns	Part of mdRNAcat; Added from NMR structure
2mhi	53	200 ns	Part of mdRNAcat; Added from NMR structure
2mqt	68	200 ns	Part of mdRNAcat; Added from NMR structure
2qus	69	200 ns	Part of mdRNAcat
2qwy	52	300 ns	Part of mdRNAcat
3cw5	77		Modified residues not parameterized
3d2v	77	200 ns	Part of mdRNAcat
3e5c	53	200 ns	Part of mdRNAcat
3egz	65		In scope; not prioritized
3npq	54		Missing residues in the pdb
3oxj	88		System size out of scope
3pdr	161		System size out of scope
3q3z	75	200 ns	Part of mdRNAcat
3r4f	66	200 ns	Part of mdRNAcat
3rg5	86		In scope; not prioritized
3slq	68	200 ns	Part of mdRNAcat
3suh	101		System size out of scope
4enc	52	200 ns	Part of mdRNAcat
4fe5	67	200 ns	Part of mdRNAcat
4frg	84	200 ns	Part of mdRNAcat
4gxy	172		Missing residues in the pdb

continued on next page

Table 4 – continued

PDB ID	Residues	Simulation length	Status
4jf2	77	200 ns	Part of mdRNAcat
4jrc	57		In scope; not prioritized
4k27	55	200 ns	Part of mdRNAcat
4kqy	119		System size out of scope
4lvw	89	200 ns	Part of mdRNAcat
4oji	54		Modified residues not parameterized
4p5j	86		System size out of scope
4pqv	68		System size out of scope
4qln	125		Missing residues in the pdb
4r4p	186		System size out of scope
4rum	94		Missing residues in the pdb
4tzx	71	200 ns	Part of mdRNAcat
4wfl	107		System size out of scope
4xwf	64	200 ns	Part of mdRNAcat
4y1m	107		System size out of scope
4yaz	84		In scope; not prioritized
4zc7	92		System size out of scope
5btp	75		Missing residues in the pdb
5dqk	48		System size out of scope
5dun	54		Modified residues not parameterized
5fjc	93	200 ns	Part of mdRNAcat
5lyu	58	200 ns	Part of mdRNAcat
5m0h	42	200 ns	Part of mdRNAcat
5ml7	96		System size out of scope
5npm	80		System size out of scope
5nwq	38	200 ns	Part of mdRNAcat
5ob3	69	200 ns	Part of mdRNAcat
5t83	90		Missing residues in the pdb
5u3g	85		System size out of scope
5une	47	200 ns	Part of mdRNAcat
6cb3	101		System size out of scope
6ck5	117		Missing residues in the pdb
6cu1	81	200 ns	Part of mdRNAcat
6d3p	45	200 ns	Part of mdRNAcat
6dmc	102		System size out of scope
6dvk	95		System size out of scope
6fz0	53		Missing residues in the pdb
6gyv	192		Missing residues in the pdb
6jq5	82		System size out of scope
6lax	55	200 ns	Part of mdRNAcat
6n2v	99		System size out of scope
6n5o	126		System size out of scope
6ol3	112		Missing residues in the pdb
6p2h	69	200 ns	Part of mdRNAcat
6prv	58	200 ns	Part of mdRNAcat
6qn3	50	200 ns	Part of mdRNAcat
6tf0	52	200 ns	Part of mdRNAcat
6ugg	77	200 ns	Part of mdRNAcat
6vmy	148		System size out of scope
6wjr	112		System size out of scope
6xkn	101		Missing residues in the pdb
7d7v	57	200 ns	Part of mdRNAcat
7d81	50	200 ns	Part of mdRNAcat
7elq	45	200 ns	Part of mdRNAcat
7eog	48	200 ns	Part of mdRNAcat

continued on next page

Table 4 – continued

PDB ID	Residues	Simulation length	Status
7jju	51	200 ns	Part of mdRNAcat
7jrr	51		System size out of scope
7jrs	129		System size out of scope
7jrt	134		System size out of scope
7k16	51	200 ns	Part of mdRNAcat
7kga	90		System size out of scope
7mky	66		Modified residues not parameterized
7mlw	128		Missing residues in the pdb
7oax	52	200 ns	Part of mdRNAcat
7qr3	69	200 ns	Part of mdRNAcat
7uq6	86		In scope; not prioritized
7v9e	68		Modified residues not parameterized
7vnv	78		Modified residues not parameterized
7wii	50	200 ns	Part of mdRNAcat
8eyu	87	200 ns	Part of mdRNAcat
8fcs	71	200 ns	Part of mdRNAcat; Added from NMR structure
8hb8	55		In scope; not prioritized
8hzl	84		System size out of scope
8its	46		System size out of scope
8uiw	128		Missing residues in the pdb
8uo6	134		System size out of scope
9bun	48	200 ns	Part of mdRNAcat
9obm	73	200 ns	Part of mdRNAcat; Added from NMR structure

A.3 Sequence and Structural-Motif Coverage

Reference populations. The reference search included RNA polymer entities from X-ray structures released by 25 September 2026, with sequence lengths of 40–192 nucleotides and reported resolution ≤ 3.6 Å. RNA entities in complexes were eligible. Sequences were obtained from the PDB canonical sequence field where available, with thymine converted to uracil. Sequences containing remaining non-ACGU symbols were excluded. Low-complexity filtering excluded sequences satisfying both a uracil fraction ≥ 0.65 and a consecutive uracil run of at least 16 nucleotides. These restrictions applied to the reference population; representatives from all 48 mdRNAcat PDB IDs were retained irrespective of these criteria.

The sequence analysis used 644 distinct sequences: 48 mdRNAcat representatives and 596 evaluation sequences. Structural descriptors were available for 820 retained rows, comprising 51 selected rows across all 48 mdRNAcat PDB IDs and 769 non-selected rows. Sequence evaluation therefore weights distinct sequences, whereas structural evaluation weights retained entity rows. The representations describe deposited sequences and structures, which may differ from edited simulation inputs.

Locally normalized coverage. For each representation, let $\mathcal{R}$ denote the combined reference and representative population, $\mathcal{M}$ the mdRNAcat representatives, and $\mathcal{P}$ the non-selected evaluation population. For $i \in \mathcal{P}$, define

$$d_i^{\mathrm{rep}} = \min_{m \in \mathcal{M}} d(i,m), \qquad s_i = \mathrm{median}_{j \in \mathcal{N}_{10}(i;\mathcal{R})}\, d(i,j), \qquad r_i = \frac{d_i^{\mathrm{rep}}}{s_i}, \tag{1}$$

where $\mathcal{N}_{10}(i;\mathcal{R})$ contains the ten nearest non-self neighbours in the combined population. Representatives are excluded from evaluation but remain eligible neighbours. The implementation excludes nonfinite ratios, including those arising from zero local scales. Writing $\mathcal{P}^*$ for the remaining evaluation rows, coverage is

$$C(t) = \frac{1}{|\mathcal{P}^*|} \sum_{i \in \mathcal{P}^*} \mathbf{1}[r_i \leq t]. \tag{2}$$

Coverage thus refers to evaluable rows rather than necessarily the complete reference population. Pointwise 95% confidence intervals are the 2.5th and 97.5th percentiles from 300 bootstrap resamples of the fixed evaluation ratios; feature fitting and neighbour distances are not recomputed during resampling.

Sequence representation. Contiguous nucleotide 3-mer counts were transformed using TF–IDF with smoothed inverse-document-frequency weights and row-wise ℓ_2 normalization. The resulting vectors were projected using truncated SVD, retaining $\min(80, \max(8, V-1))$ components, where V is the observed vocabulary size. For a complete canonical 3-mer vocabulary, this gives 63 components. Distances were Euclidean in this SVD representation. TF–IDF and SVD were fitted to the combined population.

Structural-motif representation. RNA 3D Hub loop annotations provided 13 descriptors: hairpin, internal-loop and junction count densities; three corresponding residue fractions; the mean and standard deviation of motif lengths for each class; and maximum junction order. Pseudoknot, stem and unpaired-residue descriptors were not included in the retained feature set.

In the current converter, chain length denotes the number of distinct residues occurring in downloaded loop annotations, rather than full polymer length. Residue fractions use this annotated-residue denominator. For chains mapped to the same entity, motif counts are summed, annotated lengths and residue fractions are aggregated by their maxima, and motif-length statistics by their means. Count densities divide the aggregated counts by the aggregated annotated length. Consequently, these descriptors can depend on chain multiplicity and annotation availability.

Missing descriptor values were replaced by feature medians. The imputed features were standardized using means and standard deviations fitted to the combined structural population, and Euclidean distances were used for coverage. A sensitivity analysis additionally used PCA-whitened features retaining at least 95% of standardized variance.

Descriptor shifts and visualization. For descriptor k, the plotted shift is

$$\Delta_k = \frac{\mu_{k,\mathcal{M}} - \mu_{k,\mathcal{P}}}{\sigma_{k,\mathcal{P}}}, \tag{3}$$

using available, non-imputed descriptor values and the population standard deviation of non-selected rows. Selected entity rows are weighted equally; this is not a one-vector-per-PDB comparison.

UMAP projections are used only for visualization. Background surfaces smooth the precomputed distance ratios in the embedding plane and use separate color scales. Structural ratios are clipped at the 4th and 92nd percentiles for display only. Neither this clipping nor UMAP affects the coverage curves. The analysis characterizes deposited structural motifs, not conformational coverage of the REST2 ensembles.

A.4 Simulation Parameters

This section contains the GROMACS parameters used for each step done in the pre-processing and simulations for mdRNAcat. All simulations were prepared and executed with the same protocols apart from simulation length. For the pre-processing, firstly, energy minimization was done with the parameters from Table 5. Then NVT equilibration with Table 6 and lastly NPT equilibration with Table 7 were done. For the subsequent simulations, all parameters can be found in Table 8.

Table 5: Parameters used for energy minimization.

Parameter	Value
integrator	steep
nsteps	100000
emtol	10
emstep	0.01
nstlog	1000
nstenergy	100
energygrps	System
cutoff-scheme	Verlet
nstlist	1
rlist	1
pbc	xyz
coulombtype	PME
pme_order	4
vdwtype	Cut-off
vdw-modifier	Potential-shift-Verlet
rcoulomb	1.2
rvdw	1.2
epsilon_r	1
fourierspacing	0.12

Table 6: Parameters used in the NVT equilibration.

Parameter	Value
define	-DPOSRES
integrator	md
nsteps	50000
dt	0.002
nstxout	500
nstvout	500
nstenergy	500
nstlog	500
continuation	no
constraint_algorithm	lincs
constraints	h-bonds
lincs_iter	1
lincs_order	4
cutoff-scheme	Verlet
nstlist	10
rcoulomb	1.2
rvdw	1.2
DispCorr	EnerPres
coulombtype	PME
pme_order	4
fourierspacing	0.12
tcoupl	Nose-Hoover
tc-grps	RNA Water Ion
tau_t	0.5 0.5 0.5
ref_t	300 300 300
pcoupl	no
pbc	xyz
gen_vel	yes
gen_temp	300
gen_seed	-1

Table 7: Parameters used for NPT equilibration.

Parameter	Value
define	-DPOSRES
integrator	md
nsteps	50000
dt	0.002
nstxout	500
nstvout	500
nstenergy	500
nstlog	500
continuation	yes
constraint_algorithm	lincs
constraints	h-bonds
lincs_iter	1
lincs_order	4
cutoff-scheme	Verlet
nstlist	10
rcoulomb	1.2
rvdw	1.2
DispCorr	EnerPres
coulombtype	PME
pme_order	4
fourierspacing	0.12
tcoupl	Nose-Hoover
tc-grps	RNA Water Ion
tau_t	0.5 0.5 0.5
ref_t	300 300 300
pcoupl	Parrinello-Rahman
pcoupltype	semiisotropic
tau_p	2.0
ref_p	1.0 1.0
compressibility	4.5e-5 4.5e-5
refcoord_scaling	com
pbc	xyz
gen_vel	no

Table 8: Parameters used for production simulation runs.

Parameter	Value
integrator	md
nsteps	100000000
dt	0.002
nstxout	250
nstvout	250
nstfout	250
nstenergy	250
nstlog	5000
nstcalcenergy	250
energygrps	RNA Water Ion
continuation	yes
constraint_algorithm	lincs
constraints	h-bonds
lincs_iter	1
lincs_order	4
cutoff-scheme	Verlet
nstlist	10
coulombtype	PME
pme_order	4
rlist	1.2
rcoulomb	1.2
rvdw	1.2
fourierspacing	0.12
tcoupl	Nose-Hoover
tc-grps	RNA Water Ion
tau_t	0.5 0.5 0.5
ref_t	300 300 300
pcoupl	Parrinello-Rahman
pcoupltype	isotropic
tau_p	2.0
ref_p	1.0 1.0
compressibility	4.5e-5 4.5e-5
pbc	xyz
DispCorr	EnerPres
gen_vel	no
nstcomm	100
comm-grps	RNA Water Ion

Table 9: REST2 effective-temperature ladder and λ scaling factors.

Replica	T_i [K]	λ	$\sqrt{\lambda}$
1	300.000000	1.000000	1.000000
2	304.980000	0.983700	0.991817
3	310.050000	0.967600	0.983667
4	315.200000	0.951800	0.975602
5	320.440000	0.936200	0.967574
6	325.760000	0.920900	0.959635
7	331.180000	0.905900	0.951788
8	336.680000	0.891100	0.943981
9	342.270000	0.876500	0.936216
10	347.960000	0.862200	0.928547
11	353.740000	0.848100	0.920923
12	359.620000	0.834200	0.913345
13	365.590000	0.820600	0.905870
14	371.670000	0.807200	0.898443
15	377.840000	0.794000	0.891067
16	384.120000	0.781000	0.883742
17	390.500000	0.768200	0.876470
18	396.990000	0.755700	0.869310
19	403.590000	0.743300	0.862148
20	410.290000	0.731200	0.855102
21	417.110000	0.719200	0.848057
22	424.040000	0.707500	0.841130
23	431.080000	0.695900	0.834206
24	438.250000	0.684500	0.827345
25	445.530000	0.673400	0.820610
26	452.930000	0.662400	0.813880
27	460.460000	0.651500	0.807155
28	468.110000	0.640900	0.800562
29	475.880000	0.630400	0.793977
30	483.790000	0.620100	0.787464
31	491.830000	0.610000	0.781025
32	500.000000	0.600000	0.774597

A.5 Energy Overlap Plots

This section contains the energy overlap plot for all 48 simulations that were done for mdRNAcat in Figures 8 and 9.

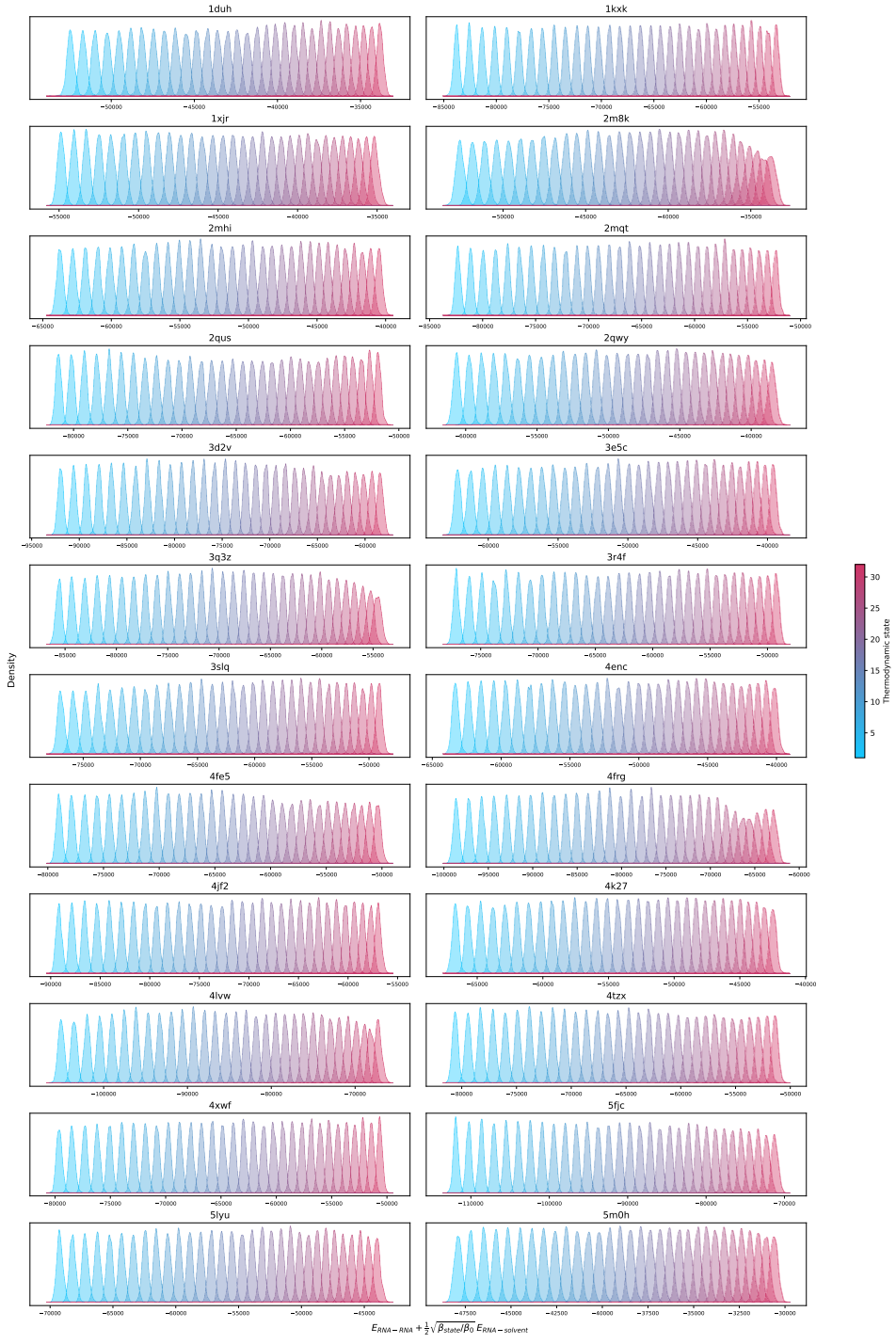

Figure 8: Density of the thermodynamic states against the energy of the first 24 systems in mdRNAcat.

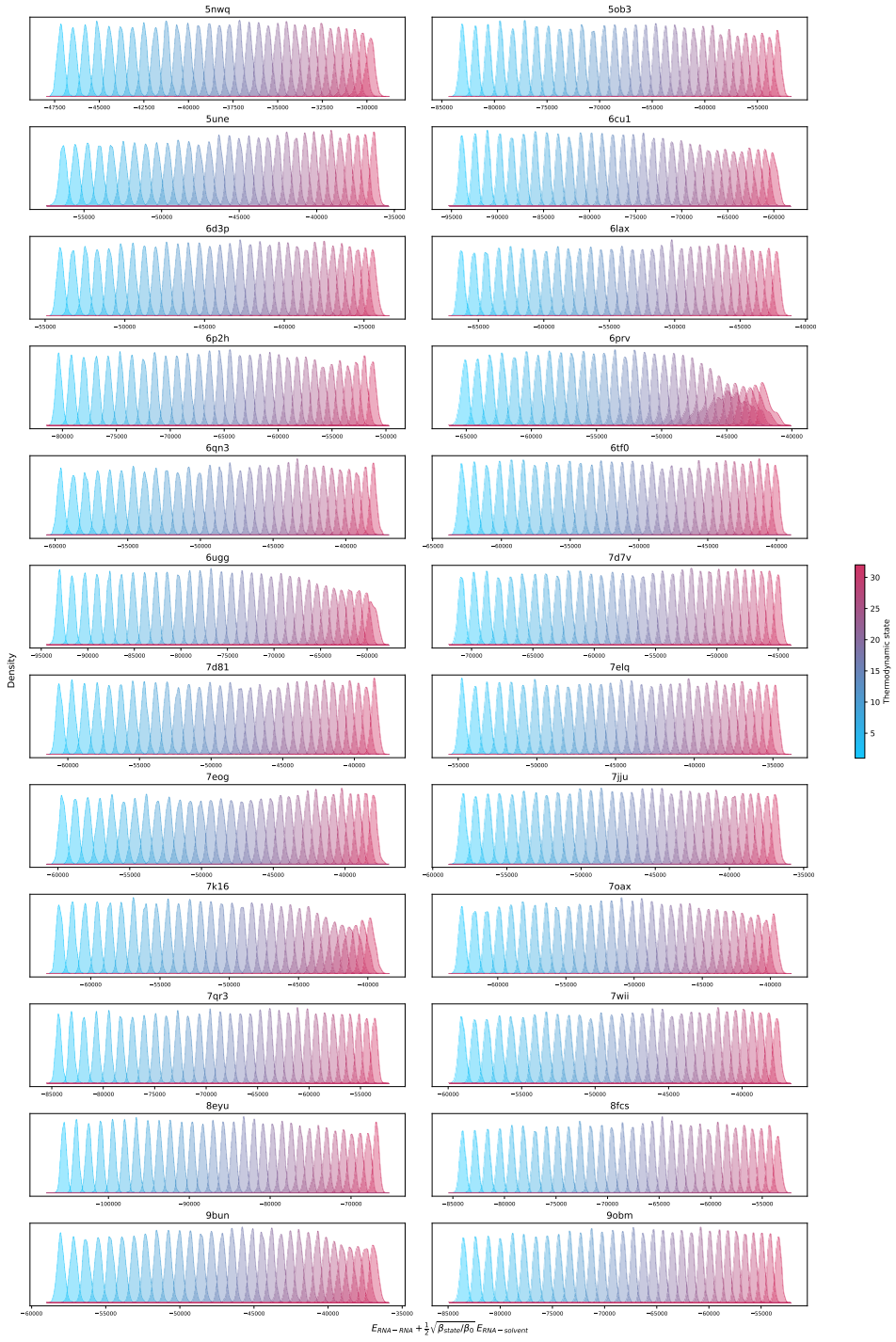

Figure 9: Density of the thermodynamic states against the energy of the second 24 systems in mdRNAcat.