ABSTRACT

Title of Dissertation: PHYSICS-INSPIRED GENERATIVE DEEP LEARNING WITH APPLICATIONS TO RNA STRUCTURE PREDICTION
Lukas Herron
Doctor of Philosophy, 2026

Dissertation Directed by: Professor Pratyush Tiwary
Department of Chemistry and Biochemistry

Biomolecular structure is often treated as a single native conformation, but many molecular systems are more accurately described as thermodynamic ensembles of interconverting structures. This thesis develops a probabilistic generative modeling framework for learning such ensembles from limited molecular data. The central premise is that generative models are most effective for molecular science when physical inductive biases are included.

The thesis begins by benchmarking existing classes of probabilistic generative models on molecular data and distributions with tunable dimensionality. These comparisons identify denoising diffusion probabilistic models as especially capable for the moderate-dimensional, multimodal landscapes common in molecular systems. Building on that result, the next study shows that diffusion models trained on replica exchange molecular dynamics data can extrapolate to temperatures not present in the training set.

The core methodological contribution, the Thermodynamic Maps framework, represents physical temperature in a diffusion model. This adjustment enables precise thermodynamic extrapolation from sparse data and is validated on both the Ising model and RNA conformational sampling. The framework is then extended to *ab initio* RNA ensemble prediction in RNAnneal, where hierarchical structure generation, thermodynamic scoring, and physics-grounded ranking are

combined to recover three-dimensional RNA conformational ensembles from sequence alone. A complementary application, DiffDock-Glide, demonstrates that a similar physics- and data-driven pipeline improves molecular docking and virtual screening outcomes. Together, these studies demonstrate that generative models can extrapolate from highly limited data when provided physical and chemical inductive biases.

PHYSICS-INSPIRED GENERATIVE DEEP LEARNING WITH APPLICATIONS TO RNA STRUCTURE PREDICTION

by

Lukas Herron

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:

Professor Pratyush Tiwary, Chair
Professor Christopher Jarzynski
Professor Yifei Mo
Professor Jeffery Klauda
Dr. John Schneekloth

© Copyright by

Lukas Herron

2026

Acknowledgments

Many people made this thesis possible, and I am grateful to all them for supporting me on my journey of scientific development. They say that it takes a village to raise a child; well it certainly takes a community to raise a scientist.

I would like to thank my advisor, Pratyush Tiwary, for his guidance and support throughout my doctoral studies. I am grateful to him and to the members of the Tiwary group, past and present, for the many stimulating conversations and discussions that shaped the ideas in this thesis. Similarly, I thank Jay Schneekloth for supporting my interest in all thing RNA - a direction I couldn't have imagined myself taking at the start of my degree.

I thank my parents, grandparents, and sister for the love, support and encouragement they have given me, from long before my doctoral degree began all the way to the present. I would also like thank Michelle for sticking with, supporting me through these years, and being a sounding board for all aspects of my life. Finally, the many long walk I have taken with our dog, Spot, gave me much needed space to think. Sometimes about science and sometimes not. For that I am grateful.

Finally, I thank the mentors who first set me on this path: Purushottam Dixit, Bing-Kan Xue, and Pablo Sartori. Each of them opened my eyes to computational science and statistical physics, and their early guidance is a large part of the reason I found my way to this work.

Table of Contents

Acknowledgments ii

Table of Contents iii

List of Tables vi

List of Figures vii

List of Abbreviations ix

List of Publications x

Chapter 1: Generative Modeling in the Biomolecular Sciences 1
1.1 Biomolecular structure as a problem of distributions . . . 1
1.2 Sampling equilibrium states of physical systems . . . 3
1.3 Classical approaches to estimating partition functions and free energies . . . 5
1.4 Generative models for prediction of molecular ensembles . . . 7
1.5 RNA: A perfect storm of conformational heterogeneity, limited data, and environmental nuance . . . 9
1.6 The role of physical structure in generative modeling for molecular science . . . 10
1.7 Thesis contributions and organization . . . 12

Chapter 2: A Comparison of Probabilistic Generative Frameworks for Molecular Simulations 14
2.1 Introduction . . . 14
2.2 Theoretical background . . . 16
2.3 Probabilistic Generative Frameworks . . . 17
2.3.1 Normalizable Architectures . . . 18
2.3.2 Neural Ordinary Differential Equations . . . 19
2.3.3 Diffusion Models . . . 20
2.3.4 Schrödinger Bridges . . . 21
2.4 Experiments . . . 22
2.4.1 Gaussian Mixture . . . 24
2.4.2 Aib_9 Dihedral Torsion Angles . . . 28
2.5 Discussion . . . 31

Chapter 3: From Data to Noise to Data: Mixing Physics Across Temperatures with Generative AI **33**
3.1 Results . . . 36
3.2 Discussion . . . 44

Chapter 4: Inferring Phase Transitions and Critical Exponents from Limited Observations with Thermodynamic Maps **47**
4.1 Targeted Free Energy Perturbation . . . 49
4.2 Leveraging Non-Equilibrium Thermodynamics . . . 51
4.3 Score-Based Models . . . 53
4.4 Thermodynamic Maps . . . 56
4.5 Critical Behavior of the Ising Model . . . 57
4.6 Exploring RNA Conformational Landscapes with Thermodynamic Map-accelerated Molecular Dynamics (TM-aMD) . . . 60
4.7 Discussion . . . 66

Chapter 5: Ab Initio Prediction of RNA Structure Ensembles with RNAnneal **69**
5.1 RNAnneal . . . 72
5.2 Unsupervised sampling and prediction of RNA structures . . . 73
5.3 RNAnneal samples experimentally resolved conformations of pseudoknot-free riboswitches . . . 77
5.4 The RNAnneal score is selective for experimentally resolved conformations . . . 79
5.5 Annealing improves classification and ranking of experimentally resolved conformations . . . 80
5.6 State-informed prediction of conformational ensembles . . . 81
5.7 Quantification of ensemble heterogeneity via per-nucleotide interaction entropy . . . 82
5.8 Evaluation of RNAnneal on experimentally resolved ***holo*** riboswitch conformations 83
5.9 Evaluation of RNAnneal on experimentally resolved ***apo*** riboswitch conformations 84
5.10 ***De novo*** prediction of ***apo*** riboswitch conformations . . . 85
5.11 Discussion . . . 86

Chapter 6: DiffDock-Glide: A Hybrid Physics-Based and Data-Driven Approach to Molecular Docking **88**
6.1 Introduction . . . 88
6.2 Binding site-conditional DiffDock . . . 90
6.3 DiffDock–Glide . . . 92
6.4 Applications . . . 93
6.4.1 Molecular docking . . . 94
6.4.2 Virtual screening . . . 97
6.5 Discussion . . . 103

Chapter 7: Outlook **105**

Appendix A: Supplementary Material for Chapter 2 **107**
A.1 Architectural Details . . . 107

A.2 Data Details . . . 109

Appendix B: Supplementary Material for Chapter 3 **111**
B.1 MD simulations setup . . . 111
B.1.1 AIB_9 . . . 111
B.1.2 GACC . . . 111
B.2 Network structure and hyperparameters . . . 112

Appendix C: Supplementary Material for Chapter 4 **113**
C.1 Network Architecture and Parameters . . . 113
C.2 Training and Sampling . . . 113
C.3 Ising Model . . . 115
C.4 G-vectors . . . 115
C.5 Ensemble Weighted Observables . . . 116
C.6 Rosetta, MD Simulations, and Adaptive Sampling . . . 116
C.7 Local Equilibrium and Linear Fluctuations . . . 117
C.7.1 Motivation . . . 117
C.7.2 Protocol . . . 118
C.8 Convergence to Equilibrium . . . 120

Appendix D: Supplementary Material for Chapter 5 **127**
D.1 Secondary Structure Prediction . . . 127
D.2 FARFAR2 . . . 127
D.3 Clustering . . . 128
D.4 Molecular Dynamics . . . 128
D.5 R- and G-vectors . . . 129
D.6 SPIB . . . 129
D.7 Thermodynamic Maps . . . 130
D.8 RNAnneal Score . . . 131
D.9 Reference Structures . . . 131
D.10 RMSD . . . 131
D.11 INF . . . 132
D.12 AUAC . . . 132
D.13 BEDROC . . . 133
D.14 Interaction Entropy . . . 134
D.15 State Ranking . . . 134

Appendix E: Supplementary Material for Chapter 6 **162**
E.1 Software and Data . . . 162

Bibliography **163**

List of Tables

6.1 Average enrichments for virtual screening against AF2-generated receptors. 100

A.1 Aib_9 MD parameters 110

D.1 Details of riboswitch sequences and reference conformations 154
D.2 Molecular Dynamics Equilibration Procedure 155
D.3 Molecular Dynamics Simulation Parameters. 156

List of Figures

2.1 Model accuracy results for the 4-modal Gaussian mixture dataset. 23
2.2 Model accuracy results for the 50-dimensional Gaussian mixture dataset. 25
2.3 Free energy difference estimation accuracy on asymmetric bimodal distributions. . 27
2.4 Speed and model size results for the 4-modal Gaussian mixture dataset. 29
2.5 Model accuracy results and generated data for the Aib9 peptide. 30

3.1 DDPM neural network architecture. 38
3.2 AIB9 structure and free energy profile. 39
3.3 Samples from REMD and DDPM at 400 K. 41
3.4 Extrapolated free energy differences at 400 K. 43
3.5 GACC samples and free energy profile at 325 K. 44

4.1 Transporting a probability distribution with a diffusion process. 52
4.2 Illustration of a thermodynamic map between systems. 55
4.3 Inferring the phase transition of the 2D Ising model from limited sampling. . . . 58
4.4 GCAA Tetraloop Conformational Landscape. 61
4.5 HIV-TAR RNA Conformational Landscape. 65

5.1 Illustrative overview of RNAnneal. 71
5.2 RNAnneal evaluated on the riboswitch dataset. 75
5.3 Conformational ensemble prediction with RNAnneal. 77

6.1 Performance of DiffDock-Glide on the Posebusters dataset. 93
6.2 Targeting the allosteric site of β-2 Adrenergic receptor with ligand KBY. 98
6.3 Target-specific enrichment for the AF2-DUD-E virtual screening benchmark. . . . 99
6.4 Analysis of poses sampled for IGF-1R 101

A.1 Hyperparameter tuning for each model. 109

C.1 Quality of linear fluctuation fit for the HIV-TAR RNA. 123
C.2 Convergence of melting temperature calculations. 124
C.3 Convergence towards equilibrium for GCAA. 125
C.4 Convergence towards equilibrium for HIV-TAR. 126

D.1 Riboswitch reference ERCs labeled with PDB accession code. 136
D.2 Interaction Network Fidelity (INF) by interaction type. 137
D.3 Evaluation of secondary structures predicted by EternaFold, LinearFold, and LinearAlifold. 138

D.4 Number of clusters vs d. 139
D.5 RMSD variation over a 100ns MD simulation launched from the reference conformation. 140
D.6 RMSD distribution of conformations sampled by RNAnneal. 141
D.7 RMSD distribution of conformations for the classification task. 142
D.8 INF variation over a 100ns MD simulation launched from the reference conformation. 143
D.9 INF distribution of conformations generated by RNAnneal. 144
D.10 INF distribution of conformations for the classification task. 145
D.11 Correlation between RMSD and INF for RNAnneal conformations. 146
D.12 AUAC and BEDROC for each combination of training features. 146
D.13 AUAC and BEDROC for RMSD classification criteria. 147
D.14 Variation in AUAC and BEDROC with scoring temperature. 147
D.15 Accumulation Curves. 148
D.16 Summary of Accumulation Curves 149
D.17 Conformational ensembles predicted by RNAnneal, set 1. 150
D.18 Conformational ensembles predicted by RNAnneal, set 2. 151
D.19 Conformational ensembles predicted by RNAnneal, set 3. 152
D.20 Conformational ensembles predicted by RNAnneal, set 4. 153

List of Abbreviations

AF2	AlphaFold2
AI	Artificial intelligence
AUAC	Area under the accumulation curve
BEDROC	Boltzmann-enhanced discrimination of receiver operating characteristic
DDPM	Denoising diffusion probabilistic model
EF	Enrichment factor
ERC	Experimentally resolved conformation
IE	Interaction entropy
INF	Interaction network fidelity
KL	Kullback–Leibler
MD	Molecular dynamics
ODE	Ordinary differential equation
PCA	Principal component analysis
PDB	Protein Data Bank
PDE	Partial differential equation
PGM	Probabilistic generative model
REMD	Replica exchange molecular dynamics
RNA	Ribonucleic acid
RMSD	Root-mean-square deviation
SDE	Stochastic differential equation
SPIB	State Predictive Information Bottleneck
TFEP	Targeted free-energy perturbation
TM	Thermodynamic Maps

List of Publications

The following manuscripts form the basis of, or are closely associated with, this dissertation.

1. Yihang Wang, **Lukas Herron**, and Pratyush Tiwary. *From Data to Noise to Data for Mixing Physics Across Temperatures with Generative Artificial Intelligence*. *Proceedings of the National Academy of Sciences of the United States of America*, 119(32):e2203656119, 2022.

2. Shams Mehdi, Zachary Smith, **Lukas Herron**, Ziyue Zou, and Pratyush Tiwary. *Enhanced Sampling with Machine Learning*. *Annual Review of Physical Chemistry*, 75(1):347–370, 2024.

3. Pratyush Tiwary, **Lukas Herron**, Richard John, Suemin Lee, Disha Sanwal, and Ruiyu Wang. *Generative AI for Computational Chemistry: A Roadmap to Predicting Emergent Phenomena*. *Proceedings of the National Academy of Sciences of the United States of America*, 122:e2415655121, 2025. DOI: 10.1073/pnas.2415655121.

4. **Lukas Herron**, Kinjal Mondal, John S. Schneekloth Jr., and Pratyush Tiwary. *Inferring Phase Transitions and Critical Exponents from Limited Observations with Thermodynamic Maps*. *Proceedings of the National Academy of Sciences of the United States of America*, 121(52):e2321971121, 2024.

5. Richard John, **Lukas Herron**, and Pratyush Tiwary. *A Comparison of Probabilistic Generative Frameworks for Molecular Simulations*. *The Journal of Chemical Physics*, 162(11), 2025.

6. Suemin Lee, Ruiyu Wang, **Lukas Herron**, and Pratyush Tiwary. *Exponentially Tilted Thermodynamic Maps (expTM): Predicting Phase Transitions Across Temperature, Pressure, and Chemical Potential*. *arXiv preprint arXiv:2503.15080*, 2025.

7. **Lukas Herron**, Jumana Dakka, Steven V. Jerome, Kun Yao, and Da Shi. *DiffDock-Glide: A Hybrid Physics-Based and Data-Driven Approach to Molecular Docking*. *Journal of Chemical Information and Modeling*, published online April 23, 2026. DOI: 10.1021/acs.jcim.5c01635.

8. Yunrui Qiu, Richard John, **Lukas Herron**, and Pratyush Tiwary. *Latent Thermodynamic Flows: Unified Representation Learning and Generative Modeling of Temperature-Dependent Behaviors from Limited Data*. *Chemical Science*, 2026.

9. **Lukas Herron**, Yunrui Qiu, Anjali Verma, Venkata Sai Sreyas Adury, Richard John, Suemin Lee, Shams Mehdi, Disha Sanwal, John S. Schneekloth Jr., and Pratyush Tiwary. *Ab Initio Prediction of RNA Structure Ensembles with RNAnneal*. *bioRxiv*, 2026.02.01.703098, 2026.

10. Pratyush Tiwary and **Lukas Herron**. *Ribonucleic Acid Tertiary Structure Prediction*. U.S. Patent Application No. 19/278,449; Publication No. US 2026/0031181, 2026.

Chapter 1: **Generative Modeling in the Biomolecular Sciences**

1.1 Biomolecular structure as a problem of distributions

Historically, structural biology focused on the native structure—the dominant, stable fold of a biomolecule. At the dawn of the field, this description was apt. The early structures of DNA, myoglobin, and tRNA, solved from the 1950s through the 1970s, were resolved by X-ray diffraction as well-folded structures [1, 2, 3]. As the number of solved molecular structures grew, determining the native structure became the central goal of experimental structure determination and, later, of computational prediction.

Concurrently, theorists sought first-principles descriptions of biomolecular folding. For RNA, dynamic-programming approaches and nearest-neighbor energetic rules showed that secondary structure could be predicted from an explicit decomposition of interaction energies [4, 5]. More generally, energy-landscape and folding-funnel pictures described folding as motion on a thermodynamic surface whose minima correspond to stable structures [6, 7]. Spin-glass-inspired analyses further shed light on how heterogeneous interactions generate ordered, disordered, and metastable phases [6, 8]. The common insight is that many biomolecules have multiple accessible configurations, each with an associated thermodynamic weight, rather than a single, dominant fold.

From the 1980s onward, structural heterogeneity was increasingly confirmed experimentally: methods such as NMR spectroscopy, single-molecule FRET, and cryo-electron microscopy revealed structural variation that is not explained by a single static model of the structure [9, 10, 11]. The goals of molecular modeling expanded beyond identifying one dominant structure, shifting toward characterization of structural ensembles.

The equilibrium probability distribution of statistical physics is the natural description: it assigns each microstate, or configuration, a Boltzmann weight. For a 3D molecular structure, a microstate is a set of atomic coordinates; for a spin model, it is a configuration of spin variables. In either case, the equilibrium distribution takes the form [12]

$$p_{\text{eq}}(\mathbf{x};\beta) = \frac{e^{-\beta U(\mathbf{x})}}{Z(\beta)} \quad \text{where} \quad Z(\beta) = \int e^{-\beta U(\mathbf{x})}\, d\mathbf{x}. \tag{1.1}$$

The numerator in Eq. 1.1 is the Boltzmann weight. It describes how the probability of microstate $\mathbf{x}$ depends on its energy $U(\mathbf{x})$ for a system maintained by an environment at inverse temperature $\beta = 1/k_B T$. At low temperatures, low-energy configurations dominate, while at higher temperatures the distribution flattens and high-energy states become accessible. The conceptual picture is intuitive, but estimating p_{eq} numerically is difficult in all but the most trivial cases. With an accessible energy function and temperature in hand, p_{eq} appears to be within reach. The obstacle is the normalization constant Z, called the partition function, whose evaluation requires an integral over a typically high-dimensional configuration space. Computing $Z(\beta)$ is of interest across physics, biology and chemistry, since the full thermodynamics of a system is encoded in the scalar-to-scalar map, a far simpler object than $p_{\text{eq}}(\mathbf{x};\beta)$. However, practical, computationally tractable methods for obtaining $Z(\beta)$ proved to be elusive due to the so-called sampling problem.

Levinthal's paradox illustrates the severity of the problem: a 100-residue polypeptide has 10^{47} configurations, assuming each residue has 3 conformational states [13]. The astronomical configuration space makes enumeration or brute-force exploration impossible. Analytical methods from statistical mechanics for estimating Z, such as the replica trick, are useful, but they typically require assumptions that are violated by real biomolecular systems [14]. Today's commonly used biomolecular energy functions like AMBER and CHARMM are necessarily complex because they must account for bonded interactions, electrostatics, van der Waals interactions, and solvent effects [15, 16].

Another complication arises when one does not work with a fully atomistic description. In practice, biomolecular structure is often represented in terms of reduced variables or collective

coordinates—for example, selected dihedral angles, radii of gyration, or contact distances. These reduced descriptions still follow equilibrium probability distributions, but the effective energy function that generates them is often not straightforwardly available. In such cases, the distribution must be learned from samples alone, without access to the energy function that generated them.

1.2 Sampling equilibrium states of physical systems

The previous section identified the Boltzmann distribution and the partition function as central objects in statistical mechanics. However, most practical calculations do not begin by evaluating the partition function directly. Instead, they begin by generating samples from the equilibrium distribution. If representative configurations can be obtained, then structural properties, thermodynamic averages, and free-energy differences can be estimated from those samples. The central computational problem is therefore not only how to evaluate the partition function, but how to explore the relevant regions of configuration space in the first place.

One major class of sampling methods is Markov chain Monte Carlo (MCMC). MCMC constructs a sequence of configurations whose stationary distribution is the desired equilibrium distribution. In a typical MCMC algorithm, each step consists of a proposal and an acceptance decision. The proposal step generates a candidate configuration from the current one. The acceptance step then decides whether the chain moves to the proposed configuration or remains where it is. The acceptance rule is chosen so that, over many steps, the chain samples the target distribution [17, 18].

A central advantage of MCMC is that it does not require the normalization constant of the target distribution. Acceptance probabilities depend on ratios of probabilities, so the partition function is never required. This makes MCMC one of the most important tools in computational statistical physics: it allows equilibrium samples to be drawn from distributions with unknown or intractable normalization [17, 18].

At the same time, MCMC dynamics are not usually physical dynamics. The "time" in a Markov chain is algorithmic time: one counts MCMC steps, sweeps, or accepted moves, not

seconds, minutes, or hours. A Markov chain may move between configurations in ways that are statistically valid but physically unrealistic. For this reason, MCMC is naturally suited to estimating equilibrium averages, but it does not generally provide kinetic information [19].

The practical difficulty in MCMC is in designing the proposal step. Small local moves, such as flipping one spin or slightly perturbing one torsion angle, tend to be accepted often, but explore configuration space slowly. Large global moves can decorrelate samples more efficiently, but they are also more likely to propose high-energy configurations and be rejected. An effective MCMC implementation must simultaneously maintain high acceptance probabilities while exploring the configuration space efficiently.

A second major class of sampling methods is molecular dynamics (MD). Rather than proposing discrete transitions between configurations, MD simulates microscopic equations of motion. Given an energy function, forces are computed from the energy landscape and used to propagate the system forward in time [19, 20]. In atomistic molecular simulations, the coordinates are atomic positions and the energy function is supplied by a molecular mechanics force field [15, 16]. With appropriate thermostats and barostats, MD simulations can be configured to sample particular thermodynamic ensembles, including the canonical and isothermal-isobaric ensembles [21, 22, 23, 24].

Like MCMC, performing MD simulations does not require the partition function directly. But unlike MCMC, MD simulations produce trajectories with an associated physical time variable. This makes them especially useful for the study of biomolecules, where the same simulation may provide both equilibrium structural information and insight into conformational dynamics. A trajectory of a protein, RNA, or ligand-receptor complex can, for instance, reveal which conformations are thermodynamically accessible and how the system transitions between them [19].

The cost is that MD must resolve the fastest motions in the system. Bond vibrations and other high-frequency modes restrict the integration timestep to femtoseconds. At a timestep of one femtosecond, a microsecond of simulation requires one billion integration steps. A microsecond may be enough to observe a fast conformational rearrangement, but many biologically relevant

events occur on longer timescales of seconds, hours, and days [19].

MCMC and MD both face similar sampling limitations. Inefficient proposals in MCMC lead to rejection or slow mixing, while MD trajectories may remain trapped in local basins because barrier-crossing events are rare. In either case, the total number of sampled configurations can be misleading, since they are not statistically independent.

These limitations motivate enhanced sampling and free-energy methods. Some approaches modify the temperature or ensemble, some bias the dynamics along selected collective variables, and others reweight samples after they have been collected [25, 26, 27, 28, 29, 30, 31, 32, 33]. Despite their differences, they all address the same challenge of sampling high-dimensional distributions. The next section summarizes classical approaches developed to overcome this sampling problem and estimate partition functions or free-energy differences in realistic systems.

1.3 Classical approaches to estimating partition functions and free energies

The difficulty introduced by Eq. 1.1 is not new. Much of computational statistical mechanics has been devoted to estimating partition functions, ratios of partition functions, or the free-energy differences derived from them. Since the absolute value of $Z(\beta)$ is rarely accessible for realistic molecular systems, most practical methods target relative quantities. For example, the free energy $F(\beta) = -\beta^{-1} \log Z(\beta)$ is discussed only in terms of free-energy difference between thermodynamic or conformational states. Classical approaches tend to focus either on improving the exploration of configuration space or extracting more accurate estimates of free-energy differences from samples that have already been collected.

Free-energy perturbation, Bennett acceptance ratio, thermodynamic integration, and targeted free-energy perturbation are the longest-standing route to computing free-energy differences [34, 35, 36, 37]. These methods apply when the relevant energy functions are known and when samples from one thermodynamic or conformational state overlap with the other. In practice, the overlap

requirement is the central difficulty. Exponential averages can be dominated by rare samples, and free-energy estimates can converge slowly when the reference and target distributions differ substantially.

One family of methods improves sampling by changing the temperature or ensemble of the system. Simulated annealing raises and then lowers the temperature to help a system escape local minima [25]. Simulated tempering and replica exchange molecular dynamics extend this idea by allowing configurations to move among multiple temperatures or thermodynamic ensembles [26, 27]. These *tempering* methods accelerate sampling by allowing high-temperature replicas to cross barriers, while maintaining low-temperature simulations to recover the free energy at a temperature of interest. However, the replicas must have similar energies for temperature exchanges to occur frequently enough to accelerate sampling. For large solvated biomolecules, maintaining this overlap can require many closely spaced replicas, which increases computational cost and limits scalability.

A second class of methods biases the dynamics along selected collective variables. Umbrella sampling, metadynamics, accelerated molecular dynamics, and related approaches modify the energy landscape so that rare transitions become frequent [28, 29, 30]. When the relevant slow coordinates are known, these methods can efficiently reconstruct free-energy profiles along those variables. The limitation is that the choice of collective variables is a strong modeling assumption. Many biomolecular systems do not evolve along one or two obviously meaningful coordinates. RNA conformational landscapes, in particular, can involve many competing base-pairing, stacking, tertiary-contact, and solvent-mediated interactions. In such cases, biasing a small set of variables may accelerate some motions while leaving other important barriers unresolved.

A final set of methods reconstructs thermodynamic information by reweighting samples after simulation. Weighted histogram analysis, multistate Bennett acceptance ratio, and related histogram-reweighting methods combine data from biased simulations or multiple thermodynamic states to estimate free-energy differences and, indirectly, ratios of partition functions [31, 32, 33]. These approaches are especially valuable when samples have been collected at different temperatures or

under different biasing potentials. Their accuracy, however, depends on sufficient overlap between the sampled distributions. If important regions of configuration space are missing, post-processing cannot recover their statistical weights.

The breadth of these approaches reflects the difficulty of the problem they aim to solve. Biomolecular free-energy landscapes are high-dimensional, rugged, and often only partially observed by experiment. Some methods require carefully chosen collective variables, while others require energy evaluations or overlap among sampled ensembles, and still others become expensive as system size increases. There is therefore no universal classical solution to estimating $Z(\beta)$, $\log Z(\beta)$, or the equilibrium distribution itself. This motivates the modern generative-model perspective: instead of selecting a small number of coordinates or relying only on reweighting, one can attempt to learn a map between a simple prior distribution and the complex molecular ensemble. The next section introduces this class of models and explains their utility for predicting molecular ensembles.

1.4 Generative models for prediction of molecular ensembles

The previous sections identify the equilibrium distribution, and the partition function that normalizes it, as central objects in biomolecular structure prediction. Classical methods improve sampling or estimate free-energy differences under specific assumptions, but they do not eliminate the need to model high-dimensional, multimodal distributions from sparse data. This problem has also attracted sustained interest in machine learning, where probabilistic generative models were developed to learn complicated probability distributions directly from data.

Probabilistic generative models (PGMs) begin with access to configurations drawn from an intractable target distribution. These may be data generated by experiment or simulation. A PGM learns a transformation between the samples from the target distribution and a simple prior distribution, conventionally a Gaussian. Generative models parameterize this transformation in many ways, but the end goal is always the same: to generate samples from the intractable target by

sampling the simple prior and to evaluate the probability of data under the target distribution.

Normalizing flows were among the earliest successful realizations of this idea [38]. They build the generative map as a composition of invertible transformations, so that the likelihood of a sample can be computed exactly through the change-of-variables formula. Exact evaluation of the likelihood is important because it permits training by likelihood maximization. However, normalizing flows are limited in expressivity, since the invertibility and likelihood calculation are made possible by restricting the mathematical operations the network performs. Moreover, it was discovered empirically that normalizing flows struggle to transform between a multimodal target distribution and a unimodal prior.

A major advance was to move the invertibility condition from the network architecture to an ordinary differential equation. In continuous normalizing flows, a neural network learns a velocity field that transports particles from Gaussian-distributed initial positions to the target distribution at a final time [39]. Importantly, the likelihood remains accessible by integrating the divergence of that velocity field along the trajectory of the generative process. This continuous-time formulation is more flexible than discrete flows, but still struggles to generate multimodal distributions.

Diffusion models address this difficulty by employing a stochastic generative process. Inspired by non-equilibrium thermodynamics, they define a forward process that progressively destroys structure by adding noise to data, and a reverse process that reconstructs data by removing noise [40, 41, 42]. Because the forward process is stochastic, the learned reverse process can reconstruct disconnected metastable states without the mode-separation problems that limit deterministic models. The broader continuous-time literature now connects diffusion models, flow-matching variants, and related methods through a shared ODE–SDE–PDE perspective on probabilistic transport.

These developments have directly impacted structural biology. AlphaFold 3, for example, incorporates a diffusion model as part of a larger AI system. The success of AlphaFold 3 and contemporary models revealed another limitation. In many problems in molecular science, the main bottleneck is no longer the expressive power of the architecture, but the amount and quality of the available data. Generative models do not extrapolate reliably to entirely new classes or modes; they

are only as good as the data on which they are trained. Protein structure prediction benefited from an unusually rich data regime. Other biomolecular structure-prediction problems, such as RNA structure prediction, have far less training data. Progress with limited high-quality molecular data therefore depends on inductive biases that promote generalization in AI and deep-learning models.

1.5 RNA: A perfect storm of conformational heterogeneity, limited data, and environmental nuance

RNA structure prediction is a recurring topic throughout this thesis because it lies near the limit of what a purely data-driven generative model can be expected to learn. The difficulty is not only that fewer RNA structures are available than protein structures, but that the missing data are a consequence of the physics of RNA itself. RNA molecules are often more conformationally flexible than their protein counterparts: the ribose-phosphate backbone has six rotatable degrees of freedom per nucleotide. The large state space conferred by the flexible backbone is stabilized by base-pairing and -stacking interactions that generally give rise to multiple alternative states. In contrast, the protein backbone is frequently summarized by two main backbone dihedral angles per residue. This additional structural freedom makes RNA ensembles broader, more heterogeneous, and more difficult to summarize by a single representative conformation [43, 44].

The same flexibility that makes RNA functionally rich also contributes to the scarcity of experimentally resolved structures. High-resolution structure-resolving methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy are most straightforward to interpret when a molecule occupies a relatively coherent structural state. For flexible RNAs, the measured signal can instead be a superposition of conformations. Rather than resolving one clean structural model, the experiment may report an average, a partial model, or no high-resolution structure at all. The result is a limited number of solved structures that are biased toward RNAs that are experimentally tractable [45, 46].

RNA structure is also sensitive to environmental context. Ions such as sodium, potassium, and

magnesium can reshape the folding landscape by neutralizing the negatively charged backbone or by stabilizing specific tertiary contacts. Some RNAs deliberately use this environmental sensitivity as the basis of function. Riboswitches regulate gene expression by adopting different structural states in response to metabolites or transition metals. Viral RNAs likewise encode structural elements whose stability and function depend on host context; dengue virus RNA elements, for example, have been studied at mosquito and mammalian temperatures. More generally, both coding and noncoding RNAs are frequently bound by protein partners or other nucleic acids, and the presence of those partners can change the relevant molecular structure [47, 48, 49].

For these reasons, RNA is not merely another version of protein structure prediction. It is a problem for which limited data, ensemble descriptions, and environmental factors compound one another. A model that succeeds only by interpolating among known structures is unlikely to be sufficient, because entire classes of RNA folds are weakly represented or absent from available structural databases. The RNA3DB dataset, for example, was developed to address the sparse and structurally biased nature of RNA structures in the PDB and reports fewer than 2,000 non-redundant examples of RNA structures–two orders of magnitude fewer than are available for proteins [50].

RNA will therefore be used repeatedly in this thesis as a biomolecular system for evaluating generative models. The goal is not only to generate plausible coordinates, but to learn distributions over structures and how they respond to the environment from sparse data. This motivates the central theme of the thesis: generative models for molecular science require physical and chemical inductive biases, especially when the target system is flexible, data-poor, and ensemble in nature.

1.6 The role of physical structure in generative modeling for molecular science

Inductive biases are structural assumptions built into a model that reduce its solution space and, by extension, the amount of data required for learning [51]. In molecular machine learning, these biases take several forms. Some are architectural: translational, rotational, and Euclidean

equivariances can be built directly into the network [52]. Others enter through the objective function, such as the addition of physics-informed loss terms that penalize violations of known constraints [53]. Probabilistic generative models offer a third route: physical structure can be placed directly into the prior distribution or into the dynamical process that transports between the prior and target distribution.

Boltzmann generators are an early example of this strategy [54]. They learn a map between a simple reference distribution and a molecular Boltzmann distribution while retaining enough structure for reweighting and free-energy estimation. In one reported application, the authors scaled the variance of the prior distribution depending on the temperature at which the training sample was obtained. This enabled the model to generate data at multiple temperatures and reweight samples across temperatures. However, Boltzmann generators require that the energy function be readily available, which limits their application to the full atomistic configuration space. Reduced descriptions instead require a generative model that learns directly from temperature-dependent data, without access to the energy function.

Extrapolation to new environmental conditions cannot be expected to occur solely on the basis of structural data. AlphaFold2 and subsequent AI structure-prediction models are trained largely on structures obtained by X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy [45, 55, 56]. When such experiments succeed, their output is a molecular structure or structural ensemble corresponding to a particular set of environmental conditions. Structure, however, is linked to the environment: ion concentrations, protein or nucleic-acid partners, cofactors, metabolites, temperature, pressure, and pH can all alter the relevant ensemble. A static structural database offers only glimpses into the full relationship between molecular identity, environment, and ensemble [47, 48, 57].

This limitation cannot be overcome simply by collecting the same type of data at every condition of interest. It took decades to assemble the current Protein Data Bank, and it is still sparse for RNAs and heterogeneous biomolecular assemblies [45, 50]. Repeating high-resolution experiments across every relevant temperature, pressure, ionic condition, ligand concentration,

and binding-partner composition would be impractical. Available structures reflect only the set of conditions and limited thermodynamic contexts in which the structure could be resolved.

Statistical mechanics provides a way to approach this problem. The equilibrium distribution describes how temperature changes the weights of configurations in an ensemble. Responses to other thermodynamic variables can be described by introducing the corresponding conjugate variables into the thermodynamic potential. For example, pressure couples to volume, and chemical potential couples to species number or concentration. In either case, the environmental variable changes the probability of conformations through a structured modification to the energy, not an arbitrary transformation.

Rather than expecting such relationships to emerge from data alone, a generative model can be biased to learn them by modifying its probabilistic structure. In a diffusion model, this can be done by changing the prior distribution or the generative dynamics so that the environmental variable is represented explicitly. This modification forms the conceptual basis for Thermodynamic Maps, a core methodology developed throughout the thesis.

1.7 Thesis contributions and organization

This thesis makes three connected contributions. First, it evaluates probabilistic generative models in molecular settings and identifies when particular model classes are useful for distributions with different dimensionality, multimodality, and data scarcity. Second, it develops the idea that thermodynamic structure can be encoded directly into generative models, culminating in the Thermodynamic Maps framework for extrapolating molecular ensembles across temperature and related environmental variables. Third, it translates these principles into practical hybrid workflows: RNAnneal for *ab initio* RNA ensemble prediction from sequence and DiffDock-Glide for molecular docking through a combination of learned proposal generation and physics-based refinement. Across these studies, the common claim is that generative models become more reliable in low-data molecular regimes when their priors, objectives, or pipelines reflect physical knowledge

and intuition.

The remainder of the thesis is organized as follows. Chapter 2 benchmarks several probabilistic generative frameworks on molecularly motivated distributions, including Gaussian mixtures with tunable complexity and Aib9 peptide torsion-angle data. These comparisons motivate the later focus on diffusion models for moderate-dimensional, multimodal molecular landscapes. Chapter 3 applies diffusion models to replica-exchange molecular dynamics data and shows that information can be mixed across temperatures to improve conformational sampling and estimate thermodynamic quantities at temperatures not directly simulated.

Chapter 4 introduces Thermodynamic Maps, the core methodological contribution of the thesis. By giving the prior distribution a thermodynamic interpretation, Thermodynamic Maps allow temperature-dependent ensembles and free-energy behavior to be inferred from limited observations. The method is validated on the two-dimensional Ising model and then applied to RNA conformational landscapes through Thermodynamic Map-accelerated molecular dynamics. Chapter 5 builds on this framework in RNAnneal, an *ab initio* pipeline for RNA structure-ensemble prediction that combines hierarchical structure generation, molecular simulation, latent representation learning, and thermodynamic scoring. Chapter 6 applies the broader hybrid physics-AI philosophy to molecular docking in DiffDock-Glide, where a learned diffusion model proposes candidate poses and a physics-based docking engine refines and ranks them. Finally, Chapter 7 discusses extensions to additional thermodynamic variables, latent generative representations, and therapeutically relevant RNA targets.

Chapter 2: A Comparison of Probabilistic Generative Frameworks for Molecular Simulations

Author contribution. This chapter is based on work with Richard John and Pratyush Tiwary. I designed the experiments, conceptualized the work, and wrote portions of the manuscript.

2.1 Introduction

In recent years, generative artificial intelligence (AI) has demonstrated a remarkable capacity to produce convincing images, text, audio, and video [58, 59, 60]. The domain of applicability of generative AI has recently extended to the molecular sciences [57], where generative AI has demonstrated the ability to predict protein tertiary structure from amino acid sequence [55, 56, 61, 62], protein-ligand complex tertiary structure from chemical identity [63, 64] and the temperature dependence of the equilibrium distribution of solvated molecular systems [54, 65, 66]. While these methods differ in many aspects, all generative models share the common goal of sampling from an unknown underlying probability distribution based on an empirical dataset.

While there are many classes of generative models, recently, probabilistic generative models have seen widespread usage. These models represent a framework that broadly encompasses flow-based [38, 67] and diffusion models [42]. The probabilistic generative framework explicitly seeks to directly model the data distribution through a series of invertible transformations (in flow-based models) or by iteratively refining noisy samples back into data space (in diffusion models), providing a flexible method for generating new data points that obey the underlying distribution of the observed data.

There is now a range of probabilistic generative models for use in different domains. Arguably, the most popular ones include Neural Spline Flows [68] (NS) models, Conditional Flow Matching [69] (CFM) models, and Denoising Diffusion Probabilistic Models [41] (DDPM). All of these have already been used for exciting and novel applications, including sound field reconstruction [70] (NS), zero-shot text-to-speech synthesis [71] (CFM), and medical image segmentation [72] (DDPM), demonstrating the utility of probabilistic generative models across different modalities. We describe these methods in Section 2.3. However, the scientific literature in this field lacks a systematic comparison of these methods for benchmark problems with tunable complexities that could establish the conditions under which one particular framework out of NS, CFM, and DDPM might be advantageous. This is particularly true for applications to molecular systems. In this work, we address this gap by carefully applying NS, CFM, and DDPM to different benchmark systems. We realize that the field is moving extremely quickly, with new variants of flow and diffusion methods appearing regularly. In this vein, we expect that the datasets used here will serve as useful benchmarks for these new methods.

Our systems include a Gaussian mixture model (GMM) and an explicit water molecular dynamics trajectory for the Aib_9 peptide [73, 74], where we collect information on the $\{\Phi, \Psi\}_i$ dihedral angles for all 9 residues. For the Gaussian mixture model dataset, we are interested in how generative model accuracy scales with data dimensionality, with data multimodality, and with training dataset size. We are also interested in exploring which model best estimates probability density differences between asymmetric modes in the training dataset. We also measure sample generation speed and model network size as data dimensionality varies. For Aib_9, we are interested in model performance on molecular dynamics data at varying levels of complexity, which we tune by looking at different residues within the peptide. We also examine model accuracy in the low training data limit for the Aib_9 dataset. Overall, our findings are:

- NS exhibits superior performance estimating probability density differences. However, NS accuracy decreases for high-dimensional data.

- CFM displays the highest accuracy at high dimensionality but diminished performance in the

presence of complex, multiple modes.

- DDPM most accurately models the complex, multimodal Aib_9 dihedral angle distribution. However, DDPM is less accurate than other methods at high data dimensionality.

2.2 Theoretical background

The microscopic probabilities of configurations of a system comprising coordinates $\mathbf{x} \in \mathbb{R}^d$ are described by the Boltzmann distribution

$$p(\mathbf{x}) = \frac{e^{-\beta U(\mathbf{x})}}{Z(\beta)} \text{ with } Z(\beta) = \int e^{-\beta U(\mathbf{x})} d\mathbf{x}, \tag{2.1}$$

where β is the inverse temperature, $U(\mathbf{x})$ is the energy and $Z(\beta)$ is the normalizing constant of $p(\mathbf{x})$, also known as the partition function. Upon first impression, Eq. 2.1 may seem trivial, but the relationship between $Z(\beta)$ and $p(\mathbf{x})$ is subtle. The moments of $U(\mathbf{x})$ with respect to $p(\mathbf{x})$ are generated by the derivatives of $\ln Z(\beta)$, i.e.

$$\frac{\partial \ln Z(\beta)}{\partial \beta} = -\langle U(\mathbf{x}) \rangle_{p(\mathbf{x})} \tag{2.2}$$

where the angular brackets denote ensemble averaging with respect to $p(\mathbf{x})$.

Since the underlying structure of $p(\mathbf{x})$ at temperature β – a potentially complex, high-dimensional probability distribution – is encoded in changes in $Z(\beta)$ – a scalar-to-scalar function – significant effort has been devoted to developing computational strategies to estimate changes in partition functions, or equivalently free energy differences, where the free energy is defined as:

$$F(\beta) = -\beta^{-1} \ln Z(\beta). \tag{2.3}$$

The derivative in Eq. 2.2 indicates that the partition function is a relational quantity – that is, changes in the partition function are thermodynamically meaningful rather than the value of the

function itself. Likewise, as a quantity derived from the partition function, free energy differences are typically of interest rather than absolute free energies.

Free energy differences are evaluated between a *target* state described by Boltzmann distribution $p(\mathbf{x})$ and a *reference* state with distribution $q(\mathbf{x})$. Assuming that the target and reference states share the same temperature (say $\beta = 1$), one may express the free energy difference as a ratio of partition functions:

$$\Delta F_{pq} = -\ln \frac{Z_q}{Z_p}. \tag{2.4}$$

Computing the free energy difference in this fashion requires evaluating the Boltzmann weights (the integrand of Eq. 2.1) using the states' energy functions. In cases where the energy function is unknown, e.g., if the configuration space comprises collective variables, then the Kullback-Leibler (KL) divergence provides an upper bound on ΔF_{pq}[75, 76]. The KL divergence between $p(\mathbf{x})$ and $q(\mathbf{x})$ is defined as

$$D_{\mathrm{KL}}(p||q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \tag{2.5}$$

and is a measurement of similarity between $p(\mathbf{x})$ and $q(\mathbf{x})$. More specifically, if $D_{KL}(p||q) = 0$, then $\Delta F_{pq} = 0$ and distributions p and q are identical.

2.3 Probabilistic Generative Frameworks

Probabilistic generative models yield samples from an intractable *target* distribution $p(\mathbf{x})$ by transforming a simpler *prior* distribution $q(\mathbf{x}')$ into the target distribution. The change of measure identity underlies probabilistic generative models: it states that the change of probability as a result of an invertible coordinate transformation $\mathcal{M} : \mathbf{x}' \to \mathbf{x}$ is

$$p(\mathbf{x}) = \frac{q(\mathbf{x}')}{|J_{\mathcal{M}}(\mathbf{x}')|}, \tag{2.6}$$

where $J_{\mathcal{M}}$ is the Jacobian of $\mathcal{M}$.

Generally, the objective of a probabilistic generative model is to find an $\mathcal{M}$ that minimizes the free energy difference between a set of empirical samples $\mathcal{D}$, and $\mathcal{M}$ applied to $q(\mathbf{x}')$.

Once obtained the map is demonstrably useful, e.g. for accelerating the convergence of free energy estimates via targeted free energy perturbation [37, 54, 66, 75, 76, 77, 78]. However, there are several approaches to optimizing $\mathcal{M}$, each with advantages and disadvantages. In Sections 2.3.1-2.3.4 we summarize neural network-based approaches to optimizing $\mathcal{M}$.

2.3.1 Normalizable Architectures

Normalizing flows [79, 80] are a class of probabilistic generative models wherein a neural network defines an invertible map f_θ with change in probability

$$\log p(\mathbf{x}) = \log q(f_\theta(\mathbf{x})) - \log \left| J_{f_\theta}(\mathbf{x}) \right| . \tag{2.7}$$

The network is optimized by maximizing (minimizing) the likelihood (free energy) of $\mathcal{D}$ under the right-hand side of Eq. 2.7. Once learned, a sample from $p(\mathbf{x})$ may be obtained by first sampling $q(\mathbf{x}')$ and then applying the inverse map f_θ^{-1}. The change in probability may be computed by evaluating the Jacobian determinant J_{f_θ}.

Computationally evaluating the Jacobian determinant is expensive; in the general case, the complexity scales cubically with the dimension d, but imposing additional structure on the transformation may simplify the calculation. For example, the complexity for a triangular Jacobian is linear in d. Normalizable architectures impose additional structure on the operations the network performs in order to simplify the determinant calculation, such as using coupling layers that alternately produce upper- and lower-triangular Jacobians [38]. In practice, however, the additional structure limits the expressivity of the network. Since the introduction of the framework, efforts have focused on balancing expressivity and computational feasibility [68, 81, 82, 83].

2.3.2 Neural Ordinary Differential Equations

Neural Ordinary Differential Equations (ODEs) [84] form the basis of diffusion and flow matching models. Neural ODEs use neural networks to model the solution of a differential equation:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = f_\theta(\mathbf{x}(t)) \tag{2.8}$$

with boundary conditions $\mathbf{x}(t=0) \in \mathcal{D}$ and $\mathbf{x}(t=1) \sim q(\mathbf{x}')$. The continuous limit of repeatedly applying the change of measure identity yields the probability flow:

$$\frac{\partial \log p(\mathbf{x}, t)}{\partial t} = -\mathrm{tr}\ J_{f_\theta(\mathbf{x}(t))} = -\nabla \cdot f_\theta(\mathbf{x}(t)), \tag{2.9}$$

which depends on the trace of the Jacobian – a computation that scales linearly with d. Similar to normalizable architectures, a neural network f_θ parameterizes the drift (right-hand side of Eq. 2.8) that minimizes the free energy difference between $\mathcal{D}$ and samples generated by f_θ.

Once parameterized, the change in probability is obtained by integrating the divergence of the probability flow over the generative trajectory, i.e.

$$\begin{aligned} \log p(\mathbf{x}(0)) &= \log q(\mathbf{x}(1)) - \int_0^1 \nabla \cdot f_\theta(\mathbf{x}(t))\, dt \\ &= \log q(\mathbf{x}(1)) - \int_0^1 \mathrm{tr}\ J_{f_\theta(\mathbf{x}(t))}\, dt, \end{aligned} \tag{2.10}$$

where the divergence can be approximated by the Hutchinson trace estimator [85].

Neural ODEs yield a change of coordinates that smoothly deforms $p(\mathbf{x})$ into $q(\mathbf{x}')$, and the neural ODE can be simulated forward or reverse in time to transport samples between the prior and target distributions and compute free energy differences. However, they are potentially difficult and expensive to parameterize since Eq. 2.8 must be simulated and backpropagation must be carried out through the simulated trajectory [84].

2.3.3 Diffusion Models

Diffusion models frame generative modeling in terms of a transport equation with an ODE solution that interpolates between $p(\mathbf{x})$ and $q(\mathbf{x}')$ [41, 42, 86, 87]. The transport equation takes the form of a Fokker-Planck equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\lambda(t)\nabla \cdot [h(\mathbf{x}, t)p(\mathbf{x}, t)] \tag{2.11}$$

describing the diffusion of a probability density under the influence of a vector-field $h(\mathbf{x}, t)$ [87]. The vector-field in Eq. 2.11 is chosen to result in a linear drift

$$h(\mathbf{x}, t) = \mathbf{x} - \nabla \log p(\mathbf{x}, t), \tag{2.12}$$

so that the diffusion has the effect of transporting an arbitrary initial density $p(\mathbf{x})$ towards a Gaussian distribution $q(\mathbf{x}')$. The convergence is asymptotic, so a time-dilation factor $\lambda(t)$ is introduced to ensure that Eq. 2.11 is sufficiently converged at $t = 1$.

The diffusion in Eq. 2.11 can equivalently be expressed as a stochastic differential equation (SDE) that transports samples from $p(\mathbf{x})$ to $q(\mathbf{x}')$:

$$\mathrm{d}\mathbf{x} = -\lambda(t)\mathbf{x}\mathrm{d}t + \sqrt{2\lambda(t)}\mathrm{d}\mathbf{B}_t, \tag{2.13}$$

where $\mathbf{B}_t$ is a Brownian motion. The linear drift in Eq. 2.12 allows for simulation free evaluation of the SDE, since the conditional path distribution $p(\mathbf{x}, t|\mathbf{x}_0, 0)$ originating from any $\mathbf{x}_0 \equiv \mathbf{x}(t = 0)$ has closed form [42, 86].

A continuous-time generative model must be both (i) reversible to transport samples from $q(\mathbf{x}')$ to those of $p(\mathbf{x})$ and (ii) invertible to guarantee that the change of measure identity may be applied. Indeed, the diffusion equation is time-reversible under the change of variable $\tau = 1 - t$,

and, remarkably, the time-reverse of the SDE in Eq. 2.13 is

$$\mathrm{d}\mathbf{x} = -\lambda(\tau)\left[\mathbf{x} + \nabla \log p(\mathbf{x}, \tau)\right] \mathrm{d}\tau + \sqrt{2\lambda(\tau)}\, \mathrm{d}\mathbf{B}_\tau. \tag{2.14}$$

The invertibility condition is satisfied when the variance of $\mathbf{B}_\tau$ is zero, with the resulting dynamics being described by the probability flow ODE:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\tau} = -\lambda(\tau)\left[\mathbf{x} + \nabla \log p(\mathbf{x}, \tau)\right]. \tag{2.15}$$

The only unknown quantity in Eqs. 2.14 and 2.15 is $\nabla \log p(\mathbf{x}, t)$ – the *score*, which must be estimated [88]. Score-based models use a neural network $\mathbf{s}_\theta(\mathbf{x}, t)$ to approximate $\nabla \log p(\mathbf{x}, t)$ from realizations of Eq. 2.13. If the score estimate is sufficiently accurate, then equations 2.14 or 2.15 can be simulated to transport samples from $q(\mathbf{x}')$ to $p(\mathbf{x})$ and Eq. 2.10 can be used to compute the free energy difference from the drift of the probability flow ODE.

2.3.4 Schrödinger Bridges

One may further desire a transport process capable of mapping samples between arbitrary densities. Obtaining such a process amounts to solving the *Schrödinger Bridge* (SB) problem [89, 90, 91, 92]. The SB problem seeks to obtain the path distribution bridging distributions $p(\mathbf{x})$ and $q(\mathbf{x}')$ that minimizes the KL divergence to a reference path distribution. Follmer [93] constructs the solution using diffusive dynamics and optimal transport: the optimal path reweights a reference Brownian path distribution with an entropically regularized optimal transport plan between $p(\mathbf{x})$ and $q(\mathbf{x}')$[90]. Two related lines of work – Stochastic Interpolants [94, 95, 96] and Flow Matching [69, 97, 98] – have been developed to approximate the SB solution numerically using neural ODEs and SDEs.

2.4 Experiments

Having examined the different flavors of probabilistic generative models, we turn now to numerical experiments to compare performance across two separate datasets. NS is an example of a normalizable architecture (Section 2.3.1), CFM is a continuous flow model (Section 2.3.2) which solves a user-selected bridge problem (Section 2.3.4), in this case obeying the optimal transport solution between the target and prior distributions, and DDPM represents the broad diffusion model class (Section 2.3.3). To compare the training, sampling, and accuracy of the NS, CFM, and DDPM models, we perform experiments on two datasets - a Gaussian mixture model in spaces of varying dimensionality and the dihedral torsion angle distribution associated with an Aib_9 molecular dynamics simulation in water (see Appendix A.2 for GMM data generation procedure and Aib_9 simulation details). We collected information on all configurational coordinates for the Gaussian mixture model, while for Aib_9 we collected information on all 9 $\{\Phi, \Psi\}$ pairs of dihedral angles. To quantify the accuracy of a given model, we begin by performing principal component analysis on the training data. After conducting training, generated samples are projected along the two principal components of the training data. This data projection is binned, and KL divergence is computed between vectors enumerating the counts in the corresponding bins.

In addition to measuring the accuracy, we also report the speed of sample generation and the number of learnable parameters, a proxy for model complexity. For each model, at least one hyperparameter controlling the neural network's depth or width was optimized, as explained in Appendix A.1. Throughout the study, each experiment was conducted three times, with independent simulations. From these three runs, the mean and variance statistics can be computed. In every graphic, the mean is plotted with error bars corresponding to the standard deviation. The complete set of experiments performed for each dataset is as follows. For the Gaussian mixture:

- Measure D_{KL} for varying data dimensionality at a fixed number of modes and amount of training data (Figure 2.1a).

- Measure D_{KL} for varying amounts of training data at a fixed number of modes and dimen-

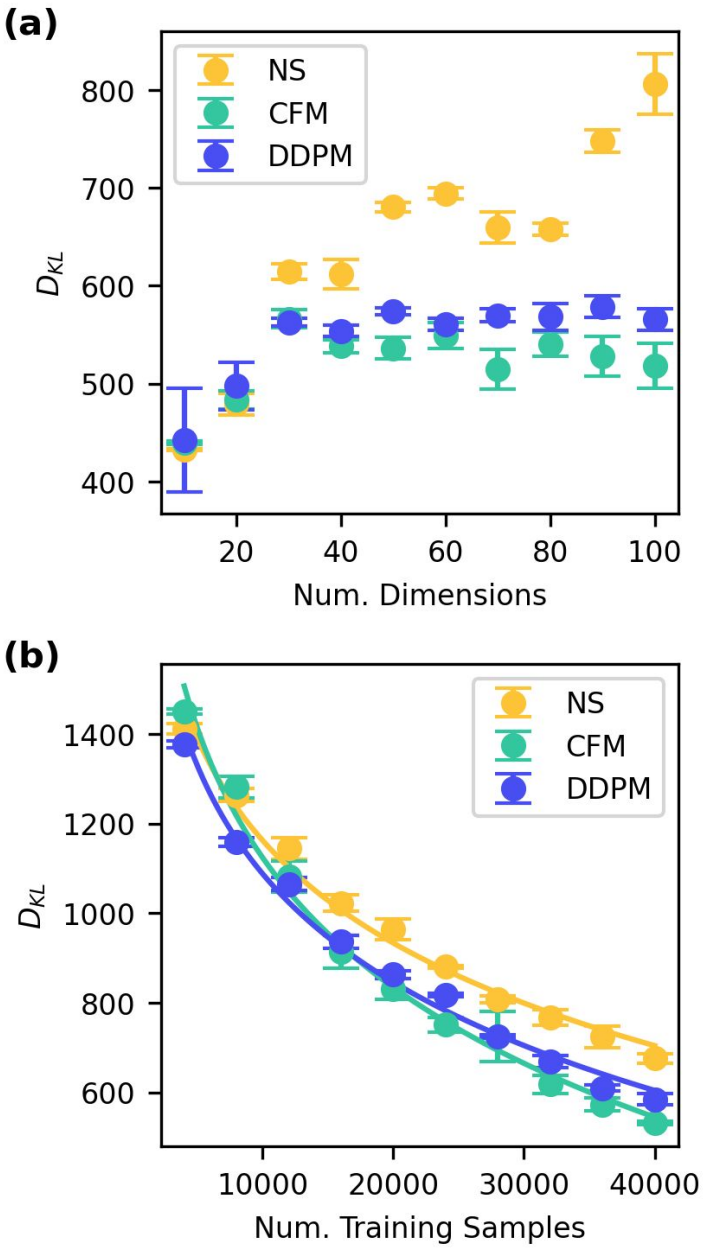


Figure 2.1: **Model accuracy results for the 4-modal Gaussian mixture dataset.** a) KL divergence (D_{KL}) comparison with analytical benchmark as a function of dimensionality. b) KL divergence comparison as a function of training dataset size at a fixed dimensionality of 50.

sionality (Figure 2.1b).

- Measure D_{KL} for a varying number of modes at a fixed dimensionality and amount of training data (Figure 2.2).
- Measure D_{KL} for varying free energy difference between modes at a fixed dimensionality, number of modes, and amount of training data (Figure 2.3).
- Measure sample generation time for varying data dimensionality at a fixed number of modes and amount of training data (Figure 2.4a).
- Measure the number of neural network parameters for varying data dimensionality at a fixed number of modes and amount of training data (Figure 2.4b).

For Aib_9:

- Measure D_{KL} for the distribution of $\{\Phi, \Psi\}_i$ for each residue i (Figure 2.5a).
- Measure D_{KL} along $\{\Phi, \Psi\}_5$ for varying amount of training data (Figure 2.5b).

2.4.1 Gaussian Mixture

The first numerical experiment we perform concerns a Gaussian mixture, a superposition of samples drawn from independent Gaussian distributions centered at different locations, with four modes but dimensionality varying from 10 to 100. After generating the data for the Gaussian mixture, we hold back 10% of the data as a test set and allow the models to train for an equal amount of time. Figure 2.1a shows that for datasets of dimensionality 40 or greater, we find a lower KL divergence, and thus higher accuracy, for CFM compared to DDPM and especially NS. The performance of CFM and DDPM seems to remain stable with no upward trend as dimensionality increases beyond 40, unlike NS, which rises sharply at very high dimensionality.

Our next GMM experiment considers varying the amount of training data the model sees and measuring model accuracy. In this test, the dimensionality is fixed at 50, and we consider a

Gaussian mixture with four modes. Figure 2.1b shows that KL divergence as a function of training dataset size varies logarithmically for all three models, and DDPM fares better than NS and CFM for low amounts of training data.

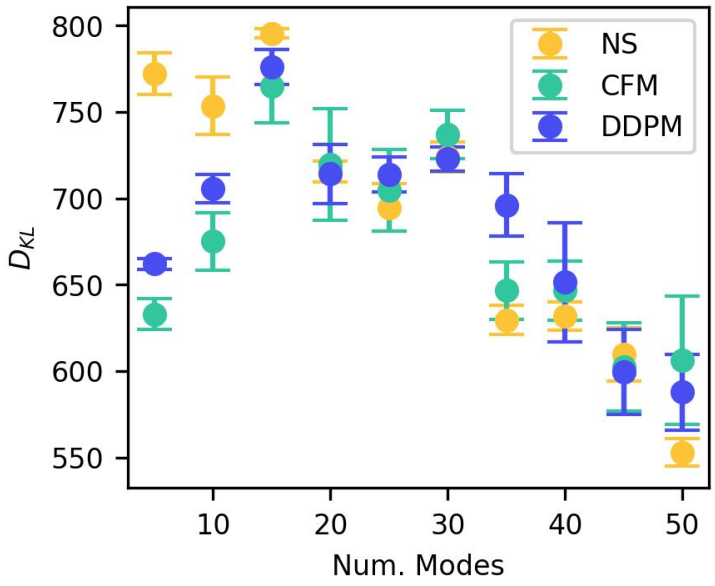


Figure 2.2: **Model accuracy results for the 50-dimensional Gaussian mixture dataset.** KL divergence (D_{KL}) comparison with analytical benchmark as a function of number of modes.

After varying the training data amount, we turn to a performance test with fixed dimension and training dataset size but a varying number of modes. For this experiment, we fixed the dimension at 50 and generated Gaussian mixtures with increasing modes ranging from 5 to 50. The general trend, shown in Figure 2.2, indicates a peak in difficulty for all models at low-intermediate multimodality, with improved accuracy as the number of modes increases. One possible reason is that as the number of modes increases, the space becomes full and more closely resembled the standard normal base distribution since the Gaussian means are randomly drawn from a standard normal distribution, making transport less costly for all three models. It is difficult to appraise an advantage for any model when examining this experiment. Perhaps CFM performs better with lower numbers of modes and NS slightly better as the modes increase and the target distribution approaches that of the normal one. In any event, we maintain that DDPM most accurately models the complex, multimodal Aib_9 molecular data (Section 2.4.2), which does not exhibit the cleanly separated modes, and does

not suffer from convergence to a standard normal distribution as seen in the Gaussian mixtures dataset.

We also note that all of the models tested only claim to estimate the density of modes present in the empirical data. The *mode discovery* task, wherein a generative model predicts additional modes is not considered in this work, and we restrict our study to training datasets with full mode coverage.

The fourth experiment performed on the Gaussian mixture dataset examines the ability of each model to reproduce free energy differences found in training data. To this aim, we designed a training dataset with a 50-dimensional bimodal Gaussian mixture where the free energy difference between the two modes varies. This construction differs from the 4-modal distribution of the other GMM experiments so that the free energy difference between the two modes can be isolated and examined. We can compute the free energy difference between the two modes by first establishing a boundary defining two domains corresponding to the two states in the PCA histogram space. To compute the free energy difference between the two modes, we first compute the partition function for each state using $Z_i = \sum_{j \in i} p_j$, where i is a state label and p_j represents the probability associated with histogram bin j. For both numerical stability and to exclude low-probability, high-free-energy bins, we impose a free-energy minimum cutoff of roughly 0.0374 kJ/mol from the corresponding energy minimum and only sum over bins meeting this criterion for either domain. The free energy difference is then $\Delta F = -\frac{1}{\beta}\ln(Z_1/Z_2)$, where the partition functions correspond to the two modes, and we have used $\beta = 1$.

Figure 2.3 shows the accuracy results and the coefficient of determination r^2 for each model compared to the training free energy difference. We conclude that NS reproduces the training free energy differences the most faithfully, followed by CFM and DDPM. It is useful to note that the least accurate model by KL divergence in our principal GMM dimensionality test outperforms the other two in this case.

Our final two GMM experiments concern sampling speed and model capacity, both measured as dimensionality varies for a Gaussian mixture with four modes. As shown in Figure 2.4a, CFM

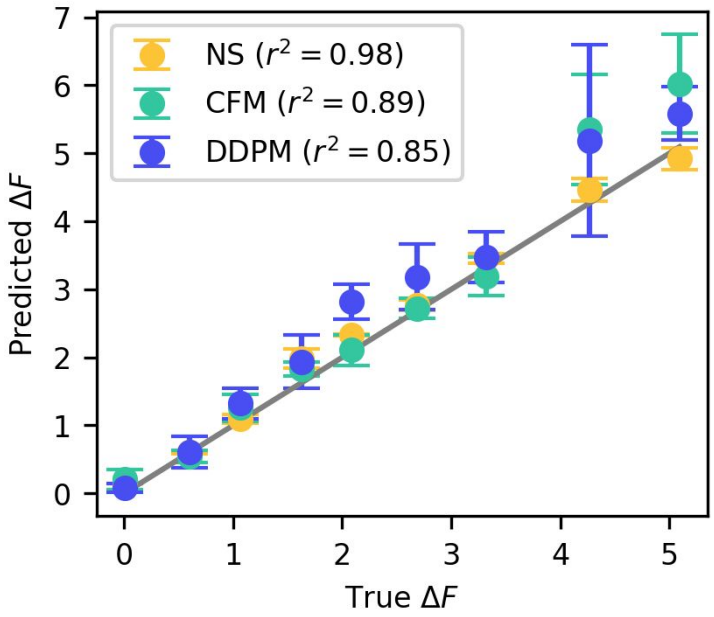


Figure 2.3: **Free energy difference estimation accuracy on asymmetric bimodal distributions.** A free energy difference of zero represents two equal Gaussian modes, while higher free energy indicates a higher level of asymmetry. r^2 is computed with the residuals of each model from the plotted line indicating training free energy difference. We impose a free energy cutoff of 0.0374 kJ/mol and note the data dimensionality is fixed at 50.

displays much faster inference than DDPM or NS. Computing the likelihood associated with a given sample is also useful in many statistical mechanics tasks. The NS architecture conveniently passes likelihoods through the same series of invertible transformations as samples, and thus, the cost of likelihood estimation is included in this experiment. In contrast, DDPM and CFM require additional probability flow ODE simulations to produce likelihoods at approximately double the cost as inference for samples alone [42]. As such, this separate process is not accounted for here, although it would not change CFM's superior speed ranking. Figure 2.4b shows the results of the model capacity measurement. CFM and DDPM employ the same predictive neural network, so their model capacity is equal. In contrast, the NS network is initially less expensive but increases rapidly in size as dimensionality climbs.

2.4.2 Aib_9 Dihedral Torsion Angles

The previous example has a limit of four modes in the underlying probability distribution. More often than not, molecular systems tend to have a very large number of modes corresponding to different metastable states. To mimic this situation, we move to a more complex system. Aib_9 is a synthetic peptide used as a model system because of its clear chirality transitions between left and right-handed forms. We simulate a molecular dynamics trajectory of an Aib_9 molecule and record all of the 9 $\{\Phi, \Psi\}_i$ dihedral torsion angle pairs (18 angles in total) as a function of simulation time (see Appendix A.2 for details of the molecular dynamics data generation procedure). In this experiment, we input a collection of 18-dimensional vectors of $\{\Phi, \Psi\}_i$ values as training data and train the models to generate samples from this target distribution. KL divergence is then computed for each residue individually by isolating the corresponding $\{\Phi, \Psi\}_i$ pair and performing principal component analysis between the held-back test set and generated samples.

The 'exterior' residues closest to the end of the peptide chain, residues 1 and 9, for example, are more flexible and thus more rapidly transition between left and right-handed orientations, sampling more of the transition states and other high-energy regions. These residues thus have a more complex distribution than the middle residues, which exhibit slower transitions [99]. This

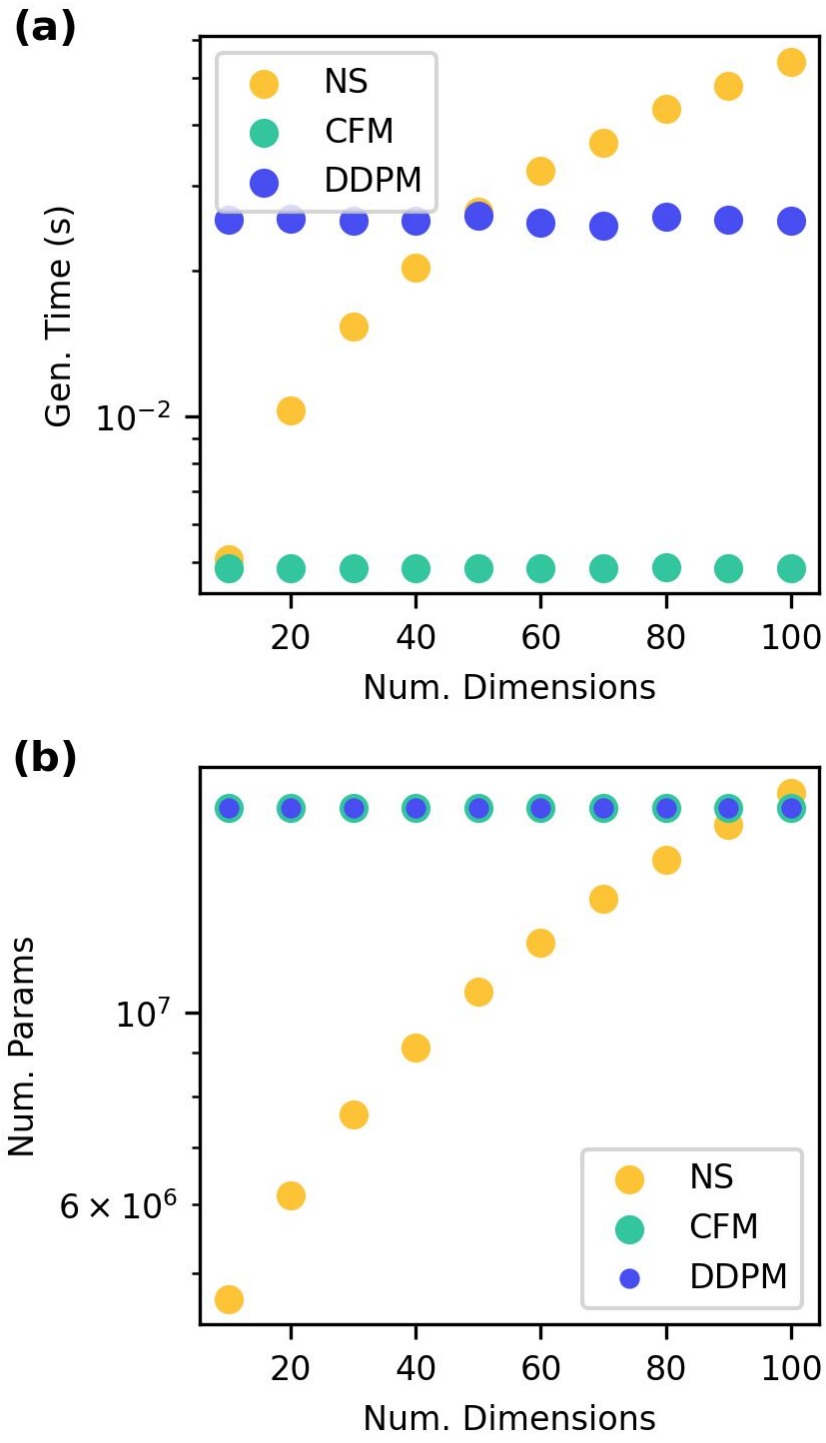


Figure 2.4: **Speed and model size results for the 4-modal Gaussian mixture dataset.** a) Single sample generation time comparison as a function of dimensionality. b) Model capacity comparison measured by total number of learnable parameters as a function of dimensionality.

behavior can be seen in the U-shaped curve of KL divergence scores, indicating the higher difficulty the models face in generating the exterior residues.

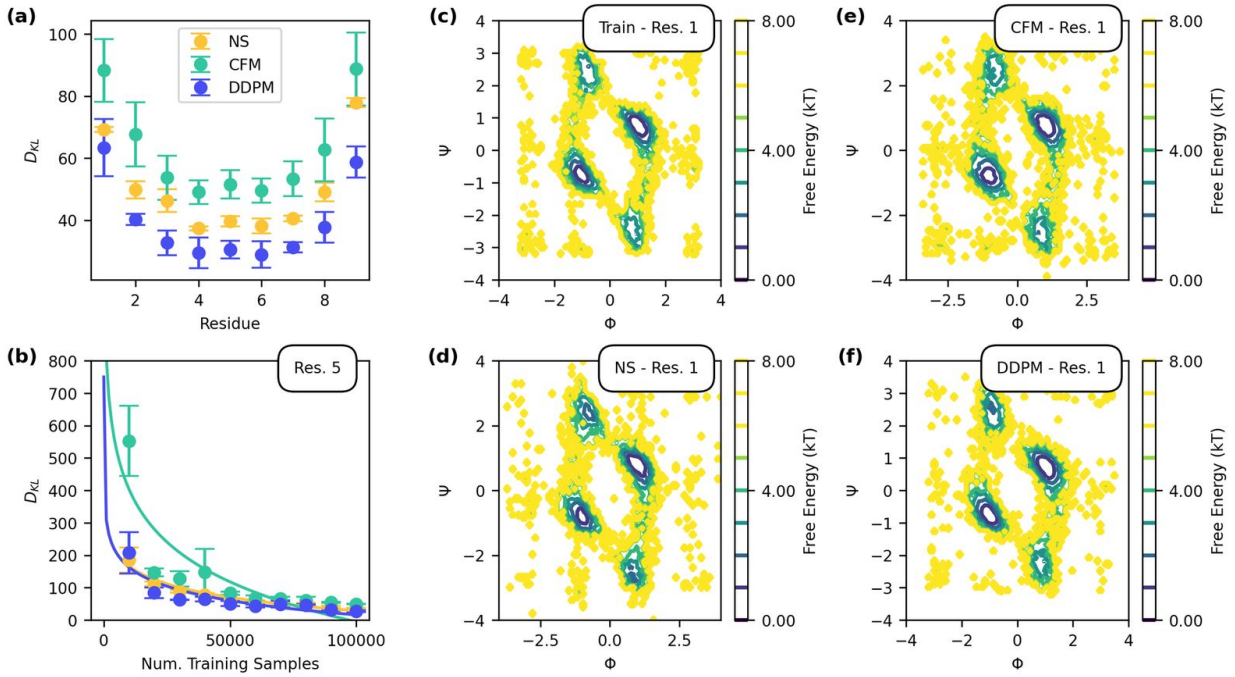


Figure 2.5: **Model accuracy results and generated data for the Aib_9 peptide.** a) KLD performance comparison as a function of residue index for the complete Aib_9 torsion angle dataset. b) KLD performance comparison for the three models and Gaussian baseline fit as a function of training dataset size for the Aib_9 torsion angle data distribution at residue 5. c) d) e) f) $\{\Phi, \Psi\}$ show free energy contour plots for torsion angle distributions at residue 1 for training data and generated data for NS, CFM, and DDPM, respectively.

Training is performed exactly as in the Gaussian mixture experiment, and accuracy is measured via the KL divergence at individual residues via the procedure described above. Correspondingly, training is done in an 18-dimensional space, and measurement is done in a 2-dimensional space, by isolating the $\{\Phi, \Psi\}_i$ torsion angles corresponding to the i-th residue in the training and generated data, performing PCA on the training data, transforming both datasets, binning, and computing KL divergence. While transforming according to 2 principal components is not strictly necessary for the already 2-dimensional data, we retained this step for consistency with the GMM experiments. Figure 2.5a shows the accuracy results as a function of residue index. We observe the highest accuracy from DDPM, with CFM performing the least well universally.

We show the results of varying training dataset sizes in Figure 2.5b. In this experiment, the residue considered was fixed as residue 5. DDPM and NS perform comparably and show better accuracy than CFM for all dataset sizes. $\{\Phi, \Psi\}_4$ plots at residue 1 are shown in Figure 2.5c-f respectively for training data and generated data for NS, CFM, and DDPM.

2.5 Discussion

In this work, we explored qualitatively and quantitatively how three different classes of probabilistic generative models perform on datasets with tunable, varied complexity. We considered three classes of models, namely Neural Spline Flows (NS) models, Conditional Flow Matching (CFM) models, and Denoising Diffusion Probabilistic Models (DDPM). This selection of models is by no means exhaustive, and recently introduced architectures such as Rectified Flow [100], Latent Diffusion Models [58], and newly improved diffusion architectures [101] stand out as prime candidates for future testing with these benchmark datasets. For our chosen models, we performed experiments on a Gaussian mixture model and molecular dynamics simulations of a 9-residue synthetic peptide undergoing chirality transitions in water. After introducing each class of generative architecture, exploring strengths and weaknesses, and examining the results of our model comparison experiments on the Gaussian mixture and Aib_9 torsion angle datasets, we may now draw some conclusions about the relative cases in which each model is the optimal choice. CFM outperforms other models for high-dimensional datasets of limited complexity, such as the Gaussian mixture model, and exhibits the fastest inference. For lower-dimensional datasets of high complexity, such as the Aib_9 torsion angle dataset, DDPM is the most accurate. For the free energy difference estimation task, NS most accurately reproduces asymmetry between modes in the training data. We hope these conclusions will help guide the selection of models for a given task depending on the characteristics of the training data, including dimensionality, complexity, and asymmetry, and how sensitive the generative problem is to accuracy, cost, and speed. We also expect that the systematic curation of datasets with quantified complexity will be helpful for future

methods developed for probabilistic generative modeling and beyond.

Chapter 3: From Data to Noise to Data: Mixing Physics Across Temperatures with Generative AI

Author contribution. This chapter is based on work with Yihang Wang and Pratyush Tiwary. I conducted and analyzed the RNA simulations.

The comparison in the previous chapter identifies diffusion models as especially well suited to the moderate-dimensional, multimodal distributions common in molecular simulations. In this chapter, diffusion models are employed to infer a family of equilibrium distribution from replica exchange molecular dynamics data.

How does one learn physics and chemistry at a certain temperature on the basis of experiment or experiments performed at certain other temperatures? The Arrhenius Equation [102] provides the simplest way to do, if one is willing to make significant simplifying assumptions about the energy landscape of the system. Often arbitrarily complex systems indeed display kinetics conforming to the Arrhenius picture [103], but frequently one also observes its violations [104]. Even when the Arrhenius picture is applicable, it only allows extrapolation of a single number i.e. reaction rate across the temperatures. It thus remains desirable to develop techniques allow inferring much more detailed information across a temperature range. These could be thermodynamic observables such as relative probabilities of a molecule's different conformations or of a crystal's different polymorphs, or more detailed kinetic observables than just an overall rate constant. Even more generally, given observations of the positions and velocities of all constituents of a $N-$body system at some set of temperatures, we seek to estimate what they will be at a temperature where the experiment or simulation was never even performed. The problem is hard because the probability distribution connecting configurational coordinates across temperatures is hard to sample from, especially since

N is extremely large for most systems of interest, equaling at least tens of thousands.

Here we propose a framework using generative Artificial Intelligence (AI) that learns to efficiently sample such high-dimensional, complex probability distributions for $N-$body molecular systems valid across temperatures. There are two central ideas guiding our framework. Firstly, we do not treat the temperature T as just a control parameter. By noting that the temperature is a measure of the average kinetic energy, we instead work directly with the fluctuating kinetic energy, using equipartition theorem to define an instantaneous, effective temperature which we still call T. For finite N not yet in the thermodynamic limit, T so defined will display significant fluctuations proportional to $1/\sqrt{N}$. Secondly, if we have experiments or simulations performed at temperatures $T_1, T_2, ...T_K$, we can view them all together as being sampled from the same, but unknown, joint probability $p(\mathbf{x}, T)$ for $N-$body configurations $\mathbf{x}$. We then use a generative AI method, specifically denoising diffusion probabilistic models (DDPM) [41, 105] to generate many more samples from $p(\mathbf{x}, T)$, given the spare, high-dimensional dataset.

To learn a $p(\mathbf{x}, T)$ from such a high-dimensional and sparse dataset, we use the denoising diffusion probabilistic model (DDPM),[41, 105] a generative artificial intelligence (AI) framework inspired by non-equilibrium dynamics to approximate and sample from very high-dimensional distributions. By learning $p(\mathbf{x}, T)$ we are referring to the ability to generate many more samples from $p(\mathbf{x}, T)$ given the data set we have from simulations or experiments performed so far. DDPM has been shown to possess the ability to infer and learn the underlying relationship from complicated data i.e. high-dimensional and with complex correlations, but also noisy and sparse [41, 105]. DDPMs learn two diffusion processes expressed through stochastic deep neural networks, which are called noising and denoising. The forward diffusion or noising converts the samples from high-dimensional, structured and unknown probability distribution into simpler and analytically tractable white noise. The backward diffusion or denoising learns the mapping back from noise to meaningful data. The central idea is that it is easy to generate numerous samples from the noisy distribution, which can then be denoised back to structured data. The process has been demonstrated to be comparable or even outperform other generative AI models for generating high

quality samples, and in its ability to model in an unsupervised way the underlying semantics behind meaningful data. [41, 105]

In spite of their promising potential, to the best of our knowledge DDPMs have not yet been used in the context of mixing information across processing conditions as done here. We are able to achieve this here through recognizing that for finite-size systems, as one has in molecular simulations, the temperature can be associated with a random variable that has significant fluctuations and not just a control parameter. The protocol developed here should be applicable quite generally to data coming from simulations or experiments at different temperatures. It should also be extensible to mixing data from control parameters other than temperatures - possibly including concentration, pressure and volume. Here we focus on a specific class of molecular simulations known as Replica exchange molecular dynamics (REMD) [27, 106] that use information generated at a ladder of temperatures to swap configurations across temperatures. REMD has been extremely powerful over the decades for the study of molecular systems with rough energy landscapes for fundamental science and practical applications. [30, 107] Numerous advances have been introduced over this basic idea in order to make it more efficient computationally [108, 109, 110, 111, 112, 113, 114, 115] and it continues to be an area of very active research. We demonstrate how DDPM applied to all-atom, femtosecond resolution REMD simulations of a small peptide and a RNA nucleotide very significantly improve the quality of data generated across temperatures, including temperatures at which the simulation was never performed and even at temperatures outside the ladder of temperatures. significantly improves the estimates of free energies made through REMD and providing accurate sampling in parts of configuration that were not previously visited in the lowest temperature replica during REMD. This could be metastable states or transition states. We also show how this can be used to generate samples at temperatures not included in the ladder of replicas. The samples we generate have thermodynamic relevance and correspond to correct Boltzmann weights as opposed to being just wild hallucinations. [41, 116] Due to its generalizable framework and simple post-processing nature of application we thus believe that this work should be extensible to a wide range of studies to supplement and extend the range of simulations/experiments without actually performing them at

all conditions of interest.

3.1 Results

Defining a fluctuating effective temperature. We work with the equivalent $p(\mathbf{x}, \beta)$ where inverse temperature $\beta = \frac{1}{k_B T}$ and k_B is Boltzmann's constant. Molecular dynamics (MD) or Monte Carlo (MC) methods allow sampling configurations $\mathbf{x}$ as per the equilibrium probability $p_\beta(\mathbf{x}) \equiv e^{-\beta U(\mathbf{x})}/Z$, where $U(\mathbf{x})$ is the potential energy of the $N-$body system and $Z = \int d\mathbf{x} e^{-\beta U(\mathbf{x})}$ is the partition function. For systems of practical interest in biology, chemistry and materials science, if T is not large enough, it becomes nearly impossible to sample reliably from $p_\beta(\mathbf{x})$ as many regions of interest in configuration space will have $e^{-\beta U(\mathbf{x})} \approx 0$. To deal with this problem, in REMD one simulates K+1 replicas of the system at temperatures $\beta = \beta_0 > \beta_1 > ... > \beta_K$. For low enough β_K or equivalently high enough temperature T_K, the sampling from $p_{\beta_K}(\mathbf{x}) \equiv e^{-\beta_K U(\mathbf{x})}/Z_K$ is expected to be more ergodic. One then periodically exchanges conformations between consecutive pairs of replica with a Metropolis-type acceptance probability that depends on the potential energies of the two replicas and their temperatures. This way even the low-temperature β_0 replica can explore configurations that it would not have otherwise visited.

We now argue that by treating the temperature just as a control parameter REMD is not making full use of the information gathered across the ladder of temperatures. In most current incarnations of REMD (excluding exceptions such as Ref. [33]), all that the higher temperatures do is to help the lower temperature replicas discontinuously appear in different locations of the configuration space. We take a slightly different view of the temperature, in REMD here but easily generalizable to other thermodynamic variables which show fluctuations for finite system size. By working with an approximate, effective temperature, we treat it as a random variable instead of a control parameter. To do so we work with an effective, instantaneous temperature of the system associated with its kinetic energy instead of the temperature of the heat bath that we expect the thermostat to enforce on average. More rigorously thus, we are sampling the joint $p(\mathbf{x}, \kappa)$ where

the per-particle kinetic energy κ is related through its ensemble average to the temperature, i.e. $\langle\kappa\rangle = \frac{3}{2\beta}$. We emphasize that our effective temperature is not the true thermodynamic temperature - however for the sake of simplicity we still use the symbol T or β to denote this effective temperature or its inverse respectively.

Our motivation in doing so is that all replicas across different temperatures can be viewed as being sampled from the same joint probability $p(\mathbf{x}, \beta)$ – as opposed to different replicas sampling from respective $p_{\beta_K}(\mathbf{x})$. This change of perspective allows us to combine together the data collected from different temperature as having arisen from the same, although intractble, probability distribution $p(\mathbf{x}, \beta)$.

Challenge in sampling the intractable joint probability across configurations and temperatures. Our task at hand now is to learn the joint probability $p(\mathbf{x}, \beta)$ given sampling that has already been performed in REMD across $\mathbf{x}-$space and temperatures $\beta = \beta_0 > \beta_1 > ... > \beta_K$. There are two main challenges in this. The first challenge comes from the curse of dimensionality: the memory or computational resources needed to track a very high-dimensional distribution function increase exponentially as the number of degrees of freedom increases. For instance, for REMD of a small 9-residue peptide in explicit water, which we study here, we exchange all 4749 atomic coordinates between the replicas, but for the purpose of analysis we set $\mathbf{x}$ as 18 Ramachandran dihedral angles. This means we already have a 19-dimensional space where binning procedures are out of the question. The second challenge comes from the sparsity of the data. Most of our samples come from high probability regions $p(\mathbf{x}, \beta)$ and we have very few samples for low-probability states, or for high-probability states at low temperature due to inefficient exchange betweem replica. In summary we have sparse sampling of data points in very high-dimensional $(\mathbf{x}, \beta)-$space and wish to construct $p(\mathbf{x}, \beta)$ from this information so that we can create many more samples at any temperature of interest.

Denoising diffusion probabilistic models (DDPM) can generate many more samples from $p(\mathbf{x}, \beta)$. The main idea behind the use of DDPM here is to learn a simple and easy-to-sample-from distribution $P_{\text{simple}}(\mathbf{x}, \beta)$ that approximates the true $p(\mathbf{x}, \beta)$. For notation simplicity, we denote

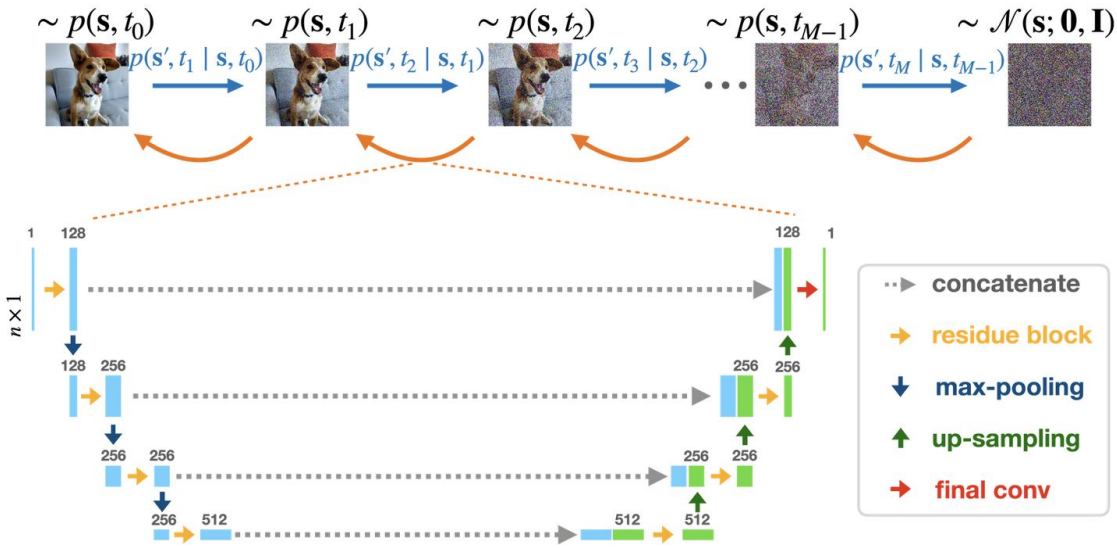


Figure 3.1: Lower panel shows the neural network architecture used in this work. It has the basic structure of the U-Net model. [117] During the diffusion process, which is indicated by the direction of blue arrows in the upper panel, noise is gradually added to the sampled data, in this case the picture of a very good boy, through a diffusion process labeled with diffusion step t_i. This changes the sampled distribution (for example, more pictures of dogs) to a simpler isotropic Gaussian distribution from which one can easily generate more samples. An AI model is then trained to reverse such a diffusion process and starting from sampled noise, learn to generate images similar to the input image by following the direction indicated by the orange arrows. It takes an 1-d array $\mathbf{s}'$, which is the noisy sample at diffusion step t_i, as input and outputs the parameterization of a Gaussian distribution to get the reversed transition kernel $q(\mathbf{s}, t_{i-1} \mid \mathbf{s}, t_i)$. Each residue block consists of three components: two 1-d convolution operations with kernel size 3 and a group normalization [118] between them. The diffusion step t_i is added to each convolutional block after being transformed by the sinusoidal position embedding [119]. The final conv label denotes the convolution operation with kernel size equal to 1. Max-pooling reduces size of the features by half, while up-sampling uses the transposed convolution to expand the size of features.

$\mathbf{s} = \{\mathbf{x}, \beta\}$ and refer to $P_{\text{simple}}(\mathbf{s})$ henceforth. DDPM does this by learning to reverse a gradual, multi-step noising process that starts with the relatively limited number of samples generated from the distribution $p(\mathbf{s})$ and diffuses to the simpler distribution $P_{\text{simple}}(\mathbf{s})$ that is easy-to-sample. For instance, $P_{\text{simple}}(\mathbf{s})$ could be an isotropic Gaussian. In addition to learning this noising process, DDPMs also learn the reverse denoising process which allows us to go back from samples generated using $P_{\text{simple}}(\mathbf{s})$ to samples that would have been generated from the underlying $P(\mathbf{s})$. Both the noising and denoising processes are modeled using diffusion processes that convert probability distributions to one another, and are implemented using the architecture shown in Fig. 3.1, which is

based on the standard architecture for DDPMs as described in Ref. [41].

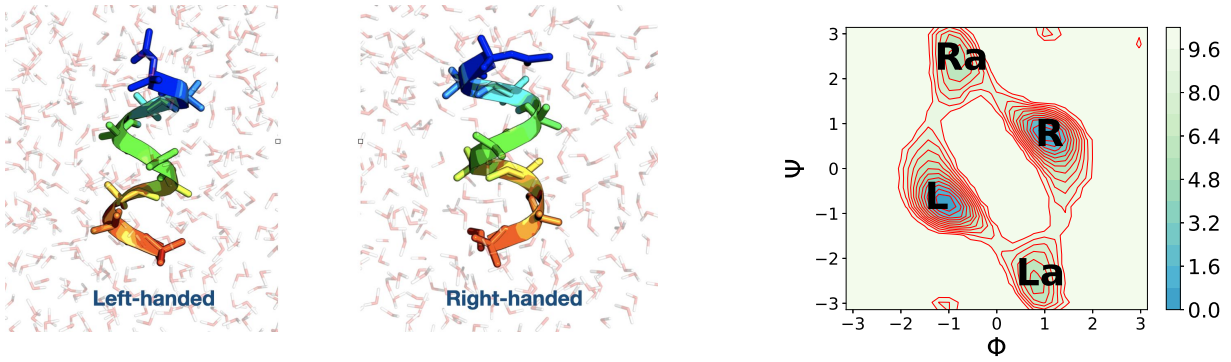


Figure 3.2: (a) AIB_9 in left-handed (L) and right-handed (R) conformations in explicit TIP3P water. (b) Free energy profile of residues (with residue 5 as an example) with two stable states labeled as L, R and two excited state labeled as La, Ra. When all residues are in L states, the peptide chain is in left-handed state and similar for right-handed state. Free energies in (b) here as well as all through the manuscript and Supporting Information are provided in units of k_BT.

Noising and denoising processes. The noising diffusion process carried out in the space $\mathbf{s} = \{\mathbf{x}, \beta\}$ that converts $p(\mathbf{x}, \beta)$ to the simpler $P_{\text{simple}}(\mathbf{x}, \beta)$ can be decomposed into M discrete steps denoted by corresponding transition probabilities $p(\mathbf{s}', t_{i+1}|\mathbf{s}, t_i)$ where $i \in [0, M]$, $P(\mathbf{s}, t_0) \equiv p(\mathbf{x}, \beta)$, and $P(\mathbf{s}, t_M) \equiv P_{\text{simple}}(\mathbf{x}, \beta)$. In DDPM, this noising diffusion process that converts sampled data to essentially noise, is set to be an Ornstein–Uhlenbeck (OU) process in which the transition probability follows an simple Gaussian form. One can then easily generate samples from $P(\mathbf{s}, t_0) \equiv p(\mathbf{x}, \beta)$. The tricky bit now is to convert these samples back to the original distribution. In Ref. 41, it was shown that this transition reversed diffusion kernel $P(\mathbf{s}, t_i|\mathbf{s}', t_{i+1})$ can also be written in a Gaussian form. A deep neural network (Fig. 3.1) is then trained using variational inference to learn the approximate reversed transition kernel $Q(\mathbf{s}, t_i|\mathbf{s}', t_{i+1}) \approx P(\mathbf{s}, t_i|\mathbf{s}', t_{i+1})$. Thus by generating samples from a normal Gaussian distribution, which we can easily do in large numbers, and then passing these through the reversed transition kernel, we can generate samples that follow the target distribution $p(\mathbf{x}, \beta)$ as desired. Note that instead of learning the joint probabilities $P(\mathbf{s_1}, \mathbf{s_2})$, it can be advantageous to learn the conditional probability $P(\mathbf{s_1}|\mathbf{s_2})$. This can be done through the protocol in Ref. 41 by adding a delta function to allow only a subset of $\mathbf{s}$ to change during the nosing and denoising process, i.e. $\delta(\mathbf{s_2} - \mathbf{s_2'})P(\mathbf{s_1}, \mathbf{s_2}, t_i|\mathbf{s_1'}, \mathbf{s_2'}, t_{i+1})$. This is very

useful when for instance we are interested in generating samples only at a certain temperature or only in certain regions of the configuration space. In the most general form of diffusion probabilistic models (DPM),[41] the reversed transition kernel $Q(\mathbf{s}, t_i|\mathbf{s}', t_{i+1})$ is considered as $Q(\mathbf{s}, t-1 \mid \mathbf{s}', t) = \mathcal{N}(\mathbf{s}; \tilde{\boldsymbol{\mu}}(\mathbf{s}', t), \tilde{\boldsymbol{\sigma}}_t(\mathbf{s}', t))$ and the neural network is trained to learn the mean $\tilde{\boldsymbol{\mu}}(\mathbf{s}', t)$ and variance $\tilde{\boldsymbol{\sigma}}(\mathbf{s}', t)$. In practice, however, there are many different ways to choose the Gaussian distribution parameterization. In Ref. [41], DDPM was introduced with a new parameterization approach to reduce the complexity of the training task. DDPM got its name because such a design makes the learning task resemble a denoising score matching procedure. [105] In Ref. 41, it was shown that with such a design, DDPM can generate samples of a quality that are comparable or even better than other generative models.

DDPM applied to Replica Exchange MD: Desirables. We now demonstrate how the above protocol can be applied to mix data collected from different temperatures and configurations in REMD and signficantly improve the quality of sampling. Specifically, we consider the following two challenging tasks:

1. Can we improve the sampling quality for the lowest-temperature replica with more accurate probability estimates than directly seen after REMD? This includes being able to generate samples in low probability regions such as transition and metastable states, and reliably estimating their free energies.

2. Can we generate samples at temperatures that are not even included in the replica ladder? This would include temperatures within the range of the replica ladder and also extrapolation to temperatures outside the range.

Peptide conformational transitions. To demonstrate the performance of DDPM, we first study a small peptide chain Aib_9 in explicit TIP3P water [120] using CHARMM36m force-field. [121] This 9 residue system (Fig. 3.2) displays rich and complex conformational dynamics [122] including the transitions between fully left-handed and right-handed helices. However, even in 4 μs unbiased MD at 400K, one can see only around 2–3 transitions between these two

dominant equiprobable conformations, and even fewer, if any, transitions to the higher energy metastable states. To improve the sampling of this system, we perform REMD with 10 replicas at geometrically spaced temperatures ranging from 400K to 518K, with temperature increased by 3% for each replica. The attempt of exchanging configurations was made every 20 ps, with acceptance rate around $1\% \sim 2\%$ between neighboring replicas, which is intentionally kept lower than what one usually has in REMD. This is because we want to show that even in the extreme cases where the number of atoms N is so large that replicas do not have enough overlap, or if one wants to reduce the number of replicas to save computational resources, our DDPM can still do a decent job of complementing REMD and reconstructing the true probability distribution at any temperature of interest. To benchmark our results, we ran unbiased MD at 400 K for $4\ \mu s$ and at 500 K for $0.6\ \mu s$.

To quantify the quality of sampling, we focus on the 18 dihedral angles corresponding to all 9 residues $(\Phi_1, \Psi_1, \Phi_2, \Psi_2, \cdots \Phi_9, \Psi_9)$. As shown in the Ramachandran plot in Fig. 3.2, the free energy surface along any pair of dihedrals is mainly characterized by four metastable states: two equiprobable low-energy ground states (L, R) and two excited states (La, Ra). The Ramachandran plots for all nine residues in the system look qualitatively similar but the middle residues of the peptide are known to be less flexible with higher energy barriers. [123, 124] We thus focus on the sampling for residues 5 and 8 as shown in Fig. 3.3. We train our DDPM on a REMD trajectory. At the lowest temperature of interest (400K), this trajectory has not yet achieved sufficient sampling. As

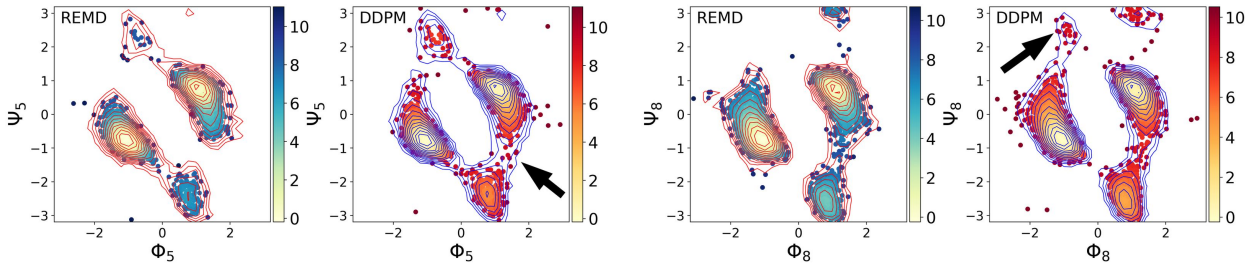


Figure 3.3: Samples (dots) from REMD (left) and DDPM (right) at 400K. The Boltzmann weights for different samples are indicated through their free energy (contour lines, separated every 0.74 k_bT). (a) Samples generated from DDPM and free energy profile projected on dihedral angles of residue 5. (b) Samples generated from DDPM and free energy profile projected on dihedral angles of residue 8. DDPM was able to generate samples in states that are not present in the training dataset for both residues, indicated with thick black arrows in the right panels.

shown in Fig. 3.3, DDPM successfully mixes information from all temperatures and configurations, and generates samples in states that are not present in the training dataset for both residues, indicated with thick black arrows in the right panels of Fig. 3.3. Specifically, for residue 5 we can see that the transition states between state R and state La, which were not being sampled in the 400K replica, are populated in samples from DDPM. For residue 8 the improvement is even more striking as the state Ra which was simply not sampled in REMD now gets populated after DDPM. Thus, DDPMs are able to accomplish the first task highlighted above. We also want to highlight that while some spurious samples are generated, seen through dots outside the free energy contours in Fig. 3.3, their Boltzmann weights are very low. Thus, our model “dreams" new configurations without hallucinating spurious configurations.

We now move to the second task from the list above, and test DDPM’s ability to generate samples by extrapolating across temperatures not considered in the ladder of replicas. To do so, we completely removed the samples from the 400 K replica in the training set and used it to train a new model, which was then used to generate samples at 400 K. Even though the lowest temperature in the training set is 412K, the model can make good predictions of free energy differences between states, as shown in Fig. 3.4. We are particularly encouraged by this finding as in general sampling at lower temperatures tends to be harder than at higher temperatures.

RNA conformational transitions. As a second example to illustrate the general applicability of our DDPM+REMD approach, we turn our attention towards sampling RNA conformational ensemble. Rare RNA structures have been previously shown to be biologically relevant [125, 126, 127], but estimating the conformational ensemble still remains computationally intractable using traditional sampling techniques [128]. RNA dynamics may occupy a wide range of timescales - from several hours for conformational changes that require breaking base pairs, to picoseconds for more continuous deformations. [127] As a consequence, identifying rare transient structures and estimating their contribution to the RNA ensemble has proved to be difficult. In this example, we consider a GACC tetranucleotide as our model system (Fig. 3.5). As a single-stranded RNA consisting of four nucleotides labeled G1, A2, C3 and C4, GACC has been previously used as a

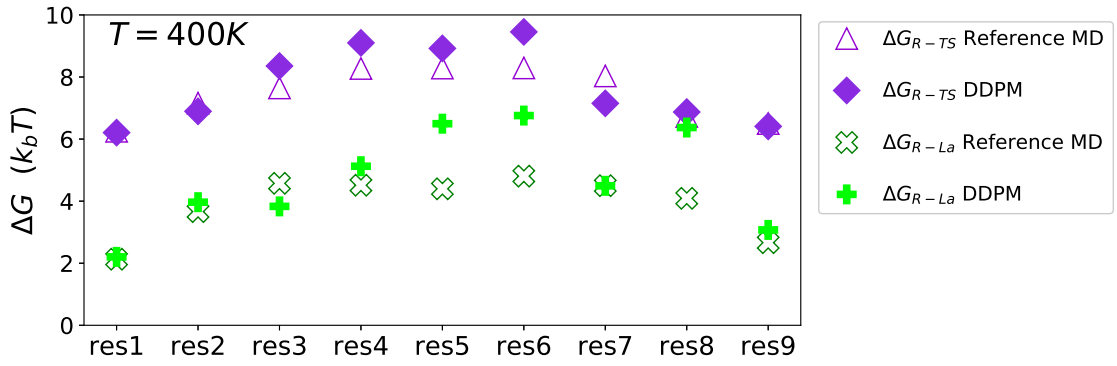


Figure 3.4: Free energy differences between different metastbale states ΔG calculated from samples from DDPM by extrapolating to a temperature at 400 K, which is lower than any temperature included in the dataset. The DDPM model was trained with REMD trajectories with 9 replicas starting at the lowest temperature 412K. Green empty crosses and purple empty triangles are reference free energy differences ΔG from 4 μs long unbiased MD at 400 K. Green filled plus signs and purple diamonds show the energy differences from DDPM. $\Delta G_{\text{L-Ra}}$ is the free energy difference between ground state L ($-2.2 < \Phi < -0.1$, $-1.6 < \Psi < 0.9$) and excited state Ra ($-1.6 < \Phi < -0.2$, $1.4 < \Psi < 3.14$); and $\Delta G_{\text{L-TS}}$ is the free energy difference between the ground state L ($-2.2 < \Phi < -0.1$, $-1.6 < \Psi < 0.9$) and a transition state ($-1.8 < \Phi < -0.8$, $0.8 < \Psi < 1.5$)

challenging test system for REMD based sampling methods for its conformational flexibility [129] Despite the fact that GACC has been widely studied, it still remains challenging to effectively sample possible configurations and is a good system to test new methods [128]. For example, in a previous study, in order to get the converged structural populations, a multidimensional replica exchange molecular dynamics (M-REMD) simulation was performed with 192 replicas with around 1 μs of simulation time per replica, thus totalling almost 192 microseconds of all-atom simulations. [129] Here we show that with DDPM, we can better estimate the free energy landscape using fewer computational resources, totalling only 12 microseconds of all-atom simulations adding all replicas.

We trained our DDPM model on 250 ns REMD trajectories from 48 replicas with temperatures ranging from 277 K to 408 K (see Appendix B.1 for details). In each frame of the trajectory, the structure of GACC is characterized by 6 dihedral angles for each nucleotide. In these REMD simulations, the sampling is not sufficient, especially for replicas at very low temperatures. Fig. 3.5 shows the free energy profile of GACC projected on the γ and ζ angles of A2 at 325 K. We can see that expected high energy metastable states indicated by black arrows in Fig. 3.5 were not sampled

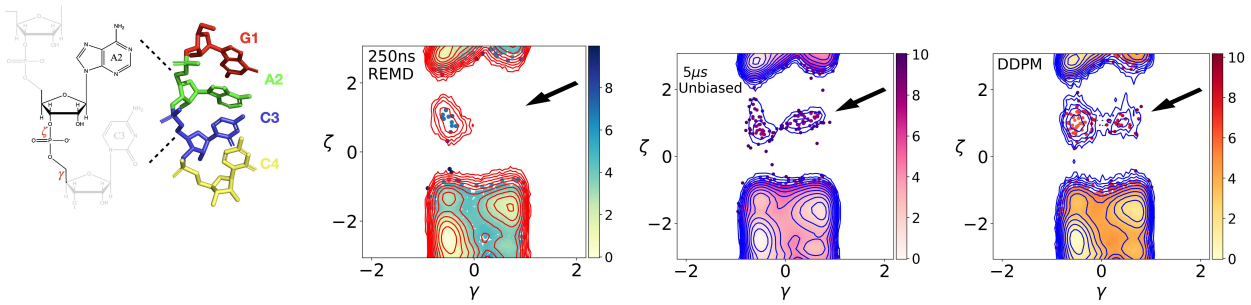


Figure 3.5: Projection of samples and free energy profile on dihedral angles ζ and γ of A2 in GACC at 325 K. The Boltzmann weights for different samples are indicated through their free energy (contour lines, separated every 0.74 k_bT). (a) The structure of GACC. (b) Samples from REMD. (c) Samples from the benchmark $5\mu s$ long unbiased MD. (d) Samples from DDPM. The metastable states that are not present in the training dataset are indicated with thick black arrows.

in REMD. In contrast, DDPM successfully generates the ensemble of such high free energy states that were never even visited in the REMD data. Here as well, any spurious configurations can be seen through dots outside the free energy contours in Fig. 3.5 with negligible Boltzmann weights.

3.2 Discussion

We have presented a generative AI-based approach that combines physics from simulations performed at different temperatures to generate reliable new molecular configurations and accurate thermodynamic estimates at any arbitrary temperature even if no actual simulation was performed there. The central idea is to not work with the thermodynamic temperature β of the system as a parameter set by the heat bath, but instead work with an effective temperature, calculated from the instantaneous kinetic energy of the system. This effective temperature shows non-trivial fluctuations for a finite size system, and on an average equals the thermodynamic temperature. Given sparse sampling from the high-dimensional space comprising configurational space coordinates and the effective temperature, we train a generative AI model that generates countless more samples of configurations at any temperature of interest. While the idea is generally applicable, here we demonstrate its usefulness in the context of the widely used replica exchange molecular dynamics (REMD) framework to improve the sampling of REMD through a post-processing framework.

We show how this significantly improves the quality of sampling at low temperatures and even generate samples in states where have not even been visited in the replicas, and at temperatures not considered in the ladder of replicas. It is worth mentioning here that a recent application [130] of normalizing flows also attempts to enhance REMD sampling through somewhat similar ideas as ours. However in that work the machinery is used to directly affect the acceptance protocol in REMD while ours is a purely post-processing scheme. Our work also bears parallel with the T-WHAM approach of Ref. 33 which uses a similar joint probability in configuration and energy space after binning them. A practical distinction is that our approach does not compute weights involving exponentials of the full system's kinetic or potential energy, thereby avoiding well-known issues with the convergence of exponentially averaged quantities [131].

The generative AI framework of DDPM used here belongs to the broad class of flow-type methods, which have been shown to have the ability to generate samples from high dimensional space with many interdependent degrees of freedom. Compared with other flow type models such as normalizing flow [54, 132] that use deterministic functions to map from an easy-to-sample distribution to target distribution, the stochastic nature of DDPM avoids the restriction of preserving the topology of configuration space and thus allows the learning of significantly more complicated distributions. At the same time the design of the transition kernels in DDPM reduces the learning task to just learning means of Gaussian kernels. This makes the training easier compared to other methods [133] while at the same time being able to learn more complicated transition kernels.

We also believe that our generative AI model, while "dreaming" thermodynamically relevant structures at different temperatures, avoids the so-called hallucinations suffered by other generative AI models, i.e. we do not generate meaningless, unphysical structures with significant thermodynamic weights [116]. We believe this is through the use of relatively simple transition kernels, which avoids overparameterization of the model, and through utilizing molecular basis functions instead of all-atom coordinates, which reduces the space that needs to be sampled. The issue of generating out-of-distribution samples that has been problematic in other methods attempting to generate molecular structure, such as the Boltzmann generator, is usually avoided by discarding the

translational and rotational degrees of freedom and reweighting the samples. [54, 133] However, calculating the weight of each sample requires knowledge of all the coordinates of a system which may also become an issue when the system contains explicit water molecules; in such a configuration space the samples will again become sparse. We finally point out that this work shows the possibility of learning generative models in the space of generic thermodynamic ensembles, by following the simple recipe that control parameters can also be viewed as fluctuating variables. As long as one is not in the thermodynamic limit – something we do not have to worry about in molecular simulations – this should be thus a practical and useful procedure for problems far beyond replica exchange molecular dynamics.

Chapter 4: Inferring Phase Transitions and Critical Exponents from Limited Observations with Thermodynamic Maps

Author contribution. This chapter is based on work that I led, from its conception through its analysis and writing, in collaboration with my coauthors.

Phase transitions are widely observed in biological, material, and social sciences. Across these disciplines, phase transitions can be defined as the emergence of higher-level, large-scale organization from the coordinated, short-range interactions between many individual constituents. Classical examples include the ferromagnetic to paramagnetic transition, boiling of water, and conformational transitions in biomolecules like proteins and nucleic acids.

Statistical mechanics, and especially the framework of energy landscapes, provides a simple and unifying way of studying phase transitions in these diverse systems [134]. In this work we are specifically interested in phase transitions in systems that stay in equilibrium throughout. For these, the Boltzmann distribution relates the probability of finding a system in a particular microscopic configuration $\mathbf{x}$ to its energy $U(\mathbf{x})$ and the system's inverse temperature β as

$$\mu(\mathbf{x}, \beta) = \frac{e^{-\beta U(\mathbf{x})}}{Z(\beta)} \quad \text{with} \tag{4.1}$$

$$Z(\beta) = \int e^{-\beta U(\mathbf{x})} d\mathbf{x}. \tag{4.2}$$

$Z(\beta)$ is a normalization constant known as the partition function, whose behavior is often associated with phase transitions. Exploration of the energy landscape is guided by competition between energy and entropy, which is encapsulated by a temperature-dependent free energy $F(\beta)$

which may be computed from the partition function as

$$F(\beta) = -\beta^{-1} \ln Z(\beta). \tag{4.3}$$

At a glance, Eqs. 4.1-4.3 suggest that the relationship between temperature, energy, microscopic probability, and macroscopic free energy is simple and tractable. It appears as if with these equations, one has the machinery to directly calculate the free energy across temperatures. By doing so for different macroscopic phases one could then obtain various thermodynamic attributes of phase transitions, including transition temperatures and phase diagrams. By calculating appropriate fluctuations, one could directly obtain response functions such as heat capacities.

In reality, however, the situation is quite complex. Studying phase transitions and their characteristics computationally is made difficult by Eq. 4.2, which requires integrating over a (usually) intractably large number of dimensions. Numerous elegant theoretical and computational schemes have been proposed over the decades to solve this problem. For example, this includes methods like free energy perturbation, Markov chain Monte Carlo (MC) methods, the replica trick and others [31, 32, 34, 35, 135, 136].

In this study, we propose a generative Artificial Intelligence (AI) based approach that characterizes phase transitions by effectively learning the temperature dependence of the partition function, and therefore the free energy. Our method, which we call "thermodynamic maps" (TM) incorporates score-based generative modeling into the framework of free energy perturbation within statistical mechanics [41, 42, 65, 86]. The central idea underlying thermodynamic maps is to map the temperature dependence of configurational ensembles of a complex system to and from the temperature dependence of a simple, idealized system. We show how such a bidirectional mapping can be obtained. This allows for efficient generation of physically realistic samples of the complex system with the correct Boltzmann weights. Within the framework of free energy perturbation, the mapping also allows for temperature-dependent free energy estimates.

We demonstrate the broad applicability of thermodynamic maps for three complex systems

where we compare against benchmarks from theory, extensive computational studies, and experiments. The first system we consider is the Ising model on a two-dimensional square lattice. With observations made at two temperatures, one deep in the paramagnetic regime and the other deep in the ferromagnetic regime, we are able to correctly infer critical behavior. We then study two different Ribonucleic Acid (RNA) systems: the GCAA tetraloop and HIV-TAR RNA [137, 138]. For both of these, we infer the temperature-dependence of the equilibrium distribution across temperatures using thermodynamic maps trained on data generated by short molecular dynamics (MD) simulations. For both RNA systems, we predict the temperature dependence of conformational ensemble and compute melting temperatures, finding agreement with computational studies and experiment.

Given this demonstration of applicability, we believe that thermodynamic maps will be found useful for the characterization of complex phase transitions in diverse systems, especially those with multiple phases. For example, local minima within the energy landscape of biomolecular systems are often biologically and functionally relevant [139], and while they have been historically challenging to characterize computationally, we are able to study them with thermodynamic maps. The computational efficiency of the learning algorithm and scalability due to not requiring samples from the global equilibrium distribution make thermodynamic maps especially suitable for studying large-scale systems exhibiting complex behavior across long timescales.

4.1 Targeted Free Energy Perturbation

To motivate thermodynamic maps, we start with relevant background work in the longstanding problem of free energy estimation. Since pioneering work by Zwanzig and Feynman in the 1950s, many frameworks have been developed to do this efficiently [32, 34, 35, 36, 140]. The guiding principle behind all such methods is that one is usually interested in differences of free energies, rather than their absolute values. Such differences can be estimated from ratios of partition functions. One relatively recent framework is that of Targeted Free Energy Perturbation (TFEP), which is

especially relevant for machine learning approaches to free energy estimation [37].

Estimates of free energy differences between states converge according to the overlap between the states. Avoiding a technical definition, the overlap between states A and B measures how likely it is to generate a sample from B by sampling from A. The central idea behind TFEP is to increase the overlap between states with an invertible mapping of the configuration space onto itself:

$$\mathcal{M} : \mathbf{x} \to \mathbf{x}'. \tag{4.4}$$

A well-chosen $\mathcal{M}$ might map $\mathbf{x} \in A$ so that A' is a subset of B. Under such a mapping, sampling from A and transforming the sample using Eq. 4.4 always generates a sample from B; it is therefore more efficient to estimate free energies under $\mathcal{M}$ and reweight back to the original configuration space. Reweighting is possible if $\mathcal{M}$ has a well-defined Jacobian $J_{\mathcal{M}}$ by using the following property: if $p(\mathbf{x})$ is the probability density of $\mathbf{x} \in A$ and $q(\mathbf{x}')$ of $\mathbf{x}' \in A'$, then the densities are related through the identity

$$p(\mathbf{x}) = \frac{q(\mathbf{x}')}{|\det J_{\mathcal{M}}(\mathbf{x}')|}. \tag{4.5}$$

Using Eq. 4.5 free energy differences computed under $\mathcal{M}$ can be reweighted to free energy differences in the original configuration space.

Clearly using Eqs. 4.4 and 4.5 depends on specifying a map that increases the overlap between states and has a tractable Jacobian. Finding a suitable mapping is difficult due to the often complex, high-dimensional nature of the configurations; as a result, the most successful approaches to TFEP represent $\mathcal{M}$ with a neural network [78, 141]. Likelihood-based models have been especially appealing since they meet both conditions [38, 84, 142], but have drawbacks of reduced expressivity and robustness [143]. Recent lines of work have improved both aspects – for example, by pursuing more expressive black-box operations or incorporating stochasticity [83, 144, 145, 146]. In the following sections, we bypass these issues of finding a suitable mapping by exploring non-equilibrium diffusion processes as general-purpose maps.

4.2 Leveraging Non-Equilibrium Thermodynamics

We begin by recounting the main results of a line of work that explores using properties of stochastic processes for general inference [42, 86, 87, 88, 147, 148]. Suppose that the diffusion process is modeled as a Fokker-Planck equation cast in the form of a Liouville equation

$$\frac{\partial}{\partial t}p(\mathbf{x},t) = -\lambda(t)\nabla\cdot[b(\mathbf{x},t)p(\mathbf{x},t)] \text{ with} \tag{4.6}$$

$$b(\mathbf{x},t) = \mathbf{x} - \nabla \log p(\mathbf{x},t), \tag{4.7}$$

where gradients are assumed to be with respect to spatial coordinates $\mathbf{x}$ [87, 149]. Eq. 4.6 is a partial differential equation (PDE) describing the evolution of a probability distribution $p(\mathbf{x},t)$ towards a Gaussian distribution under the influence of the vector field $b(\mathbf{x},t)$ (Eq. 4.7). Time is dilated by the noise schedule $\lambda(t)$ – a monotonically increasing scalar-valued function that ensures the process converges sufficiently close to a Gaussian distribution at $t = 1$. Equivalently, the process can be expressed in terms of the evolution of a particle using a stochastic differential equation (SDE) of the form

$$\mathrm{d}\mathbf{x} = -\lambda(t)\mathbf{x}\mathrm{d}t + \sqrt{2\lambda(t)}\mathrm{d}\mathbf{B}_t, \tag{4.8}$$

where $\mathbf{B}_t$ is a standard Brownian motion and $\mathbf{x}$ is implicitly a function of t.

In Figure 4.1 we solve Eqs. 4.6 and 4.8 numerically in one spatial dimension. Figure 4.1A shows the solution of Eq. 4.6 when $p(x,0)$ is a mixture of Gaussians. In this case, it is possible the evaluate the PDE exactly since $\log p(x,t)$ is known for all t. The dynamics of Eq. 4.6 are presented in the center panel, where an initial distribution is transported to a single Gaussian; during the process the barriers that originally separated the states are destroyed. Similarly, the trajectories generated by the SDE begin at points on $\log p(x,t)$ (marked as stars) and end in a Gaussian distribution at $t = 1$. The diffusion process demonstrably increases overlap by mapping all states onto a single reference state, and does so for arbitrary initial conditions. However, if the

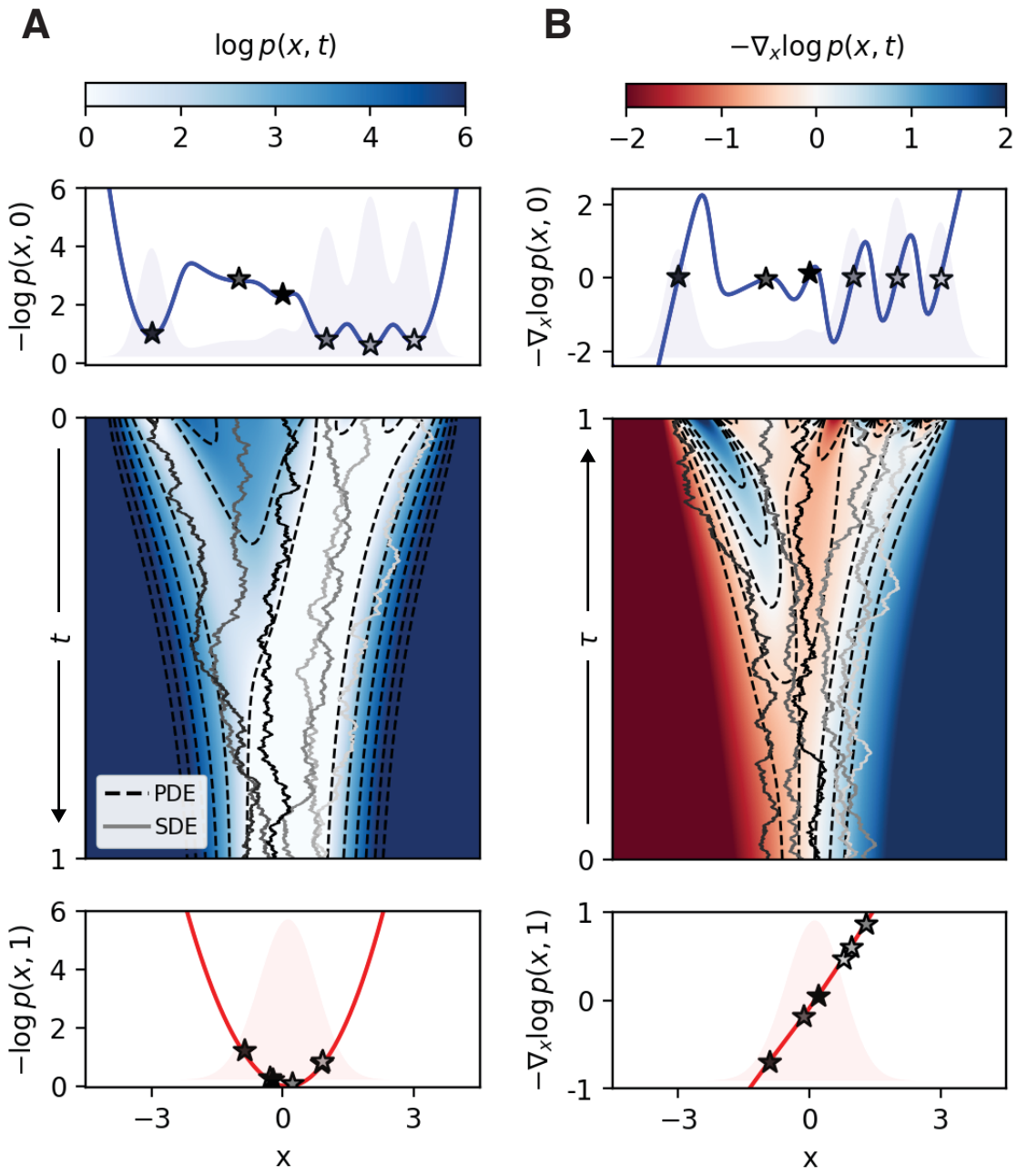


Figure 4.1: **Transporting a probability distribution with a diffusion process. A** The top panel shows the negative log-likelihood for a mixture of Gaussians. The center panels show the dynamics of a diffusion process defined as a PDE (contours) and an SDE (solid lines). Over a time interval from 0 to 1 the diffusion process relaxes to a Gaussian distribution. **B** The same dynamics are depicted in terms of the score. In both panels, the dynamics of the diffusion process can be viewed forward or backward in time (denoted by t and τ respectively).

diffusion process is to be a valid TFEP map it must meet the invertibility requirement in Section 4.1

Indeed, Eq. 4.6 is invertible under the time-reversal $\tau \equiv 1 - t$. Remarkably, it can be shown that the time-reverse of Eq. 4.8 also exists [150] and is an SDE of the form

$$\mathrm{d}\mathbf{x} = -\lambda(\tau)\left[\mathbf{x} + \nabla_{\mathbf{x}} \log p(\mathbf{x}, \tau)\right] \mathrm{d}\tau + \sqrt{2\lambda(\tau)}\mathrm{d}\mathbf{B}_{\tau}. \tag{4.9}$$

Note that the drift term is the same as in Eqs. 4.6 and 4.7 but with the opposite sign.

While particles simulated under Eq. 4.8 approach equilibrium, particles under the influence of Eq. 4.9 are driven out of equilibrium. When the variance of $\mathbf{B}_{\tau}$ is zero, Eq. 4.9 reduces to an ordinary differential equation (ODE) with a well-defined Jacobian, implying the existence of inverse mappings $\mathcal{M}_t$ and $\mathcal{M}_{\tau}$. For the SDE the inverse mapping exists in a probabilistic sense since Eq. 4.9 samples from $p(\mathbf{x}, 0)$. In either case, the existence of such an inverse mapping is critical to the class of models studied here. Here on, t will denote the forward-time process and τ the time-reverse process.

Since the diffusion process meets the TFEP criteria we are now faced with evaluating the forward and reverse processes using the PDE in Eq. 4.6 or the pair of SDEs in Eqs. 4.8 and 4.9. In either case, one must estimate either the *log-likelihood* – $\log p(\mathbf{x}, t)$, or the *score* – $\nabla \log p(\mathbf{x}, t)$. Generative models can be broadly categorized based on which quantity they estimate: likelihood-based models solve the PDE by parameterizing the log-likelihood [54, 84, 142], while score-based models solve the SDEs by parameterizing the score [41, 42]. The relationship between the likelihood and the score is depicted in Figures 4.1A and 4.1B. In this work we focus primarily on score-based models.

4.3 Score-Based Models

Score-based models (SBMs) have emerged as a robust method for generating independent samples from intractable probability distributions [42]. Using the reverse diffusion in Eq. 4.9, SBMs transport a simple *prior* distribution to a potentially complex *target* probability density that

is characterized by a collection of samples $\mathcal{D}$. The model approximates the score of Eq. 4.8, which is used to simulate Eq. 4.9 – having the effect of transporting samples from the prior distribution to the target. In an abuse of notation, we define the target distribution as $p(\mathbf{x}) \equiv p(\mathbf{x}, t = 0)$ and the prior distribution as $q(\mathbf{x}) \equiv p(\mathbf{x}, t = 1)$.

The main advantage of defining the diffusion process with the form in Eq. 4.8 is that realizations of the forward diffusion at t can be generated without simulation; in [42] the distribution of positions originating from $\mathbf{x}_0 \equiv \mathbf{x}(t = 0)$ after time t is derived from Eq. 4.8 as

$$\mathbf{x}(t) \sim \mathcal{N}\left(\mathbf{x}_0 e^{-\int_0^t \lambda(s)ds}, \mathbf{I}(1 - e^{-\int_0^t \lambda(s)ds})\right), \tag{4.10}$$

and it can be shown that $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ is proportional to the mean displacement vector from $\mathbf{x}_0$ to $\mathbf{x}(t)$ [151]. The corresponding distribution for the reverse-time diffusion does not have a closed-form solution; to evaluate Eq. 4.9 one must estimate the score for all $\mathbf{x}$ and τ and simulate from an initial condition $\mathbf{x}(\tau = 0)$.

A SBM uses a neural network to model $\mathbf{s}_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ from displacement vectors produced by Eq. 4.10. The network predicts the score by minimizing the score-matching objective function:

$$\underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{\mathbf{x}_0 \in \mathcal{D}} \mathbb{E}_{t \in \mathcal{U}(0,1)} ||\mathbf{s}_\theta(\mathbf{x}(t), t) - \nabla_{\mathbf{x}} \log p(\mathbf{x}, t)||^2. \tag{4.11}$$

Once trained, estimates of the score predicted by the network can be used to simulate Eq. 4.9. If the estimates are sufficiently accurate, then the endpoints of the trajectories launched from $q(\mathbf{x})$ are independent and identically distributed samples from $p(\mathbf{x})$. In the next section, we provide an alternative construction wherein $q(\mathbf{x})$ is viewed as the distribution arising from the dynamics of a physical system.

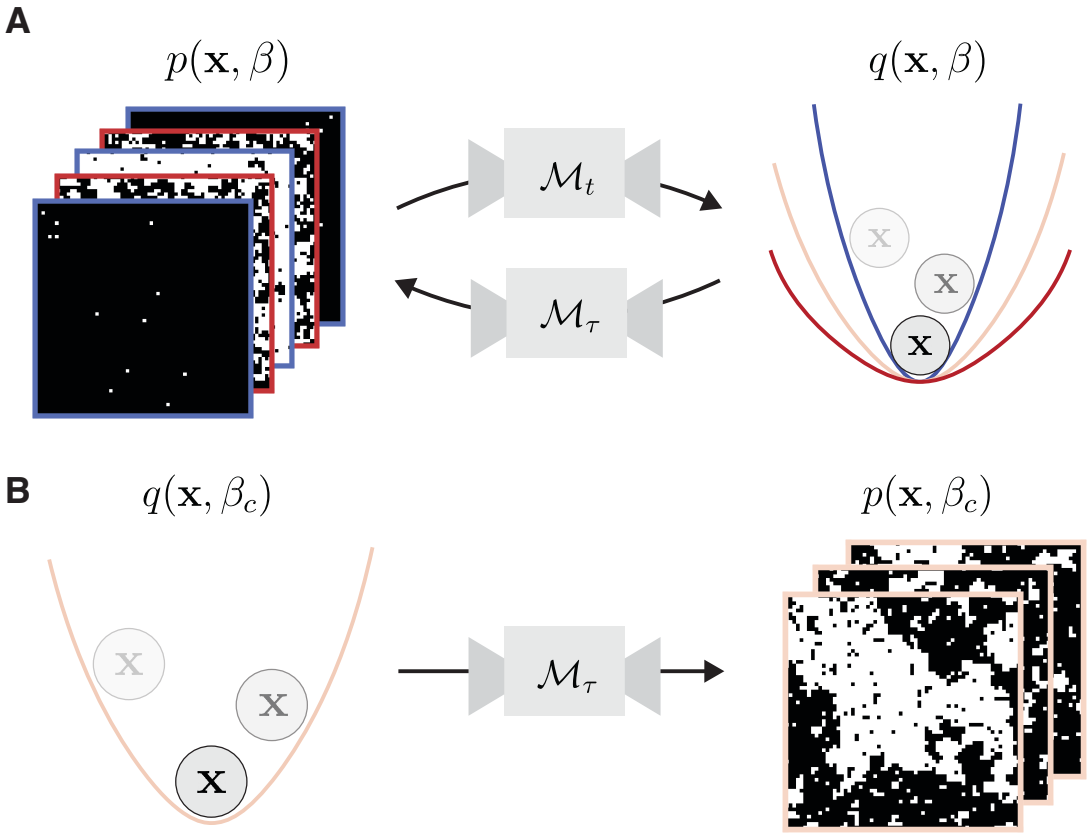


Figure 4.2: **Illustration of a thermodynamic map between systems. A** The thermodynamic map is parameterized as a score-based model transporting densities from a target distribution $p(\mathbf{x}, \beta)$ to a prior distribution $q(\mathbf{x}, \beta)$. In the schematic, data from an Ising model at two temperatures is mapped onto a simple harmonic oscillator. The generative process $\mathcal{M}_\tau$ is parameterized from realizations of the forward process $\mathcal{M}_t$. **B** Once learned, the thermodynamic map allows samples of the complex system to be generated from the simple prior system at any temperature – even those showing non-trivial behavior – by sampling the prior at the correct temperature.

4.4 Thermodynamic Maps

The Einstein-Smoluchowski relation states that temperature is proportional to the diffusivity of a Brownian motion [152]. Accordingly, we introduce temperature to the SBM SDEs (Eqs. 4.8 and 4.9) by augmenting the positions $\mathbf{x} \in \mathbb{R}^d$ with auxiliary inverse temperature-type variables $\boldsymbol{\beta} \in \mathbb{R}^d$. Together, the coordinates and temperatures form a state vector $(\mathbf{x}, \boldsymbol{\beta})^\top \in \mathbb{R}^{2d}$, and the thermodynamic map is parameterized in the joint $\mathbf{x}$-$\boldsymbol{\beta}$-space as a SBM of the form

$$\begin{pmatrix} \mathrm{d}\mathbf{x} \\ \mathrm{d}\boldsymbol{\beta}^{-1} \end{pmatrix} = -\lambda(t) \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\beta}^{-1} \end{pmatrix} \mathrm{d}t + \sqrt{2\lambda(t)} \begin{pmatrix} \sqrt{\boldsymbol{\beta}_0^{-1}} \\ \mathbf{1} \end{pmatrix} \mathrm{d}\mathbf{B}_t \quad \text{and} \tag{4.12}$$

$$\begin{pmatrix} \mathrm{d}\mathbf{x} \\ \mathrm{d}\boldsymbol{\beta}^{-1} \end{pmatrix} = -\lambda(\tau) \left[\begin{pmatrix} \mathbf{x} \\ \boldsymbol{\beta}^{-1} \end{pmatrix} + \mathbf{s}_\theta(\mathbf{x}, \boldsymbol{\beta}^{-1}, \tau) \right] \mathrm{d}\tau + \sqrt{2\lambda(\tau)} \begin{pmatrix} \sqrt{\boldsymbol{\beta}_0^{-1}} \\ \mathbf{1} \end{pmatrix} \mathrm{d}\mathbf{B}_\tau. \tag{4.13}$$

Both x and $\boldsymbol{\beta}$ fluctuate during the dynamics of Eqs. 4.12 and 4.13, relaxing to $\mathcal{N}(\mathbf{0}, \boldsymbol{\beta}_0^{-1})$ and $\mathcal{N}(\mathbf{0}, \mathbf{1})$ respectively at $t = 1$. The value of the auxiliary variables $\boldsymbol{\beta}$ at $t = 0$ is denoted by $\boldsymbol{\beta}_0$ and is a constant that sets the temperature of the diffusive dynamics for each component of $\mathbf{x}$ by rescaling the Brownian motion. The construction ensures that the diffusion process for $\mathbf{x}$ relaxes to a normal distribution at temperature $\boldsymbol{\beta}_0^{-1}$.

A thermodynamic map maps the thermodynamics of a target system $p(\mathbf{x}, \boldsymbol{\beta})$ onto the thermodynamics of a simple prior system $q(\mathbf{x}', \boldsymbol{\beta}')$, where in both systems each degree of freedom is coupled to separate baths defined by $\boldsymbol{\beta}$. Our construction is motivated by treating the prior distribution as the equilibrium distribution arising from the dynamics of a physical system. If the prior is a Gaussian distribution then the corresponding physical system is a simple harmonic oscillator, and generating samples from $p(\mathbf{x})$ at $\boldsymbol{\beta}$ amounts to sampling from the harmonic oscillator $q(\mathbf{x}')$ at the corresponding $\boldsymbol{\beta}'$. Details on the network architecture, training, and inference are

provided in Appendices C.1 and C.2.

In the generalized construction above we have assumed that each degree of freedom of $\mathbf{x}$ is coupled to a separate bath defined by the entries of $\boldsymbol{\beta}$, but in most cases physical systems are coupled to a single bath. If samples of $p(\mathbf{x}, \boldsymbol{\beta})$ are from the global equilibrium distribution of an ensemble defined by β_{bath}, each entry of $\boldsymbol{\beta}$ should be equal to β_{bath}. However, global equilibrium distribution is often not attainable in practice (e.g. in molecular dynamics simulations). In this case, allowing each degree of freedom of $\mathbf{x}$ to be coupled to separate baths $\boldsymbol{\beta}$ allows the global equilibrium distribution to be inferred from samples gathered under a local equilibrium distribution. In Appendix C.7 we discuss how the global equilibrium distribution can be inferred from local equilibrium samples if one assumes that $\boldsymbol{\beta}$ varies linearly with β_{bath}. In Section 4.6 we demonstrate the validity of the assumption by inferring the global equilibrium distribution from short, independent molecular dynamics simulations of two RNAs.

4.5 Critical Behavior of the Ising Model

We apply thermodynamic maps to the most widely studied formulation of the classical Ising model as a system of interacting spins arranged on a two-dimensional square lattice, without the presence of an external magnetic field. Arguably, this model serves as a prototypical example of a simple system with a complex phase transition. Each spin can have one of two states with value $\sigma = +1$ or $\sigma = -1$, and the spins interact with their nearest neighbors through an interaction term J and Hamiltonian given by

$$\mathcal{H} = J \sum_{\langle i,j \rangle} \sigma_i \sigma_j, \tag{4.14}$$

where $\langle i, j \rangle$ denotes nearest-neighbor pairs. When $J < 0$ the model exhibits ferromagnetic behavior in its ground state. Going forward, we set $J = -1$ without loss of generality.

As the temperature increases, an Ising model in two or higher dimensions transitions from an ordered magnetic phase to a disordered paramagnetic phase. For our set-up of a two-dimensional Ising model on a square lattice with $J = -1$, this critical temperature is known to be $T_c \approx 2.27$

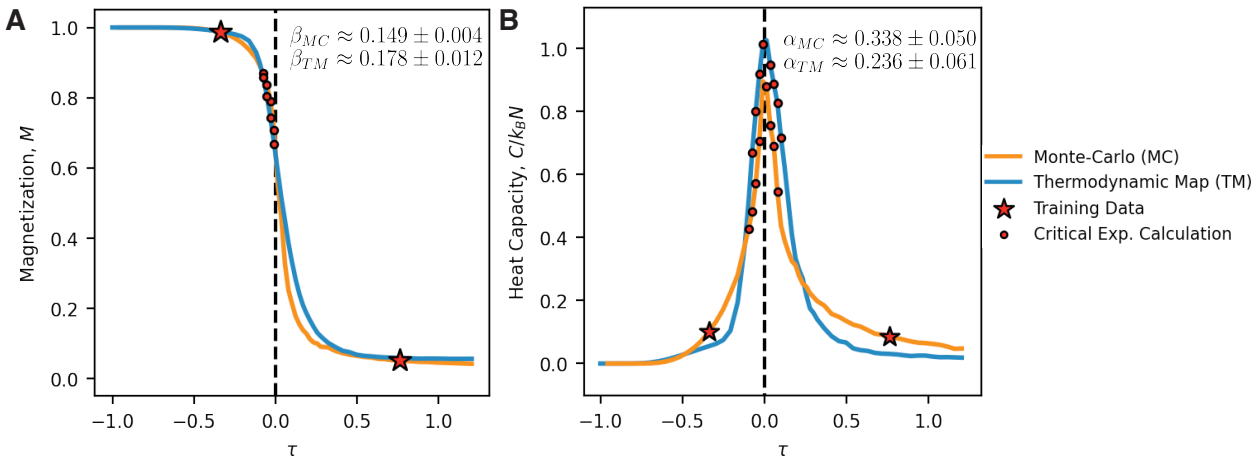


Figure 4.3: **Inferring the phase transition of the 2D Ising model from limited sampling. A** The magnetization is plotted for samples of a 32×32 square Ising model generated through MC sampling (orange) and the thermodynamic map (blue). The thermodynamic map predicts change in magnetization at T_c when trained on samples generated at $T = 1.5$ and $T = 4$ (red stars). Similar results were obtained (not reported) for a symmetric choice of training data temperatures. **B** The heat capacity of samples generated from MC sampling (orange) and the thermodynamic map (blue) is plotted. The thermodynamic map infers the divergence in the heat capacity, numerically computed for the red dots, when trained on the same samples as panel A (red stars).

[153]. The two phases can be distinguished from each other with the magnetization order parameter, M, defined as the absolute value of the average of the spins. Well above T_c, as $T \to \infty$, the spins are equally likely to be $+1$ or -1, regardless of their neighbors, resulting in a net magnetization of zero. Well below T_c, all spins in the lattice align, leading to a magnetization of 1. Power law scaling is a signature of critical behavior, and near the critical temperature the magnetization M and heat capacity C of the Ising model scale as:

$$M \sim |\tau|^{\beta} \quad \text{and} \quad C \sim |\tau|^{-\alpha} \quad \text{where} \quad \tau = \frac{T - T_c}{T_c}, \tag{4.15}$$

and with critical exponents $\alpha = 0$ and $\beta = 0.125$ (not to be confused with the inverse temperature variables $\boldsymbol{\beta}$), which can be derived analytically in the thermodynamic limit [154]. For systems that are not solvable, the critical exponents must be measured numerically, which is often done through MC sampling, wherein proposals for spin flips are generated and their acceptance or rejection is determined based on the detailed balance condition. Near the critical temperature, the presence of

long-range correlations causes MC dynamics to slow down exponentially, leading to difficulty in sampling the phase transition [155].

We investigate the ability of thermodynamic maps to infer such critical behaviors by generating configurations of an Ising model through MC sampling across temperatures and using the data from two temperatures asymmetrically spaced about T_c to train a TM (details provided in Appendix C.3). In the formalism of Eqs. 4.12 and 4.13 the lattice configurations are $\mathbf{x}_0 \in \mathbb{R}^{32\times32}$ and each entry of $\boldsymbol{\beta}_0^{-1} \in \mathbb{R}^{32\times32}$ is the bath temperature; a straightforward choice since MC sampling guarantees the global equilibrium distribution. Once the TM is trained, we generate configurations at each temperature by evaluating Eq. 4.13 with each entry of $\boldsymbol{\beta}_0^{-1}$ equal to T_{bath}. The generated continuous-valued configurations $\mathbf{x}$ are then transformed into discrete spins by $\sigma = \text{sgn}(\mathbf{x})$, and the average magnetization and heat capacity at each temperature is estimated directly from the spin configurations. Figure 4.3 shows the behavior of M and C for the MC and TM-generated samples. The TM infers the correct value of T_c ($\tau = 0$) even based on limited, deliberately misleading training data and generates samples with divergences in M and C.

Figure 4.3A shows the behavior of the magnetization for MC and TM-generated samples across temperatures, with critical exponents measured as $\beta_{MC} \approx 0.149 \pm 0.004$ and $\beta_{TM} \approx 0.178 \pm 0.012$. Similarly, Figure 4.3B shows a divergence in the heat capacity at $T_{MC} \approx 2.25$ and $T_{TM} \approx 2.30$, with critical exponents $\alpha_{MC} \approx 0.338 \pm 0.050$ and $\alpha_{TM} \approx 0.236 \pm 0.061$. For both exponents, the MC and TM exponents agree with each other, although far from the ideal value due to finite size effects.

These results highlight the ability of our model to infer the non-trivial thermodynamic behavior of phase transitions without being shown samples from the transition region; the physical meaning of temperature in the prior system has been transferred to the complex system, complete with the statistics properties associated with critical behavior. The correct prediction of T_m, even with asymmetrically spaced training temperatures, indicates our model has learned the physics of the Ising model, not merely the distribution of structures at each temperature.

4.6 Exploring RNA Conformational Landscapes with Thermodynamic Map-accelerated Molecular Dynamics (TM-aMD)

To show the broad applicability of thermodynamic maps, we now study conformational transitions and melting in two different ribonucleic acid (RNA) systems. Studying atomic-resolution conformational ensembles of RNAs through molecular dynamics simulations has proved crucial for understanding RNA structural dynamics, yet remains challenging due to the disordered, glassy nature of RNA energy landscapes [156, 157].

Increasing evidence points towards some RNAs having glassy energy landscapes where multiple minima are separated by high barriers [158, 159, 160]. The ruggedness of the landscape results in many competing degrees of freedom and no clear-cut separation of timescales within the dynamics [159, 160, 161]. The most striking feature of glassy energy landscapes as they relate to polymers is the presence of conformational heterogeneity at equilibrium [159, 162]; while the conformational landscape of proteins is often dominated by a single well-defined fold (energy minimum), the RNA conformational landscape may not be dominated by a single structure [163, 164]. This difference is analogous to the single magnetized phase of an Ising model and the multiple phases of long-ranged spin glasses [136, 165, 166, 167]. Since multiple members of the ensemble contribute substantially to the free energy and other thermodynamic observables, RNAs are best described as a weighted ensemble of conformers [164, 168].

Exploring energy landscapes by biasing dynamics along a small number of slow degrees of freedom has proven successful for exploring the conformational landscape of proteins, but the lack of timescale separation and ensemble nature of the RNA conformational landscape violates fundamental assumptions of dimensionality made in biasing methods [29]. On the other hand, thermodynamic maps have the advantage of being able to learn the conformational landscape directly in a much higher-dimensional configuration space.

We train thermodynamics maps on limited information generated by bioinformatic approaches and multi-ensemble molecular dynamics simulations to efficiently characterize RNA conformational

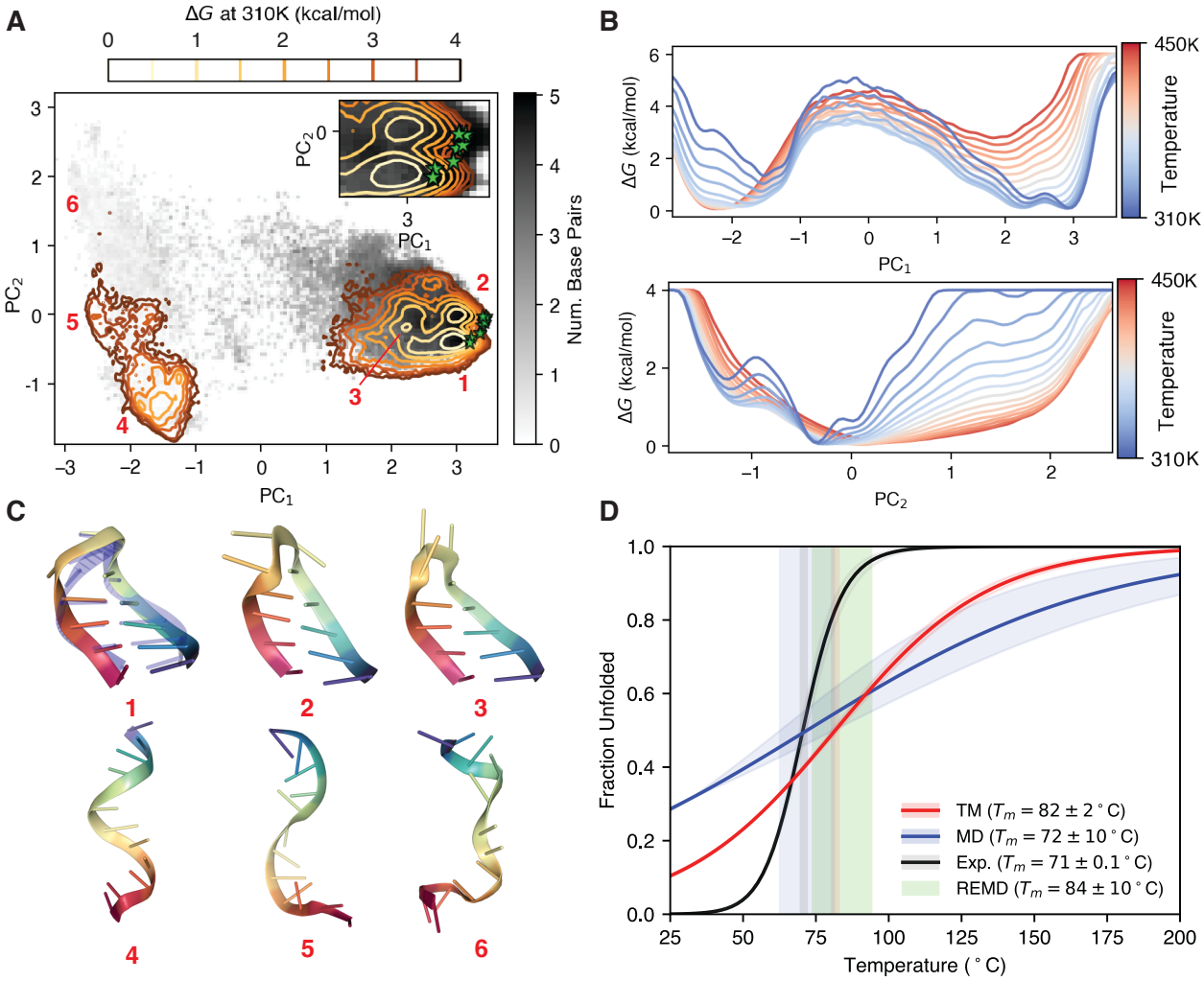


Figure 4.4: **GCAA Tetraloop Conformational Landscape. A** Joint distribution of the first two principal components of G-vectors for GCAA. Contours representing TM-generated samples at 310K are overlaid on MD samples shaded by the number of base pairs, with selected regions of the conformational landscape annotated as 1-6. The 10 NMR conformers reported in the Protein Data Bank (PDB: 1ZIH) are depicted as green stars. **B** Temperature-dependent free energy profiles along the first two principal components shown in panel A. Colors indicate temperatures ranging from 310K to 450K at 10K intervals. **C** Representative structures sampled from labeled minima in panel A, with a representative NMR conformer shown in translucent blue. The structures are colored from blue to red from the 5' to 3' end. **D** A fraction unfolded curve is obtained by reweighting the MD data in panel A with the TM conformational ensemble for each temperature in panel B, and fitting a two-state model. Uncertainties are computed from the last three iterations of TM-aMD. The cutoff for folded and unfolded states is three base pairs to match a reference REMD study [156].

ensembles. Our starting point is the physics and knowledge-based potential that is central to Rosetta to generate a putative conformational ensemble [169]. These structures serve as the starting point to

explore a more realistic energy landscape through all-atom, explicit solvent molecular dynamics performed over a range of temperatures. Between rounds of molecular dynamics simulation, the global equilibrium distribution is inferred using thermodynamic maps alongside assumptions of local equilibrium (see Appendix C.7). Initial conditions are re-sampled from the inferred equilibrium distribution using a geometric RNA order parameter. Here on, we refer to this protocol as thermodynamic map-accelerated Molecular Dynamics (TM-aMD).

Although agreement between the input equilibrium distribution and the output from the thermodynamic map is a necessary, though not sufficient condition that the true equilibrium distribution has been attained, we leave rigorously addressing convergence to equilibrium to future work. Here we take solace in the numerical results for the two challenging test systems, described next.

GCAA Tetraloop

With a shift in perspective towards viewing RNAs as dynamic entities, there has been interest in studying the variation in dynamics between different tetraloops. Studies combining computation and experiment have demonstrated that even so-called simple tetraloops can exhibit rich dynamics [170]. We study the GCAA tetraloop, a well-studied model system, enabling us to compare the equilibrium distribution generated by our model with extensive molecular dynamics simulations and experimental data [156, 171, 172]. The GCAA Tetraloop is a small, highly-stable, 12-nucleotide RNA sequence that adopts a hairpin structure consisting of an eight-nucleotide helix and a four-nucleotide loop (PDB: 1ZIH) [137]. Consistent with an ensemble perspective, the variable arrangement of nucleotides in the loop gives rise to alternative conformations, which we investigate with TM-aMD.

The thermodynamic map learns to generate RNA structures represented as G-vectors – an internal coordinate system for RNAs that effectively clusters distinct folded states [173]. The principal components of G-vectors have been shown to be a convenient visualization of RNA structural diversity, which we use to guide adaptive sampling. Further information on G-vectors,

along with details of TM-aMD can be found in Appendices C.4 and C.6.

Figure 4.4 summarizes the result of nine iterations of our enhanced sampling procedure, with a total of 50μs of simulation. Although we performed extensive MD simulations reaching long timescales, we still used two orders of magnitude less compute compared to our reference millisecond replica exchange simulation, even with sub-optimal scheduling of TM learning with MD simulation [156]. Fig. C.2A suggests that TM-aMD would benefit from more frequent reseeding of simulations.

In Figure 4.4A, we present the projection of the learned equilibrium distribution onto the first two principal components of the G-vectors. The contours represent the free energy landscape inferred by the thermodynamic map, while the shaded regions represent the distribution of structures observed in the simulation, shaded by the number of base pairs. The 10 conformers reported in the Protein Data Bank (PDB), represented as green stars, lie within the most dominant TM-predicted cluster. Figure 4.4B shows the learned, temperature-dependent free energy profiles along each of the principal components in Figure 4.4A. The first principal component clearly shows the reweighting of the folded and unfolded states with temperature. Figure 4.4C shows the structures associated with the minima in Figure 4.4A. The first three conformations are abundant at 310K and are consistent with MD studies [156, 171] and experiment (an example NMR conformer is depicted in blue). The last three represent unfolded states that are stabilized by base stacking and are weakly present in the 310K ensemble. The first principal component corresponds to the folding-unfolding transition, while the second captures the conformational heterogeneity of the loop region.

Figure 4.4D displays the melting curve of the GCAA tetraloop computed from the last three rounds of MD simulation, with the melting curve of the learned equilibrium distribution. Both curves exhibit agreement, matching the range of melting temperatures predicted from a millisecond Replica Exchange Molecular Dynamics (REMD) simulation conducted with the same force-field [156]. However, both the TM-generated melting curve and the range of melting temperatures from the reference are higher than the experimental melting temperature, indicating that current force-fields do not yet accurately capture the temperature dependence of the tetraloop ensemble.

HIV-TAR RNA

The HIV-TAR RNA is an extensively studied, 29 nucleotide RNA hairpin from the HIV genome that displays rich conformational diversity. The secondary structure consists of lower and upper helices separated by a three-nucleotide bulge with an apical loop closing the upper helix (PDB: 1ANR) [138]. The disordered loop and bulge regions mediate interactions with proteins and small molecules [174]. We investigate the conformational landscape of the HIV-TAR RNA using six iterations of TM-aMD, requiring a total of 70μs of simulation time. We infer relative free energies between the dominant conformer observed through NMR spectroscopy and alternative conformers, which are unattainable through MD simulation alone.

The conformational landscape predicted by the learned TMs, projected along the principal components of the G-vectors is shown in Figure 4.5A. Structures corresponding to the NMR ensemble (green stars) are well-separated from other misfolded structures within the two-dimensional projection. The temperature dependence of the free energy along each principal component is shown in Figure 4.5B, where we find that the free energy barrier separating the NMR conformers from unfolded states reaches a height of 5kcal/mol. Figure 4.5C depicts representative samples from each cluster. The first cluster shows agreement with the NMR conformers, with the first reported conformer shown in translucent blue. The other clusters correspond to varied secondary structure motifs. Generally, across the conformational landscape, we find that the contribution of folded states to the conformational ensemble diminishes as temperature increases. This can be clearly shown by computing a fraction folded curve (Figure 4.5D), which shows agreement between melting temperatures of the MD and TM-derived ensembles, and the experimental melting temperature [175].

Our results support the idea that RNAs evolve over a rugged free energy landscape punctuated by many long-lived states [176]. We find that the native state is indeed the most stable at physiological temperatures (310K), but misfolded states consisting of secondary structure elements still have substantial contributions to the ensemble. Minima of the energy landscape corresponding to states 2-5 in Figure 4.5 differ from the NMR conformers (state 1) by less than 1kcal/mol. Overall, our

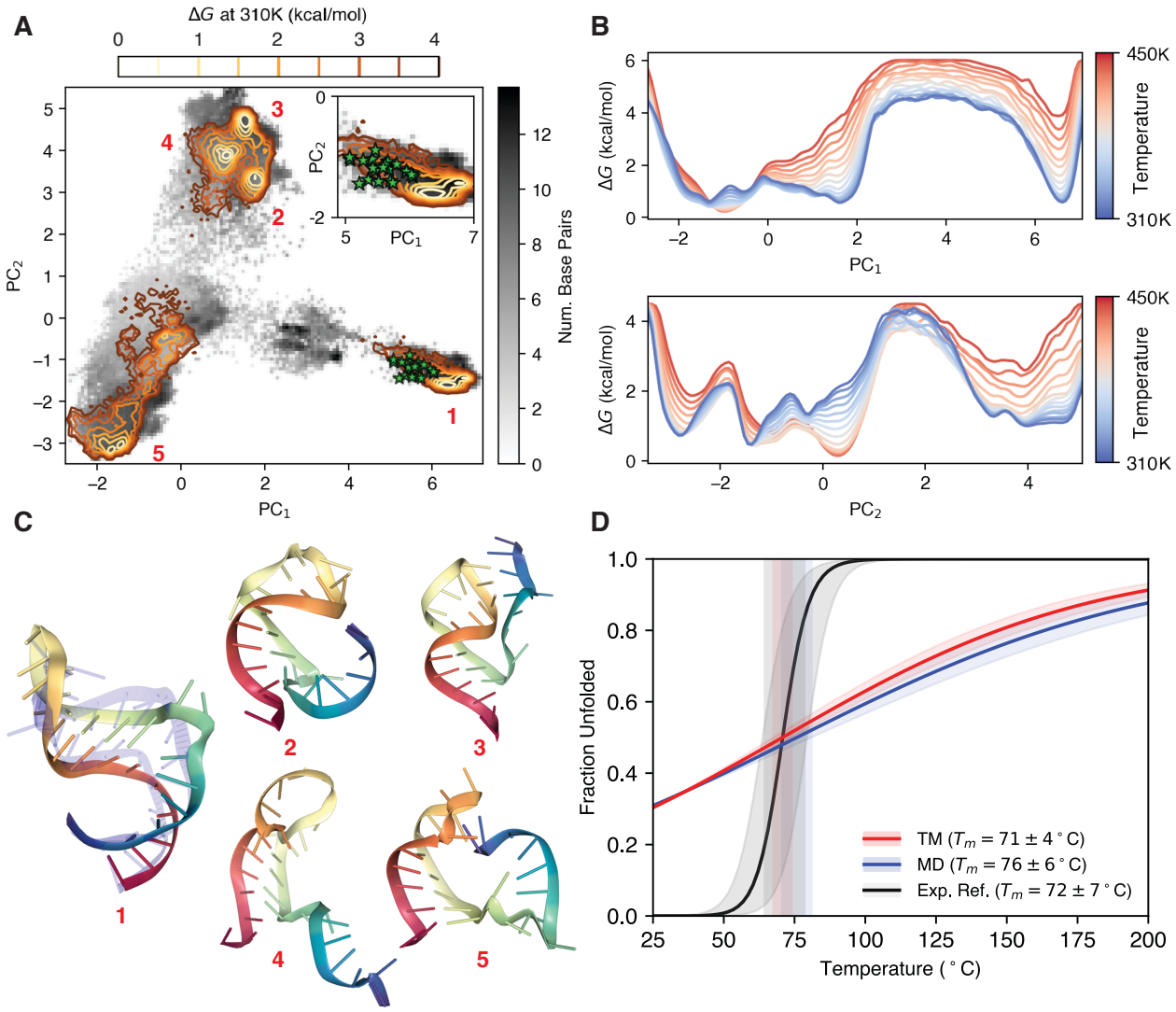


Figure 4.5: **HIV-TAR RNA Conformational Landscape.** **A** Joint distribution of the first two principal components of G-vectors for HIV-TAR. Contours representing TM-generated samples at 310K are overlaid on MD samples shaded by the number of base pairs, with basins of the TM landscape labeled 1-5. The 20 reported NMR conformers in the Protein Data Bank (PDB: 1ANR) are depicted as green stars. **B** Temperature-dependent free energy profiles along the first two principal components are shown in panel A. Colors indicate temperatures ranging from 310K to 450K at 10K intervals. **C** Representative structures sampled from labeled clusters in panel A, with a representative NMR conformer shown in translucent blue. The structures are colored from blue at the 5' end to red at the 3' end. **D** A fraction unfolded curve is obtained by reweighting the MD data in panel A with the TM conformational ensemble for each temperature in panel B, and fitting a two-state model. Uncertainties are computed from successive iterations of the TM-aMD algorithm showing agreement between MD and TM-predicted melting temperatures (see Fig. C.2B). The cutoff for folded and unfolded states is nine base pairs, as determined by the NMR conformers).

findings are in agreement with theoretical studies of RNA folding pointing towards a rugged energy landscape [160, 176].

Clearly, our findings are dependent on the accuracy of the simulation force field, which for RNAs have not yet reached experimental accuracy. Moreover, TM-aMD is incapable of reaching experimental folding timescales. Despite this, the fraction folded curve for HIV-TAR agrees with experimental results [175]. Improved forcefields are out of reach due to the difficulty of attaining equilibrium through simulation; a potential application for TM-aMD may be accelerating convergence to equilibrium, allowing for more rapid evaluation of force-field parameters.

4.7 Discussion

We have demonstrated that thermodynamic maps (TMs) are capable of inferring the non-trivial behavior of the free energy with temperature from limited data for the two-dimensional square Ising model. Additionally, we also show how thermodynamic maps may be integrated with molecular dynamics simulation to enhance sampling. In this section, we outline prospective, potential applications for thermodynamic maps. Our experiments for inferring critical behavior in the Ising model suggest a number of potential research directions that might shed light on some general principles underpinning thermodynamic maps. We outline three directions below.

In this work, we have not explored the mechanism by which thermodynamic maps learn phase transitions. In Section 4.5, as the Ising lattice grows larger the ability to infer the phase transition from only two observations diminishes. Inversely, as the lattice size decreases, the phase transition is learned more consistently. These observations suggest a connection between fluctuations due to finite-size effects and robustness of the algorithm. Spin-glasses with multiple phases present a setting for exploring these connections without the additional complexity molecular dynamics brings.

A second avenue worth exploring the extent to which thermodynamic maps infer dramatic changes in Boltzmann weights across temperature. Here too, spin-glasses are an appealing system;

some spin-glasses exhibit temperature chaos, where, in the thermodynamic limit, the weights change discontinuously with temperature [177]. In the framework provided in the introduction, this behavior manifests as a phase transition (i.e. a discontinuity in $Z(\beta)$). Recently, descriptions of temperature chaos have been generalized to finite systems, where it is prevalent on a characteristic length scale that diverges as the change in temperature becomes infinitesimal [178].

Another direction closely related to those outlined above is determining if thermodynamic maps can capture critical behavior arising from correlation functions. The order parameters explored in this work (magnetization and heat capacity) are measured by averaging over the entire system, but critical exponents are derived from the asymptotic behavior of correlation functions that capture the microscopic details of fluctuations. To measure some critical exponents, computing correlation functions is unavoidable.

In terms of applications, our results in Section 4.6 obtained by TM-aMD indicate that thermodynamic maps are well suited for inferring equilibrium distributions from swarms of short, independent simulations. We show that TM-aMD is capable of exploring the conformational landscape orders of magnitude faster than REMD, and inferring global equilibrium from samples gathered under local equilibrium. Still, TM-accelerated MD stands to benefit greatly from optimization. And since TMs allow for truly parallel simulations across arbitrary temperatures, TMs are poised to make efficient use of distributed computing resources [179].

As pointed out in Section 4.6, we believe that our implementation of TM-aMD is hindered by the scheme used to adaptively re-sample initial states. The one we employ aims for agreement between the TM and MD-generated distributions. However, in practice it is more productive to balance thermodynamic accuracy with sampling of diverse conformations. Methods such as FAST and REAP quantify this trade-off as a reward function which is to be optimized by the choice of adaptive sampling strategy [180, 181].

Finally, as presented, thermodynamic maps are not informative of kinetics since the samples they generate are independent. In principle, one may use the height of the barrier separating states and recrossing related dynamical corrections to estimate rates [182]. Still, we do not expect this

approach to be useful without some form of further enhanced sampling [183].

Overall, we have presented thermodynamic maps as an exciting way of integrating score-based generative modeling with the theoretical framework of statistical mechanics. We have demonstrated the capability of thermodynamic maps to learn physics by predicting the critical behavior associated with the Ising Model despite limited, deliberately misleading sampling. We also use thermodynamic maps to accelerate Molecular Dynamics (TM-aMD) and show that TM-aMD efficiently infers the equilibrium distribution of two model RNA systems using a fraction of the computational resources required by replica exchange simulations. The TM-aMD method, combined with experimentally measured melting curves, might also prove useful for improving force fields for RNA and similar systems. Based on these results, we believe that thermodynamic maps are suitable for widespread use and have great potential for further theoretical development and computational optimization.

The next chapter builds on this thermodynamic framework by embedding Thermodynamic Maps in a full *ab initio* RNA structure-prediction pipeline, RNAnneal, where they serve as the central component for ranking candidate RNA structures.

Chapter 5: **Ab Initio Prediction of RNA Structure Ensembles with RNAnneal**

Author contribution. This chapter is based on work that I led and carried out in nearly all of its aspects, with the help of my coauthors.

RNA was discovered to be the messenger of genetic information, carrying the code for polypeptide sequences from the nucleus to the ribosome, in 1961. [184] Six decades later, our understanding of RNA has expanded to include noncoding, molecular recognition, and enzymatic functions, establishing the transcriptome as a complex web of functional, interacting RNA molecules that orchestrate gene expression and disease at multiple levels. [185] Interrogation of the transcriptome continues to yield insights that are critical for understanding basic biology and designing next-generation therapeutics. [186, 187] The approval of small molecules such as risdiplam [188] and successes of mRNA vaccines [189, 190] highlight the role of RNA as both a target and therapeutic modality.

Understanding how RNA sequence gives rise to three-dimensional structure and function remains a frontier for theoretical and computational science. Accurate structure models would inform decision making in RNA therapeutic development in a similar manner to proteins, where atomic resolution structure models are frequently considered when evaluating the therapeutic potential of protein targets. These models, however, are rarely available for RNAs. To illustrate, approximately 100 RNA-only structures are deposited in the Protein Data Bank (PDB) annually compared to 10,000 for proteins. [45] Furthermore, most PDB entries are solved by high-resolution methods like nuclear magnetic resonance spectroscopy (NMR), X-ray crystallography, or cryo-electron microscopy. These methods operate by detecting an average signal from the conformational ensemble to generate a structural model, and are therefore most accurate when the ensemble is homoge-

neous. In the case of less ordered structures, the detector receives a superposition of the individual conformations' signals that can be difficult to deconvolve. The structural heterogeneity of RNAs therefore presents a severe limitation to such ensemble averaged methods. New experimental and computational methods are required to accurately characterize such heterogeneous conformational ensembles. [46, 191, 192] Here, we introduce RNAnneal as a computational method for predicting RNA structural ensembles. RNAnneal integrates modeling principles from generative deep learning, statistical physics, and molecular simulations into an *ab initio* conformational sampler and a deep learning score model.

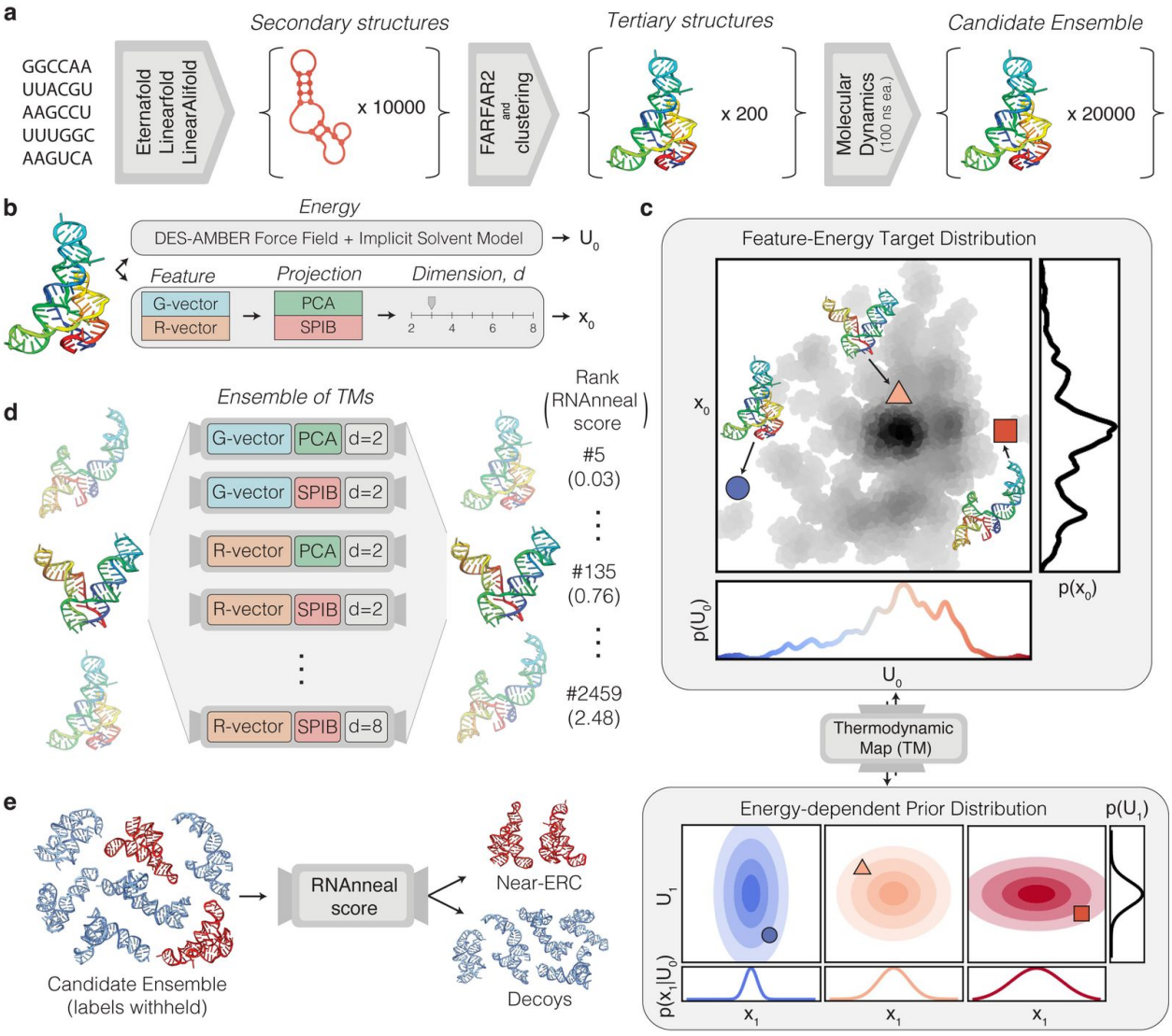


Figure 5.1: **Illustrative overview of RNAnneal. a.** The conformational sampling component of RNAnneal is illustrated. Starting with the primary sequence, a large number of secondary and 3D structures are predicted in succession. The 3D conformations are clustered into representatives before each is simulated for 100ns using implicit solvent molecular dynamics. Together, the simulation frames form a set of diverse candidate conformations that will be scored by RNAnneal. **b.** Energetic features (upper track) and latent representations (lower track) are extracted from each conformation. The latent features are constructed by projecting geometric descriptors (G- and R-vectors) into a reduced space of dimension d using Principal Component Analysis (PCA) and a State Predictive Information Bottleneck (SPIB) model. **c.** The structure of a thermodynamic map [66] (TM) is illustrated: latent representations, $\mathbf{x}_0$ are mapped onto an energy-dependent prior distribution, while the energies, U_0 themselves are mapped onto a unit normal distribution. **d.** TM models are trained on geometric features of dimension d produced by PCA and SPIB. The RNAnneal score is derived from the log-likelihood of the candidate conformations under the models. **e.** The RNAnneal score is suited for the task of classifying experimentally resolved conformations (ERCs) from decoys.

5.1 RNAnneal

RNAnneal is an unsupervised, *ab initio* method for predicting RNA structural ensembles that integrates generative deep learning with statistical physics and Molecular Dynamics (MD). RNAnneal approaches structure prediction in the spirit of a needle-in-a-haystack classification problem: a candidate structural ensemble is generated by broadly sampling secondary structures, assembling them into 3D models, selecting representatives, and performing additional conformational sampling with MD simulations (Fig. 5.1a). If the candidate ensemble contains experimentally resolved conformations (ERCs), then structure prediction amounts to picking them out of the ensemble. To this end, an RNAnneal score is developed based on a physics-inspired generative model, and used to rank candidate structures (Fig. 5.1b-e). The RNAnneal score is parameterized from the candidate conformations by an ensemble of Thermodynamic Maps, a type of generative model that infers the dependence of the conformational ensemble on environmental variables such as temperature or pressure (Fig. 5.1b-c). Thermodynamic maps are probabilistic generative models based on the denoising diffusion framework [86], and have been shown to accurately predict molecular equilibrium distributions from sparse data collected across different thermodynamic conditions. [66, 193, 194] Consistent with these works, one may tailor the RNAnneal score to a user-specified temperature by annealing the ensemble of score models.

An essential quality of any RNA structure prediction method is the ability to generalize beyond ERCs, which form the set of known RNA structures. In computational modeling, ERCs are experimentally validated structure models that are serve as ground-truths for computational structure prediction methods. Most approaches are explicitly parameterized, trained, and validated against these experimentally determined structures. However, ERCs capture only a subset of the full conformational ensemble, and assessing model generalization beyond ERCs is critical yet challenging because the true structural ensembles are unknown. Unlike supervised methods that require labeled training data, RNAnneal learns a score function de novo for each sequence, avoiding issues of overfitting to previously observed structures. Within this framework, one may evaluate

the score model on ERCs and examine if the model selects for them consistently. If they are well predicted, then the highest-scoring decoy structures may be reconsidered as potential alternative conformations.

We compared conformations sampled by RNAnneal to validated ERCs for 16 riboswitches and found agreement in their interaction compositions, measured by relative fractions of stacking, Watson-Crick and non-Watson-Crick contacts. (Fig. 5.2a-b). After simulating the ERCs for 100ns, we found their structures to be dynamic. This informed our decision to use the interaction network fidelity (INF) and root mean squared deviation (RMSD) together as complementary evaluation metrics. RMSD captures global geometric deviation but is sensitive to alignment, while INF captures interaction fidelity independent of alignment (Fig. 5.2c-f). In light of the observed flexibility, we evaluated RNAnneal against the simulated ERCs and found that pseudoknot-free (PK-free) near-ERCs were consistently sampled. Following this evaluation, we demonstrate that the unsupervised RNAnneal score prioritizes ERCs over decoys. For each RNA, 20,000 structures (including the ERC) were sampled, and RNAnneal was employed to select ten representative structures for each of the riboswitches. For 11/16 riboswitches, a representative with RMSD < 10 Å or INF > 0.75 from the ERC was selected. Lastly, we introduce the interaction entropy as a per-nucleotide measure of structural heterogeneity within an ensemble (Fig. 5.3a-b).

5.2 Unsupervised sampling and prediction of RNA structures

RNAnneal predicts RNA conformations by sequentially sampling and scoring candidate structure models. As such, the first step is to assemble a diverse, physically plausible candidate set of 3D models (Fig. 5.1a). With only the primary sequence as input, RNAnneal employs physics-based and bioinformatics programs to generate a large number of secondary structures (Eternafold [195], Linearfold [196] and LinearAlifold [197]; see Secondary Structure Prediction), which are then assembled into all-atom 3D models using the FARFAR2 algorithm (see FARFAR2). [195, 196, 197, 198] Next, 200 diverse, representative models of the 3D structure are selected by clustering using

geometric features and ranking conformations based on their Rosetta scores (see Clustering). To produce a final set of candidate structures, each of the representative conformations is simulated for 100ns via implicit solvent MD (see Molecular Dynamics).

Next, we describe the RNAnneal score, which is extracted from an ensemble of generative models trained on energy values (see Molecular Dynamics) and latent representations of candidate conformations (Fig. 5.1b). To arrive at the latent representations, geometric features (see R- and G-vectors) are extracted from the candidate conformations and subsequently encoded as latent representations by Principal Component Analysis (PCA) and a State Predictive Information Bottleneck [199] (see SPIB) model; the principal components capture the highest-variance geometric features, while the SPIB projections preserve the slowest processes of the MD simulations and learn state boundaries that are metastable over a user-specified timescale.

Then, an ensemble of Thermodynamic Maps [66] (TMs; see Thermodynamic Maps) is trained to estimate the log-likelihood of the candidate conformations based on their energies and latent representations (Fig. 5.1c). [66] In this work, we train 28 TMs on combinations of features (R- and G-vectors) and projections (SPIB and PCA) with 2 to 8 latent dimensions, d (Fig. 5.1d). It is worth noting that the energy values are consistent across all models, even though the latent representations of a given conformation vary from model to model. Finally, for a given conformation, the TM log-likelihoods are averaged to produce the RNAnneal score (see RNAnneal Score). Our evaluations found that the RNAnneal score can reliably distinguish between ERCs and decoys (Fig. 5.1e and Fig. 5.2c-d).

Since the TM models represent the energy scale of the conformational landscape as a temperature, one may anneal the RNAnneal score by adjusting the temperature of the prior distribution of the TMs during the log-likelihood calculation (see RNAnneal Score). We investigated the classification power of the RNAnneal score as a function of temperature and found that ERCs are most prioritized at low temperatures, which is consistent with ERCs lying low on the free energy landscape (Fig. D.14).

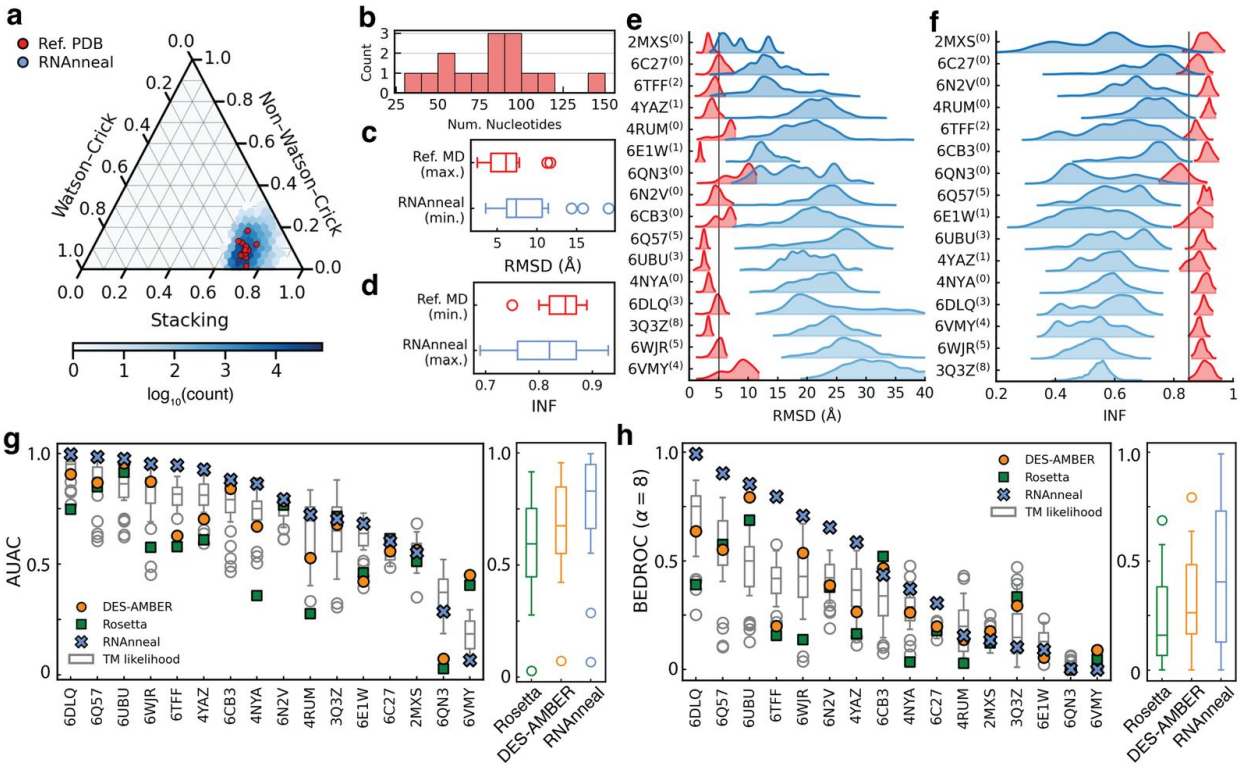


Figure 5.2: **RNAnneal evaluated on the riboswitch dataset. a.** Comparison between the interaction composition of ERC (red) and RNAnneal conformations (blue). **b.** Distribution of sequence lengths for the riboswitch set. **c.** Maximum RMSD values from MD simulations of the ERC are compared to the minimum RMSD among the RNAnneal conformations. (Median [IQR]; Ref. MD: 6.1Å [$4.2 - 7.4$Å]; RNAnneal: 7.4Å [$6.2 - 10.7$Å]). **d.** Minimum INF values from MD simulations of the ERC are compared to the maximum INF among the RNAnneal conformations (Median [IQR]; Ref. MD: 0.85 [$0.83 - 0.87$]; RNAnneal: 0.80 [$0.76 - 0.87$]). **e.** RMSD comparison between simulations of the ERC and conformations predicted by RNAnneal. **f.** INF comparison between simulations of the ERC and conformations predicted by RNAnneal. Conformations with INF≥ 0.85 (vertical line) are classified as ERCs when evaluating the RNAnneal score (g and h). The superscripts denote the number of PK base pairs in the reference conformations (e and f). **g.** Comparison of AUAC values on the ERC classification task (Median (IQR); Rosetta: 0.59 [$0.45 - 0.75$]; DES-AMBER: 0.67 [$0.55 - 0.85$]; RNAnneal: 0.83 [$0.66 - 0.95$]). **h.** Comparison of BEDROC values for the ERC classification task (Median (IQR); Rosetta: 0.16 [$0.07 - 0.38$]; DES-AMBER: 0.26 [$0.17 - 0.48$]. RNAnneal: 0.40 [$0.13 - 0.73$]).

RNAnneal's structure prediction performance is evaluated on a set of 16 experimentally determined riboswitch structures originally curated in Ref. [200] (see Reference Structures; Table D.1 and Fig. SD.1). Riboswitches are typically found in the 5'-untranslated region of bacterial mRNAs and are *cis*-acting regulatory elements that control gene expression by up- or down-regulating transcription or translation. [48] The basis of riboswitch function is conformational:

aptamer domains fold into structurally distinct "bound" (*holo*) and "free" (*apo*) states in the presence or absence of a cognate ligand, which in turn transmits a signal that alters expression. Riboswitches bind their cognate ligands with high specificity through higher order structural motifs, including multi-helix junctions, loop-loop interactions, and pseudoknots. Such motifs are rich in non-WC interactions and their structures are typically challenging to model prospectively. [201, 202, 203]

The Root Mean Squared Deviation (see RMSD) and Interaction Network Fidelity (see INF) are employed to measure structural discrepancy between RNAnneal conformations and ERCs. The RMSD measures the average change in atomic positions after optimal alignment of two conformations, while the INF measures the correlation between their nucleobase interaction networks. Both metrics are required because minor differences in the interaction network (e.g., at a helix junction) may result in large coordinate changes and therefore a large RMSD. The INF is more robust to such variations since no alignment is required.

To summarize, we found that RNAnneal consistently samples PK-free near-ERCs, and its score prioritizes them over decoys. Next, we coarse-grained the RNA structures predicted by RNAnneal into states with the goal of predicting weighted ensembles of representative conformations. Then, to assess conformational heterogeneity, we introduced the interaction entropy to quantify the variability of a given nucleotide's interaction patterns within an ensemble. Finally, we show that annealing the RNAnneal score improved ERC classification for most of the riboswitches

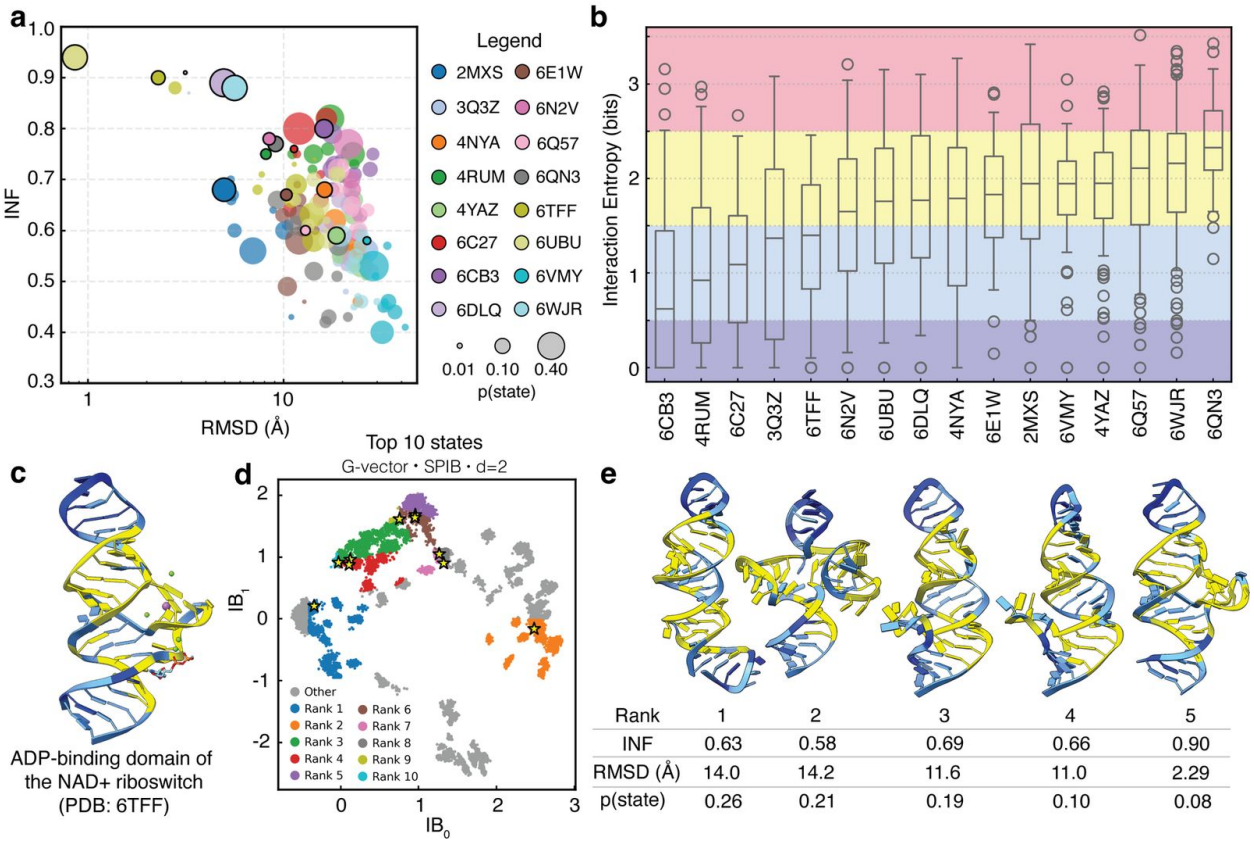


Figure 5.3: **Conformational ensemble prediction with RNAnneal. a.** Ten-state conformational ensembles are shown for each riboswitch. The location of the circles conveys the INF and RMSD of the representative conformer for the state, while the size of the circles conveys the state population. Points circled black indicate the representative with the minimum Deviation Index [204] (RMSD/INF) from the ERC. **b.** Distribution of interaction entropies for each riboswitch. **c.** Reference ERC for the NAD^+ riboswitch colored by the RNAnneal interaction entropy. **d.** State boundaries and representative conformers (stars) projected along the Information Bottleneck (IB). **e.** Five highest-ranked representatives for the NAD^+ riboswitch; the INF and RMSD from the ERC and state population are provided for each representative.

5.3 RNAnneal samples experimentally resolved conformations of pseudoknot-free riboswitches

First, we assess if RNAnneal's conformational sampling procedure (Fig. 5.1a) produces conformations that are similar to ERCs. We compared the interaction composition of conformations sampled by RNAnneal to the reference ERCs and found that 60-80% of interactions are stacking,

10-30% are Watson-Crick (WC), and 0-20% are non-WC for both sets (Fig. 5.2a). The interaction composition of the ERCs is consistent, even though their lengths vary from 28 to 149 nucleotides (Fig. 5.2b).

Next, the structural similarity of RNAnneal conformations to the ERC was evaluated using RMSD and INF. To assess the stability of the ERCs with respect to thermal fluctuations, we simulated each one for 100ns (see Molecular Dynamics) and monitored how the RMSD and INF changed over the course of the simulation (Fig. 5.2c-f). We examined the furthest excursion from the reference conformation, as measured by the maximum RMSD (RMSD_{max}) and minimum INF (INF_{min}), and compared to the closest RNAnneal conformation, as measured by RMSD_{min} and INF_{max}. The median RMSD_{max} among all reference structures was 6.1Å and the median INF_{min} was 0.85; RNAnneal sampled similar conformations, with a median RMSD_{min} of 7.4Å and a median INF_{max} of 0.80 across the set (Fig. 5.2c-d).

Ten of the ERCs contain PKs, which are excluded from the secondary structure prediction algorithms we employ. Consequently, the performance of RNAnneal's candidate ensemble degrades as the number of PK base pairs in the ERC increases (superscript in Fig. 5.2e-f). Only one PK-containing ERC (6TFF [205]) was adequately sampled by RNAnneal (RMSD < 5Å and INF > 0.85) due to conformational sampling provided by FARFAR2 and MD. RNAnneal's sampling was more consistent for the PK-free RNAs, with 5/6 ERCs (2MXS [206], 6N2V [207], 4RUM [208], 6CB3 [209], and 6QN3 [210]) being reproduced.

A Wilcoxon signed-rank test was employed to evaluate the statistical significance of the difference in means between RMSD_{max} and RMSD_{min} (and INF_{min} and INF_{max}), which yielded an ambiguous p-value of 0.08 (0.13 for the INFs). A clearer picture arises when comparing the RMSD and INF values separately for the PK-containing and PK-free ERCs. Here, the test yielded a p-value of 0.01 for the RMSDs (0.02 for the INFs) among the PK ERCs, and a value of 0.30 (0.63 for the INFs) for the PK-free ones. Based on these results, we concluded that PK-free ERCs were generally modeled well by RNAnneal, while PK-containing ERCs were not, highlighting the ongoing challenges of modeling PKs.

Correlation between INF and RMSD values of RNAnneal conformations, computed separately for each riboswitch, yielded an r of -0.49 ± 0.22 ($\mu \pm \sigma$) which was indicative of weak correlation (Fig. SD.11). This reflects the need for both alignment-based and alignment-free measures of structural fidelity for dynamic molecules such as RNA. Indeed, some ERCs exhibited large changes in RMSD during simulation while still maintaining the interaction network (4RUM [208] and 6VMY) and vice versa (6E1W [211]). The INF was observed to be more robust to variations in structure caused by thermal fluctuations compared to the RMSD, which is consistent with previous work [204]. Notably, the non-WC interaction network was found to be highly dynamic for several of the ERCs (Fig. SD.2).

5.4 The RNAnneal score is selective for experimentally resolved conformations

The performance of the RNAnneal score was benchmarked on the task of classifying ERCs and decoy conformations, with comparisons to two other state-of-the-art scores: the Rosetta score and DES-AMBER force field energy (see FARFAR2 and Molecular Dynamics). For this evaluation, we create a new set of structures by adding conformations from the reference simulations to the pool of conformations sampled by RNAnneal, and then ordering the set based on their scores (Fig. SD.10).

Classification performance was quantified by measuring how well each score prioritized near-ERCs (INF ≥ 0.85) over decoys (INF < 0.85) using two metrics: Area Under the Accumulation Curve (AUAC; see AUAC) and Boltzmann-Enhanced Discrimination of Receiver Operator Characteristic (BEDROC; see BEDROC). [212] The AUAC is dominated by contributions from the lowest-ranked models, while BEDROC emphasizes classification accuracy among the highest-ranked structures via a tunable parameter α (equal to 8). Both metrics vary from zero to one, with the latter indicating perfect classification of ERCs over decoys. The RNAnneal score broadly outperformed the Rosetta and DES-AMBER scores on the classification task in both the AUAC and

BEDROC metrics (Fig. 5.2g-h). RNAnneal had a median AUAC of 0.83 which was greater than the 0.67 for DES-AMBER or 0.59 for Rosetta. Similarly, the median BEDROC for RNAnneal was 0.40 compared to 0.26 for DES-AMBER or 0.16 for Rosetta. Statistical significance was assessed using the Wilcoxon signed-rank test, which yielded p-values of 0.02 (AUAC) and 0.02 (BEDROC) when comparing the Rosetta score to DES-AMBER, 0.008 and 0.03 for RNAnneal and DES-AMBER, and 0.004 and 0.008 for RNAnneal and Rosetta. These results provide evidence that the RNAnneal score differs from the Rosetta and DES-AMBER scores at a significance level of 0.05.

There was little variation in AUAC or BEDROC between different features and projections; instead, most of the variation is due to the latent dimension, d (Fig. SD.12). Similar results were obtained when RMSD ≤ 5Å is used as a classification criteria for defining near-ERCs (Fig. SD.13). Interestingly, when the TM models performed well on average, we observed an effect where the ensemble-averaged RNAnneal score had superior classification accuracy to the average, and sometimes even the best model (Fig. 5.2g-h).

Although PK-containing ERCs were not well-sampled by RNAnneal, the RNAnneal score was able to classify them apart from decoys. The ERCs with the five highest RNAnneal AUAC and BEDROC values (6DLQ [213], 6Q57 [214], 6TFF [205], 6UBU [215], and 6WJR [216]) all contained PK base pairs (Fig. 5.2g-h). Still, not all PK-containing ERCs were successfully predicted by the RNAnneal score at the temperature used for the main results ($T = 1$); some ERCs were better prioritized at lower or higher temperatures (Fig. SD.14).

5.5 Annealing improves classification and ranking of experimentally resolved conformations

RNAnneal encodes the scale of the energy landscape via a temperature parameter in the TM algorithm (see Thermodynamic Maps). The structure of the thermodynamic map reflects the following intutition: conformations that are low-lying on the energy landscape are sampled at low temperatures, and therefore ought to be mapped onto a low-variance prior, while those which

are high on the energy landscape and sampled at hotter temperatures should be mapped onto a high-variance prior. To prioritize low-energy conformations, one may evaluate the RNAnneal score with a low-variance generative prior (see RNAnneal Score). The main results are presented for $T = 1.0$, which corresponds to one standard deviation above the minimum potential energy value. However, there were multiple instances where altering the temperature value improved RNAnneal's performance on the ERC classification task.

The AUAC and BEDROC values improved for 11/16 riboswitches as the temperature was annealed from $T = 10$ to $T = 1$ (Fig. SD.14). The AUAC and BEDROC plateau as the inference temperature was decreased further for eight riboswitches (2MXS [206], 6C27 [217], 6DLQ [213], 6N2V [207], 6Q57 [214], 6TFF [205], 6UBU [215], and 6WJR [216]), while classification performance continued to improve down to $T = 0.1$ for three (6E1W [211], 6QN3 [210], and 6VMY [218]). The AUAC and BEDROC values showed little response to annealing for two riboswitches (4NYA [219] and 4YAZ [220]), and while performance degraded slightly for three others (6CB3 [209], 4RUM [208], and 3Q3Z [221]), the overall change in classification performance was relatively small.

5.6 State-informed prediction of conformational ensembles

As stated previously, RNA structure is best understood in terms of conformational ensembles consisting of weighted states and representative conformations. ERCs provide critical and informative snapshots but do not elucidate the full ensemble. In addition to sampling 3D structures, ensemble prediction requires state assignments and the selection of representatives–a process that often relies on intimate knowledge of the system in question and human intuition. RNAnneal automates the process of generating a conformational ensemble by ranking states predicted by SPIB based on their average RNAnneal score, and then selecting the highest-scoring conformation from within each state as a representative structure (see State Ranking).

We predict ten-state ensembles for each of the 16 riboswitches and measure the state popula-

tion and deviation from the ERC (RMSD and INF) for each representative. Note that the combined decoy and ERC dataset from the AUAC and BEDROC evaluation is employed here so that the model has an opportunity to predict ERCs in the ensemble, even if they were not sampled by RNAnneal. Ultimately, we found that RNAnneal predicted near-ERC structures ($\mathrm{RMSD} < 5$Å and $\mathrm{INF} > 0.85$) as representatives for 5/16 riboswitches, and 11/16 had representatives that satisfy a more relaxed similarity criteria ($\mathrm{RMSD} < 10$Å or $\mathrm{INF} > 0.80$) to the ERC (Fig. 5.3a). The top five representative conformers for each riboswitch are reported in Fig. SD.17-D.20.

5.7 Quantification of ensemble heterogeneity via per-nucleotide interaction entropy

The simulations of the riboswitch ERCs revealed their interaction networks to be dynamic, especially in terms of non-WC interaction patterns (Fig. SD.2). Consistent with this observation, the riboswitch ensembles predicted by RNAnneal contained heterogeneous structures spanning wide ranges of RMSD and INF values (Fig. 5.3a). To assess heterogeneity in nucleobase interaction patterns across an ensemble, we develop the per-nucleotide interaction entropy.

The interaction entropy for a given nucleotide in a structural ensemble is the Shannon entropy of its Leontis-Westhof base-base contact classifications (see Interaction Entropy). When the interaction entropy is zero, the corresponding nucleotide is found in the same base-base interaction across all conformations in the ensemble, and as the interaction entropy increases, the nucleobase participates in more varied interactions. When measured in bits, an interaction entropy equal to H for nucleotide i implies that i participates in 2^H interactions across the ensemble; based on this, we bin the interaction entropy around integer values (colors in Fig. 5.3b).

The distribution of interaction entropies for the riboswitch ensembles predicted by RNAnneal indicate that a few of the riboswitches have mostly well-determined interaction patterns (light and dark blue in Fig. 5.3b), but most were predicted to have many nucleotides with variable interaction patterns (yellow and red in Fig. 5.3b). To aid in visualization of heterogeneity, the interaction

entropy may also be used to color each nucleotide using the ranges indicated in Fig. 5.3b, as is done in Fig. SD.17-D.20.

5.8 Evaluation of RNAnneal on experimentally resolved *holo* riboswitch conformations

RNAnneal was evaluated on fourteen *holo* riboswitches representing ligand-bound states that span transition-metal and natural-ligand complexes. Across this diverse set, RNAnneal frequently recovered conformations resembling experimentally resolved structures.

RNAnneal successfully predicted near-ERC representative structures for the transition-metal riboswitches (4RUM [208], 6CB3 [209], and 6N2V [207]). The interaction network of the ERC for the cobalt-binding riboswitch (4RUM [208]) was reproduced in representative conformations predicted by RNAnneal (INF of 0.82 and RMSD of 8.1Å in Fig. 5.3a). Predicted structures for the cadmium- and manganese-binding riboswitches (6CB3 [209] and 6N2V [207], respectively) were slightly less accurate in terms of INF (0.80 for 6CB3 [209] and 0.78 for 6N2V [207]) and RMSD (10.6Å for 6CB3 [209] and 8.5Å for 6N2V [207]). All three riboswitches had regions were the interaction entropy was low, indicating consistent structure across the ensemble (Fig. 5.3b). Visual inspection of the ERCs colored by the RNAnneal interaction entropy reveals that nucleobases with variable interaction patterns are in close proximity to bound transition metals in the *holo* structure (Fig. SD.17-D.20).

The rest of the *holo* ERCs bind natural ligands, which present a greater variety of interaction modes due to their diverse chemistry and spatially extended nature. RNAnneal recovered near-ERCs as representatives for four of the riboswitches, as quantified by an INF value greater than 0.85 and RMSD less than 5Å (3Q3Z [221], 6TFF [205], 6UBU [215], and 6DLQ [213]). Superimposing the RNAnneal interaction entropy on each of the ERCs reveals variable interaction patterns for nucleobases close to the binding site(s) (Fig. 5.3c and SD.17-D.20). The next grouping includes two riboswitches (2MXS [206] and 6QN3 [210]) for which the RNAnneal representatives had

moderate accuracy with respect to the ERC ($0.7 < \text{INF} < 0.85$ and $5\,\text{Å} < \text{RMSD} < 10\,\text{Å}$). The INF for the neomycin sensing riboswitch (2MXS [206]) is low due to the relatively small number of interactions present; RNAnneal predicted multiple conformations with RMSDs between 5 and 10Å, indicating good prediction accuracy. The glutamine II riboswitch (6QN3 [210]) was solved via X-ray crystallography as a dimer, but was predicted as a monomer by RNAnneal. As a result, a large fraction of the bases in the ERC appear to be unpaired but actually form inter-chain interactions. Nonetheless, RNAnneal predicts a representative conformation less than 10Å in RMSD from the ERC. Remarkably few nucleobases had a consistent interaction pattern, indicating that the monomer may not have a dominant stable fold (Fig. 5.3b and SD.17-D.20).

RNAnneal successfully recovered near-ERC representatives for 9/14 holo riboswitches. The remaining 5 systems (4NYA [219], 4YAZ [220], 6E1W [211], 6Q57 [214], 6VMY [218]) showed lower agreement with the ERCs, as quantified by an INF value less than 0.7 and a RMSD greater than 10Å (Fig. 5.3a). Compared to the previous groupings, these riboswitches were predicted to have many nucleotides with high interaction entropy, consistent with a more flexible or dynamic structural ensemble (Fig. 5.3b). In general, nucleobases in close proximity to ligand binding sites have higher interaction entropies (Fig. SD.17-D.20).

5.9 Evaluation of RNAnneal on experimentally resolved *apo* riboswitch conformations

Riboswitch ERCs are typically representative of the *holo* state due to increased conformational flexibility (i.e., heterogeneity) in the absence of a ligand. Consequently, the *apo* conformations of riboswitches are generally poorly characterized. Fortunately, two riboswitch ERCs have been solved in the *apo* state (6C27 [217] and 6WJR [216]). For the S-adenosyl methionine (SAM) III riboswitch (6C27 [217]), multiple representative conformers with INF greater than 0.8 and RMSD between 10 and 15Å were predicted by RNAnneal (Fig. 5.3a). Similarly, RNAnneal predicted a representative conformation less than 5Å in RMSD from the ERC for the flavin mononucleotide

(FMN) riboswitch in the apo state (6WJR [216]). The RNAnneal ensemble for SAM III riboswitch is somewhat homogenous as indicated by lower interaction entropies; the FMN riboswitch however, is more heterogeneous, likely due to PK base pairs present in the ERC which were poorly modeled during the structure sampling step (Fig. 5.3b)

5.10 *De novo* prediction of *apo* riboswitch conformations

While experimentally resolved *apo* riboswitch structures are valuable benchmarks, most systems lack such reference structures, necessitating a *de novo* approach to characterize their conformational landscapes. RNAnneal is uniquely suited for this task since it is an unsupervised model.

We focus on the NAD^+ riboswitch as a representative example and note that similar reasoning may be applied to other RNAs. This choice is motivated by three factors: the *holo* ERC (6TFF [205]; Fig. 5.3c) was predicted *de novo* by RNAnneal (Fig. 5.2e–f), found to have a fragile interaction network (Fig. SD.2), and was prioritized by the RNAnneal score (Fig. 5.2g–h).

Inspection of the location of the ten representative conformations in the SPIB latent space associated with the NAD^+ riboswitch generated using RNAanneal reveals only five distinct clusters. The clusters are represented among the top five (of ten total) representative conformations (Fig. 5.3d). The first- and second-ranked states span the first information bottleneck (IB) component, suggesting that they are endpoints of the slowest structural transition. The representative conformation for the top-ranked state forms a single continuous helix with a six nucleotide overhang, while the second-ranked state forms a cloverleaf structure (Fig. 5.3d). The third-, fourth-, and tenth-ranked states cluster together and exhibit an internal loop that propagates up the 5' side of the helix as the 5' and 3' ends pair together. The internal loop moves further up the helix in the fifth- through ninth-ranked states, where the representatives closely model the holo ERCs (both with and without the PK contacts). The transition from the *holo* ERC to the second-ranked state is unclear, but one may reasonably infer that the transition involves the loop propagating further up the helix.

5.11 Discussion

RNAnneal introduces an unsupervised, *ab initio* framework for predicting three-dimensional structural ensembles of RNA. The strong performance of RNAnneal on RNA structure prediction tasks may be attributed to multiple factors, including a hierarchical approach to sampling the structural ensemble, geometric and dynamics-based structure featurization, and deep learning models that incorporate physics-inspired inductive biases. The high performance of RNAanneal on structurally challenging systems indicates that that RNAnneal will be an essential tool for rapidly predicting three-dimensional conformations of RNA molecules.

Determining how to venture beyond experimental structures of both RNAs and proteins is critical in the post-AlphaFold2 world. RNAnneal generalizes beyond known ERCs by employing *ab initio* sampling to generate training data, and unsupervised deep learning to distinguish between experimentally resolved and decoy structures. Although in this manuscript RNAnneal has been benchmarked exclusively on the challenging case of riboswitches, the method is expected to generalize well since the structure generation pipeline relies exclusively on minimally parameterized physical models which have been previously evaluated on a variety of systems. [156, 198]

RNAs are dynamic, flexible entities containing both persistently structured elements such as helices and dynamic elements such as internal loops. Variability of these structures is captured by RNAnneal's ensemble output, and flexibility hotspots are conveyed by the interaction entropy. One may examine the riboswitch ERCs colored by their predicted interaction entropies in Fig. SD.17–D.20. Visual inspection of the structure datasets reveals that regions of moderate-to-high interaction entropy (yellow and red nucleotides) frequently coincide with the presence of stabilizing molecules such as metabolite ligands (e.g., Neomycin in 2MXS [206] or PreQ1 in 6E1W [211]), metal ions (e.g., Mn^{2+} in 6N2V [207] or Cd^{2+} in 6CB3 [209]), or crystallization additives such as cobalt hexammine (e.g., in 6C27 [217] or 6UBU [215]). Regions with high interaction entropy may be important for ligand binding, allosteric regulation, or functional switching.

While RNAnneal is already demonstrated to be capable of tertiary structure ensemble pre-

diction for different problems, it is a general and highly modular framework with many further directions for improvement. First, magnesium ions that facilitate folding by both neutralizing the phosphodiester backbone and binding to specific sites in folded RNA [43], are not explicitly treated by RNAnneal. Future work should focus on both more appropriately modeling Mg^{2+} implicitly and predicting the location of explicit, bound Mg^{2+}. [47]Second, employing templates for hard-to-sample motifs such as kink turns and multi-helix junctions [222, 223, 224] may improve the quality of the sampled 3D structures, and would not bias the RNAnneal score since it is both unsupervised and parameterized *ab initio*. Third, the RNAnneal score may be extended to environmental variables like pressure and chemical potential, which have already been incorporated into the Thermodynamic Maps framework. [193] Finally, the coverage of the candidate structure pool could be improved by including outputs from PK-aware secondary structure prediction algorithms [225, 226, 227, 228, 229] and candidates models from physics-based [228, 229, 230, 231] and deep-learning [56, 232, 233, 234] tertiary structure prediction methods.

Overall, RNAnneal provides a target-specific method to generate and score 3D RNA structure ensembles from sequence alone that makes ensemble-level RNA structure prediction tractable and interpretable, and will help shine light on the sequence-structural ensemble relationships that underlie RNA diversity and function.

Chapter 6: DiffDock-Glide: A Hybrid Physics-Based and Data-Driven Approach to Molecular Docking

Author contribution. This chapter is based on work that I conceived and carried out, in collaboration with my coauthors.

6.1 Introduction

The previous chapter introduced RNAnneal, a modeling pipeline that uses a series of physics-based models to generate candidate RNA structures and a probabilistic generative model to score them. This chapter explores the reverse arrangement through DiffDock-Glide, a molecular docking method that uses a probabilistic generative model proposes candidate protein-ligand poses and a physics-based model to score them.

Structure-based drug design is a paradigm of drug discovery that uses structural information to rationally design small molecules that bind to targets with high specificity and affinity. [235] Computational approaches have proven indispensable for reducing the vast space of possible molecules to a small number of potentially high-affinity binders. [236, 237, 238] Historically, the most successful computational strategies have been physics-based molecular docking models that explicitly account for energetic [239, 240, 241, 242] and entropic [243] contributions to the binding free energy.

But, in recent years, alternative approaches rooted in deep learning have emerged, all of which share a core assumption that the relevant chemical interactions are represented in structural and evolutionary data and can be implicitly modeled by a neural network. In a foundational

application of deep learning to biology and chemistry, AlphaFold2 (AF2) modeled protein structure to unprecedented accuracy by training an artificial intelligence system on coevolutionary information and the structural contents of the Protein Data Bank (PDB). [55] Since AF2 there has been a dramatic increase in deep learning methods for predicting structures that arise from molecular interactions [56, 62], including ones specifically focused on molecular docking. [63, 64]

However, comparative studies find that, at present, classical docking methods significantly outperform deep learning ones. [244] Deep learning methods struggle to predict binding poses for receptors from families outside the training data [244] and fail to model the molecular interactions that stabilize binding with high fidelity [245], indicating that the amount and diversity of currently available training data is insufficient to model physical chemistry more accurately than first-principles descriptions.

Classical learning theory attributes the performance gap to deep learning models having many more free parameters than are necessary to capture the latent variables. The bias-variance trade-off quantifies an inverse relationship between model complexity and capacity for generalization: the reduction in fitting error (bias) due to introducing more tunable parameters is concomitant with an increased ability to memorize noise (variance) present in the data. [246] Deep learning models operate in the low-bias high-variance regime, and as a result they are prone to overfit their training data and struggle to generalize to out-of-distribution settings. [247]

Models in the natural sciences are less prone to overfitting: free parameters are kept to a minimum through the inclusion of conceptual prior knowledge, i.e. inductive biases. [248] In fact, incorporating task-specific inductive biases into deep learning models is a popular approach to improving generalization ability [249, 250], and generative diffusion models are particularly promising for introducing such biases via modifications to the diffusion dynamics. One may, for instance, define equivariant dynamics [251], restrict the diffusion to a manifold [252, 253], bias the generative dynamics [254], or represent thermodynamic quantities. [66, 193]

In this work, we present a hybrid physics-based and deep learning approach to molecular docking with biophysical and chemical inductive biases. We use DiffDock [64]–a diffusion model–

to sample putative binding poses, then score and refine the poses with Glide [241]– a physics-based molecular docking algorithm. We modify the generative dynamics so that the location of the binding site may be specified in a manner that introduces negligible computational overhead and does not require retraining the model.

We benchmark our method, called DiffDock-Glide, on the Posebusters dataset, demonstrating improved generalization to novel sequences and sampling of near-native binding poses compared to DiffDock alone. We also virtual screen compounds from the DUD-E [255] dataset against AF2-generated receptor structures and obtain enrichment values that broadly surpass those obtained from Glide or DiffDock alone.

The hybrid molecular docking model has two stages. The first stage involves generating a set of potential binding poses using the DiffDock model and a modified generative process that optionally restricts generation to a putative binding site. The second stage involves refining and scoring the generated poses using Glide.

6.2 Binding site-conditional DiffDock

DiffDock is a diffusion model that predicts binding poses for a small-molecule given its chemical identity and a target protein structure. It is a *blind* docking algorithm – anywhere inside or on the target protein's surface is a potential binding site. But, in a realistic structure-based drug discovery setting, one often knows the location of the binding site and would like to explore poses therein. To this end, we bias DiffDock's generative process towards the location of the site using the particle guidance framework. [254] We will briefly summarize the core operating principles of DiffDock before introducing the biasing mechanism.

DiffDock incorporates internal symmetries of molecular conformers as inductive biases. During conformer generation, perturbations to molecular degrees of freedom are represented as a product space $\mathbb{P}$ of groups: $\mathbb{T}(3) \times SO(3) \times SO(2)^m$; $\mathbb{T}(3)$ is the group of 3D-translations; $SO(3)$ is 3D-rotations; and $SO(2)^m$ are 2D-rotations of m torsion angles. Clearly, the actions of $\mathbb{P}$

on a particular conformer will result in new conformers; it is then natural to ask if two arbitrary conformers are connected through an action of $\mathbb{P}$. Indeed, Corso et. al. show that a typical pair of conformers lie on a manifold $\mathcal{M}$ defined by actions of $\mathbb{P}$. [64]

DiffDock is a score-based diffusion model [42]–that is, a stochastic differential equation–with dynamics on $\mathcal{M}$:

$$\mathrm{d}\mathbf{x} = -\lambda(\tau)\left[\mathbf{x} - \sigma^{\mathbb{P}}\mathbf{s}_{\theta}^{\mathbb{P}}(\mathbf{x}, \mathbf{y}, \tau)\right]\mathrm{d}\tau + \sqrt{2\lambda(\tau)\sigma^{\mathbb{P}}}\mathrm{d}\mathbf{B}_{\tau}^{\mathbb{P}}, \tag{6.1}$$

where $\mathbf{B}_{\tau}^{\mathbb{P}}$ is Brownian motion on $\mathcal{M}$, $\lambda(\tau)$ is a time dilation factor to accelerate convergence of the diffusion process, and $\sigma^{\mathbb{P}}$ is the scale of the noise perturbations. The vectors $\mathbf{x}$ and $\mathbf{y}$–containing the positions of the heavy ligand and C_{α} protein atoms respectively–are inputs to a score network $\mathbf{s}_{\theta}$. The score network acts as a time-dependent vector field on $\mathcal{M}$ that generates binding poses from noise vectors by simulating Eq. 6.1 from $\tau = 0$ to $\tau = 1$.

The DockGen benchmark highlighted that DiffDock struggles to predict binding poses for out of distribution binding pockets. [256] To address this, we condition Eq. 6.1 on a putative binding site by biasing the translational component of the diffusion with the distance between the ligand's center of geometry $\mathbf{r}$, and the center of the binding site $\mathbf{r}_0$:

$$\mathbf{s}_{\theta}^{\mathbb{T}(3)}(\mathbf{x}, \mathbf{y}, \tau) \to \mathbf{s}_{\theta}^{\mathbb{T}(3)}(\mathbf{x}, \mathbf{y}, \tau) - \gamma\nabla\Phi(\mathbf{r}, \mathbf{r}_0). \tag{6.2}$$

A flat-bottomed harmonic potential Φ characterizes the geometry and location of the binding site:

$$\Phi(\mathbf{r}, \mathbf{r}_0) = \begin{cases} 0 & \text{if } d \leq r_{\min} \\ (d(\mathbf{r}, \mathbf{r}_0) - r_{\min})^2 & \text{if } r_{\min} < d < r_{\max} \\ d(\mathbf{r}, \mathbf{r}_0) \cdot (r_{\max} - r_{\min}) & \text{if } d \geq r_{\max}. \end{cases} \tag{6.3}$$

The function d measures the distance between $\mathbf{r}$ and $\mathbf{r}_0$, $r_{\min}$ is the inner radius, and the outer radius $r_{\max}$ limits the gradient for numerical stability. The strength of the potential is controlled by γ.

When $\gamma = 0$ the generative process in Eq. 6.1, is *unconditional*, i.e. the center of geometry $\mathbf{r}$ freely explores the entire protein surface; in contrast, when $\gamma > 0$, the generative process is *conditional* and the center of geometry is biased by Φ towards a spherical region of radius r_{min} centered around $\mathbf{r}_0$. The orthogonality of the groups factoring $\mathbb{P}$ guarantees that the rotational degrees of freedom of the generative process are unaffected by the translational bias.

6.3 DiffDock–Glide

Since researchers are often interested in a small subset of representative poses, the samples from DiffDock ought to be scored and rank-ordered based on their plausibility. The original DiffDock implementation trains a separate neural network, called a confidence model, to predict a score that indicates if a pose is likely to be less than 2Å in root mean square deviation (RMSD) from a crystal pose. We take a different approach in using the ranking mechanism to introduce further inductive biases and evaluate pose quality without reference to the generative model or its training dataset.

Buttenschoen et. al. have previously shown that subtle features of real binding poses that are absent from the generated poses can be recovered by minimization under a molecular mechanics force field. [244] Accordingly, we use Glide's post-docking minimization pipeline to score and refine the generated poses. The pipeline optimizes ligand-receptor and intramolecular ligand interactions through rigid-body roto-translations and rotatable bonds optimizations under the OPLS-AA force field. [241] The minimized complexes are then ranked using Glide's scoring function, which has been empirically parameterized to approximate the binding free energy by rewarding or penalizing interactions, e.g. hydrogen bonding, van der Waals, hydrophobic, electrostatic interactions and internal strain energy. [241]

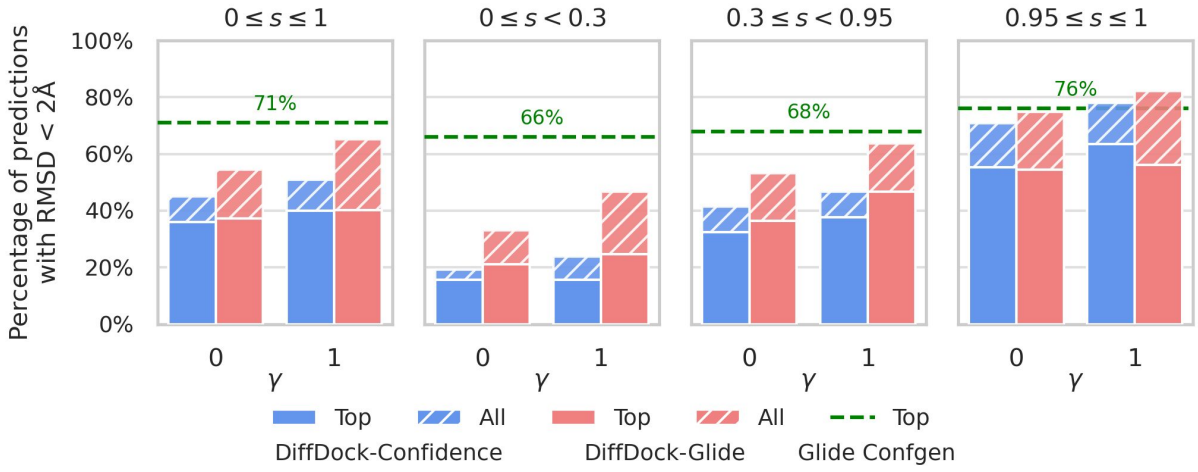


Figure 6.1: **Performance of DiffDock-Glide on the Posebusters dataset.** The percentage of Posebusters (PB) targets that have an RMSD within 2Å of the crystal pose is presented for three docking methods: DiffDock-Confidence, DiffDock-Glide, and Glide Confgen. For the DiffDock-based methods samples are generated with $\gamma = 0$ (no conditioning) and $\gamma = 1$ (binding site conditioning). The percentage of targets for which the top-ranked pose by the confidence model and Glide score (solid bars) is crystal-like is compared to the lowest RMSD selection among all poses (hatched bars). Glide Confgen (green line) is presented as a baseline. The leftmost panel summarizes performance for the entire PB dataset, and the remaining panels stratify the PB dataset by sequence similarity (s) to PDBBind General Set v2020.

6.4 Applications

We explore two applications of the hybrid DiffDock–Glide model: pose prediction and virtual screening.

- Section 6.4.1 evaluates DiffDock-Glide's molecular docking performance on the Posebusters dataset. Our results indicate that binding site conditioning and energy minimization improve generalization to receptors outside the training set and increase the fraction of targets for which a pose less than 2Å in RMSD from the crystal pose is sampled. We also demonstrate a potential use case by using the $\gamma = 1$ model to dock against an allosteric binding site that the $\gamma = 0$ model does not recognize.

- In Section 6.4.2 we use DiffDock-Glide to virtual screen binding and non-binding compounds from the DUD-E library against receptor structures generated by AF2. We report enrichments

surpassing those obtained from DiffDock or Glide alone.

6.4.1 Molecular docking

Molecular docking aims to predict the binding pose of a small molecule with a receptor. The task poses a significant challenge due to the diversity of physical and chemical interactions that contribute to binding stability. We benchmark DiffDock-Glide on the Posebusters molecular docking dataset. [244]

Posebusters benchmark

The Posebusters (PB) dataset is a collection of protein-ligand complexes that is curated to serve as a benchmark for deep learning docking methods: PDBBind General Set v2020. [244] The dataset provides 308 diverse, high-resolution complexes that are not present in PDBBind v2020. Besides the curated set of complexes, Posebusters implements consistency checks that valid binding poses must satisfy. These checks consider factors such as steric clashes, internal energy, and chemical consistency. A pose that passes all these checks is classified as PB-valid. Similarly, we define a Glide-valid (G-valid) condition by determining whether a pose meets Glide's internal consistency requirements. Since Glide rejects poses with extreme values in the terms of its scoring function, any pose that receives a score from Glide is considered G-valid.

To compare the performance of different ranking methods we sample 40 poses for each PB target using DiffDock with $\gamma = 0$ and $\gamma = 1$. The poses are then ranked by the confidence model and Glide score after energy minimization. For the $\gamma = 1$ models, residues within 5Å of the crystallized ligand are considered part of the binding site, and their coordinates parameterize the binding site in Eq. 6.3: $\mathbf{r}_0$ is the center of geometry, $r_{\min}$ is three times the standard deviation of the pairwise inter-residue distances, and $r_{\max}$ is $r_{\min} + 10$Å.

In Figure 6.1 we summarize the performance of DiffDock-Confidence (blue), DiffDock-Glide (red), and Glide Confgen (green) on the PB benchmark. The solid bars display the percentage of PB targets for which a crystal-like (<2Å RMSD) structure is the top ranked pose by each method,

while the hatched bars denote the lowest RMSD selection among all sampled poses. The dashed green line denotes the performance of Glide Confgen, indicating that both DiffDock-Glide and DiffDock-Confidence fall short of state-of-the-art.

We observe that $\gamma = 1$ has the expected effect of yielding a greater fraction of native-like poses compared to $\gamma = 0$. Likewise, we find that energy minimization increases the fraction of targets for which a crystal-like pose is sampled. Consistent with previous studies that identified artifacts in samples from DiffDock, we find that less than 0.1% of the sampled poses are G-valid without minimization. [244, 245] The most striking result is that the fraction of native-like poses selected according to the Glide score (solid red bars) is far below the ideal value (hatched red bars), reflecting the limitations of pose refinement via energy minimization.

In the remaining panels we stratify the results by maximum sequence similarity, s, to proteins in PDBBind v2020. We find that higher sequence similarity correlates with more accurate predictions, and the positive effect of energy minimization and binding site conditioning is greatest when $s < 0.95$. Notably, both DiffDock-Glide and DiffDock-Confidence are capable of sampling sub-2Å poses for a similar fraction of targets compared to Glide Confgen, provided that they have high ($s \geq 0.95$) sequence similarity.

Docking into Allosteric Sites

Traditional small-molecule drug design operates on targeting the primary orthosteric binding site of a receptor to modulate its activity. Orthosteric sites typically exhibit well-defined, highly conserved pockets that bind endogenous signaling molecules with high specificity and affinity. [257] However, certain "undruggable" proteins lack well-defined binding pockets, posing a significant challenge for traditional drug discovery. Allosteric modulation offers a compelling alternative by targeting allosteric sites that are spatially separate from the orthosteric site, but can nonetheless affect the protein's function upon ligand binding. [258] Targeting allosteric sites in G-protein coupled receptors, for instance, has led to improved target selectivity and reduced off-target effects. [259]

Although allosteric sites have therapeutic potential, current deep learning approaches remain focused on orthosteric targets. DiffDock's training set contains both ortho- and allosteric complexes, but is dominated by the former due to the historical bias towards orthosteric drug discovery and limitations of structure probing techniques like X-ray crystallography, which readily resolve well-defined orthosteric sites but struggle to capture the dynamic and flexible structures of allosteric sites. [259]

In this section, we demonstrate how biasing DiffDock's generative process enables targeting of allosteric sites. We design a cross-docking experiment centered around a β-2 Adrenergic G-coupled protein receptor domain containing both ortho- and allosteric sites. The reference complex (PDB: 6N48) is a *holo* crystal structure of the β-2 Adrenergic receptor in complex with two ligands: BI-167107 is bound to the orthosteric site while Compound-6FA (KBY) occupies the allosteric site, which is shallow and lies exposed on the membrane interface of the receptor. [260] The ligand of interest, KBY, engages in π–π stacking interactions with side chain residue PHE1133 and interacts with hydrophobic residues on the membrane-exposed face of the receptor's helical bundle. The target (PDB: 6KR8) is an NMR-derived model containing an ensemble of 10 conformers that represent the *apo* state. [261] We select the first conformer as a representative *apo* structure, and backbone alignment with the *holo* conformer reveals an RMSD of less than 2Å. It is worth noting that the *holo* structure is present in the DiffDock training set whereas the *apo* structure is not, though it is included in 2019 version of the PDB.

Forty poses of KBY docked to the *apo* conformer were sampled using DiffDock-Confidence ($\gamma = 0$) and DiffDock-Glide ($\gamma = 1$) with the biasing potential centered on the allosteric site. Figure 6.2a depicts the distribution of poses sampled by both methods. The orange sphere sits at the centroid of the poses sampled by DiffDock-Confidence, all of which occupy the orthosteric site, while the blue sphere is the centroid of the poses generated by DiffDock-Glide, all of which are within 5Å of the allosteric binding site after aligning the *holo* ligand (yellow sphere). In this example, the unguided DiffDock-Confidence model always docks KBY at the orthosteric site. The introduction of a simple biasing potential to the generative process in DiffDock-Glide enables one

to target the allosteric site.

Figure 6.2b depicts a pose sampled by DiffDock-Glide ($\gamma = 1$) alongside the aligned *holo* (orange) and *apo* (purple) allosteric sites. The *apo* intracellular loop 2 (THR146 and LYS147) is in a different orientation relative to the *holo* conformation, resulting in a narrower binding site. Consequently, we expect the final refined pose to deviate from the *holo* ligand. Indeed, Glide resolves clashes with THR146 and LYS147 during post-docking minimization by including protein side chain residues that are not present in DiffDock, and, upon aligning the *holo* binding site to the *apo* conformer, we find that the DiffDock-Glide pose deviates from the *holo* pose. Still, a primary interaction is shared between the *holo* and DiffDock-Glide poses: both form π-π stacking with residue PHE133 (blue dotted line). The DiffDock-Glide pose forms additional stabilizing interactions between the ligand and LYS147: a hydrogen bond (yellow dotted line) and a salt bridge (magenta dotted line) form between the ligand's carboxylate and the positively charged ϵ-amino group on the side chain residue.

6.4.2 Virtual screening

In silico hit identification uses computational methods to discover potential drug candidates, or hits, from chemical libraries. A successful virtual screening effort effectively prioritizes binding ligands (actives) over non-binders (decoys). Two classes of techniques are often used to screen large libraries for compounds likely to bind to a target protein.

Ligand-based methods are typically used for lead compound optimization when the structure of the target protein is unknown. They rely on chemical features of known actives to search for compounds with similar binding modalities. Techniques like quantitative structure-activity relationship (QSAR) are particularly effective in prioritizing biologically active compounds by using properties of other known actives. [262]

On the other hand, structure-based approaches like molecular docking rely on high-resolution protein structures to predict the binding interactions of compounds. Classical docking tools explore potential binding modes by optimizing the orientation of the candidate ligand in the protein's

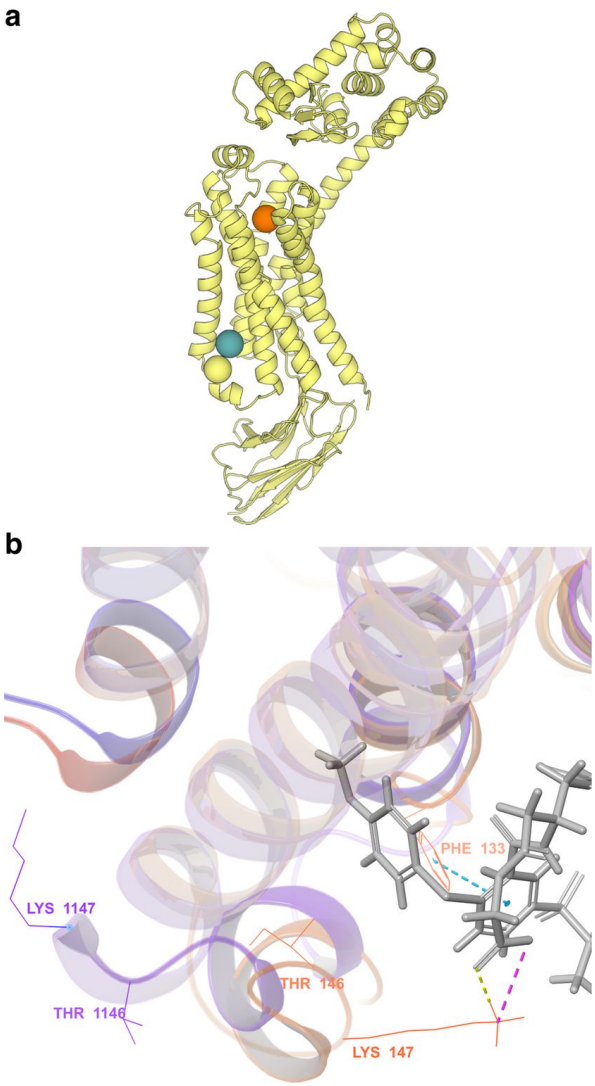


Figure 6.2: **Targeting the allosteric site of β-2 Adrenergic receptor with ligand KBY.** (a) The *holo* structure (PDB: 6N48) is shown. DiffDock-Confidence ($\gamma = 0$) samples all 40 generated poses at the orthosteric site (orange sphere), while DiffDock-Glide ($\gamma = 1$) samples poses (denoted by the blue sphere) within 5Å of the *holo* allosteric site (yellow sphere). (b) The allosteric sites of the *holo* (orange) and *apo* (purple) receptors are aligned and superimposed alongside a pose sampled by DiffDock-Glide ($\gamma = 1$). Stabilizing interactions recovered by Glide are indicated by dotted lines.

binding site according to a scoring function. Well-known tools like AutoDock Vina [239, 263], DOCK [242], Glide [241], and GOLD [240] use distinct algorithms for both pose generation and scoring. [264] In the following section, we explore if DiffDock-Glide is suitable for structure-based virtual screening.

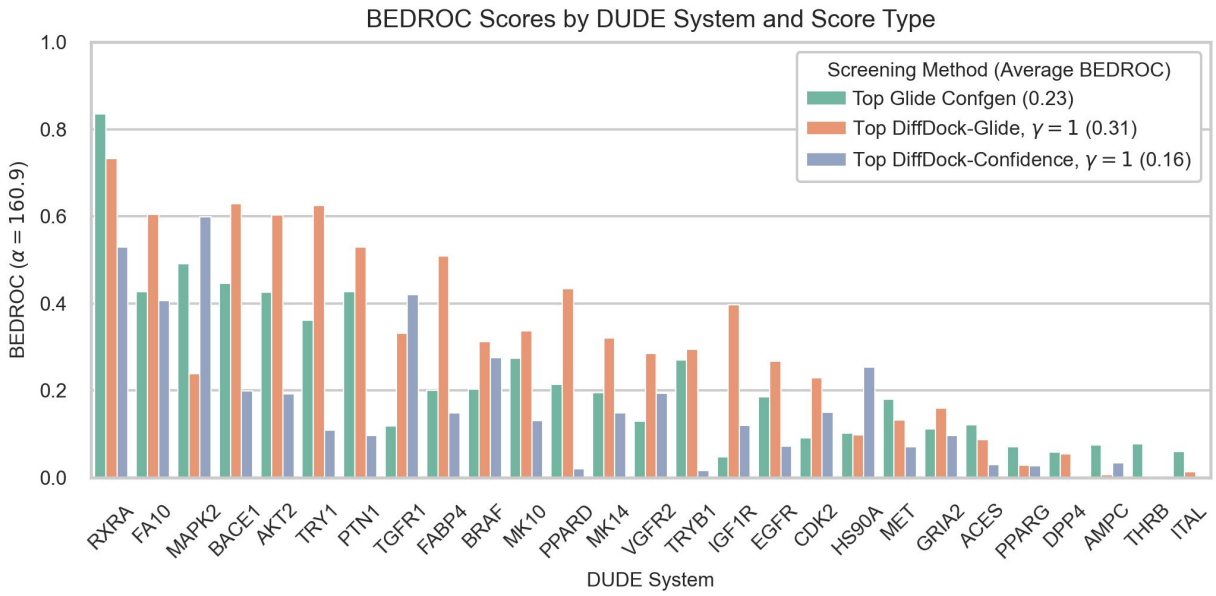


Figure 6.3: **Target-specific enrichment for the AF2-DUD-E virtual screening benchmark.** BEDROC ($\alpha = 160.9$) enrichments are displayed for the 27 AF2-DUD-E target. Five poses are sampled per compound by DiffDock, and scored by the confidence model and Glide's post-docking minimization pipeline. Compounds are ranked according to the top confidence (blue) and Glide (orange) score among the five poses for the BEDROC calculation. BEDROC enrichments from Glide Confgen (green) are provided as a reference.

AF2-DUD-E benchmark

Most molecular docking tools perform best on high-quality ligand-receptor structures when searching for binding poses. DiffDock, however, is unique in that it infers binding poses solely based on the C_α coordinates. Accordingly, we believe that DiffDock-Glide should perform well when screening against low-quality receptor structures, e.g. those with misplaced sidechains. To test our hypothesis, inspired by a previous study that showed AF2-generated receptor structures are not generally suitable for virtual screening with molecular docking [265], we screen active and decoy compounds in the DUD-E dataset against receptor structures generated by AF2 using DiffDock-Confidence, DiffDock-Glide, and Glide Confgen.

The DUD-E dataset is a standard benchmark in virtual screening for evaluating a method's ability to distinguish between actives and decoy compounds. [255] It contains proteins from chemically diverse families, each with a set of validated active compounds and physicochemically

similar but structurally distinct decoy compounds.

Screening Method	EF (↑)	BEDROC (↑)
DiffDock-Confidence ($\gamma = 0$)	1.07	0.02
DiffDock-Confidence ($\gamma = 1$)	9.24	0.16
Glide Confgen	13.1	0.23
DiffDock-Glide ($\gamma = 0$)	15.5	0.28
DiffDock-Glide ($\gamma = 1$)	**16.2**	**0.31**

Table 6.1: Average enrichments from virtual screening DUD-E compounds against AF2-generated receptors. The EF values are computed with a top 1% cutoff. Likewise, the BEDROC values are computed with $\alpha = 160.9$ to emphasize enrichment in the top 1% of rankings.

We screen against the 27 structures from the AF2 database for which the corresponding DUD-E structure does not contain cofactors at the binding site or clashes with the *holo* ligand. The structures are prepared using the truncation procedure outlined by Zhang et. al. [265] For each structure-compound pair, five poses are sampled, minimized, and scored with the same parameters as the docking study in Section 6.4.1, except that residues constituting the binding site are chosen based on the DUD-E crystal structure and $r_{\min}$ and $\mathbf{r}_0$ are derived from the corresponding residue geometry in the AF2 structure.

Table 6.1 reports the average EF [266] and BEDROC [212] ($\alpha = 160.9$) enrichments. Intuitively, EF1% measures how concentrated actives are in the top 1% of rankings compared to the entire set of compounds. Similarly, BEDROC ($\alpha = 160.9$) non-linearly emphasizes early enrichment due to actives among the highest ranked compounds. For both measures, a larger value indicates a greater concentration of active compounds among the top ranks. The disproportionately low enrichment values for DiffDock-Confidence ($\gamma = 0$) are due to it being a blind screening method. The corresponding DiffDock-Glide ($\gamma = 0$) model is not blind since Glide penalizes the rank of poses generated outside the binding site. The top performing DiffDock-Glide ($\gamma = 1$) model has the most inductive biases among all the DiffDock-based methods, and even outperforms Glide Confgen. Figure 6.3 compares enrichment on a per-target basis for Glide Confgen, DiffDock-Glide ($\gamma = 1$), and DiffDock-Confidence ($\gamma = 1$). While the maximum observed enrichment across all targets from Glide Confgen, DiffDock-Glide ($\gamma = 1$) provides the greatest enrichment for over half

of the targets. It is also noteworthy that the purely data-driven DiffDock-Confidence ($\gamma = 1$) model outperforms Glide Confgen for seven targets.

Case Study for IGF-1R

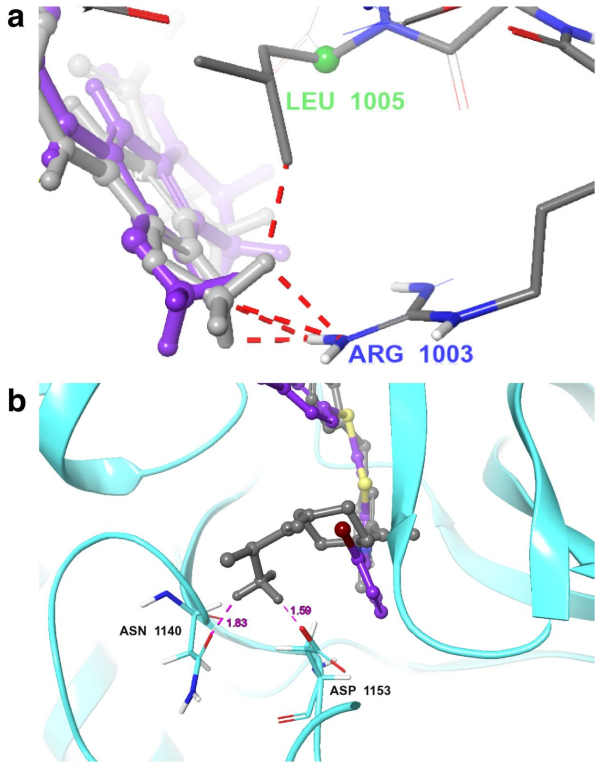


Figure 6.4: **Analysis of poses sampled for IGF-1R** (a) DiffDock-Glide poses before (gray) and after (purple) minimization are shown for ligand CHEMBL461876. Interatomic distances less than 3Å (red dotted lines) indicate clashes between the receptor and DiffDock pose at LEU1005 and ARG1003. (b) The DiffDock-Glide pose (purple) for CHEMBL251795 fails to establish a potent bidentate salt bridge with residues ASN1140 and ASP1153 in the active site. The Glide-Confgen pose (gray) is correctly oriented to form the salt-bridge. Interactions with interatomic distances measuring less than 3.5Å are depicted as magenta dotted lines.

DiffDock-Glide demonstrates significantly higher enrichment of active compounds compared to Glide Confgen for the IGF-1R virtual screening target (see Figure 6.3). Glide scores of the top 1% of ranked compounds range from -9.5 to -7.6 for Glide Confgen and -9.1 to -6.9 for DiffDock-Glide ($\gamma = 1$), indicating that both methods sample favorable binding poses. For reference, scores greater than -6 are generally associated with poor docking poses. [241]

Although the ranges of scores are similar for both methods, the chemical composition of their top 1% compounds differs. DiffDock-Glide ranks 39 active compounds with 33 unique scaffolds among the top 1%. Notably, 91.4% of the top 1% compounds contain pyrimidine-pyrrole cores. Glide Confgen, on the other hand, recovers 11 actives with 11 unique scaffolds in the top 1% and does not show preference for any particular DUD-E chemical series: only two of the top-ranked actives have pyrimidine-pyrrole cores, and the others include, for example, thiazoles and benzimidazoles. Even though the composition of the top 1% of compounds is different from DiffDock-Glide, Glide Confgen still assigns valid (i.e. negative) scores to all the top actives identified by DiffDock-Glide, which suggests that the Glide score is unable to prioritize actives over decoys in this instance. A more detailed analysis reveals that Glide Confgen fails to model the formation of buried hydrogen bonds with nitrogen acceptors. The scores for IGF-1R are dominated by hydrophobic contributions that are not balanced by suitable penalties for protein desolvation. This is a known limitation of Glide Confgen and is addressed by GlideWS, formerly known as WScore. [243]

To demonstrate the improvement that pose minimization brings to DiffDock-Glide, we analyze performance for active compound CHEMBL461876 with and without minimization. Figure 6.4a depicts the un-minimized (gray) and minimized (purple) poses. The un-minimized pose undermines interpretation of the binding mode due to an over-packed binding pocket and unrealistic contacts with residues ARG1003 and LEU1005. Atom C17 of the ligand clashes with the guanidinium group of ARG1003 and CD2 of LEU1005 due to interatomic distances less than 3Å. Pose minimization improves the rank of CHEMBL461876 from 372 to 3 by resolving the steric clashes.

Still, there are instances where Glide Confgen successfully models essential interactions that DiffDock-Glide fails to capture. We focus on active compound CHEMBL251795, which is stabilized by a buried salt-bridge formed between the protonated amide and charged carboxylate on residue ASP1153, and a hydrogen bond between the ligand and ASN1140. Figure 6.4b depicts poses sampled by Glide Confgen (gray) with the bidentate salt-bridge and DiffDock-Glide (purple) without this key interaction. The top Glide Confgen pose is ranked as the top active with a score

of -9.5. DiffDock-Glide does not sample the salt-bridge, and although minimization proceeds successfully, it yields a comparatively poor score of -4.7.

6.5 Discussion

In this work we have developed a hybrid approach to molecular docking called DiffDock-Glide that improves the out-of-distribution performance of data-driven docking. The core idea is to use a deep generative model to sample from a learned distribution of binding poses and subsequently relax the sampled poses with a physics and chemistry-based energy function that contains inductive biases. We conclusively demonstrate that providing such biases dramatically improves the performance of DiffDock. The primary limitations of DiffDock-Glide are sampling crystal-like poses, especially for out-of-distribution receptors, and selecting a crystal-like pose from the sampled distribution.

We found that the confidence and Glide scores are unable to reliably select for such poses even if they are present among the generate samples. Physics-based pose refinement beyond energy minimization may lead to more accurate pose ranking. We also found that DiffDock-Glide is especially suitable for virtual screening against AF2 receptor structures, though it should be noted that all the screened receptors except *PPARD* are found in DiffDock's training set. At the time of writing, there is no virtual screening equivalent of the Posebusters set, so the effects of overfitting could not be systematically investigated. Though based on the Posebusters analysis, we anticipate that DiffDock's performance will degrade for out-of-distribution targets, and the decline can be mitigated via the inclusion of appropriate biases.

Co-folding models that predict the structure of entire complexes have garnered much attention since the publication of DiffDock. Nevertheless, we still believe DiffDock to be the most suitable base model for our approach: while models like Boltz1 [267] and AlphaFold3 [56] are indeed diffusion models, their dynamics unfold in the full coordinate space and are less straightforward to control compared to DiffDock's torsional representation which respects molecular symmetries.

On a final note, we would like to emphasize that our modifications to DiffDock's generative dynamics are independent of the model's weights and therefore do not require retraining to implement, nor do they present substantial computational overhead during sampling. Overall, we introduce DiffDock-Glide as a practical solution for structure-based drug discovery applications that incorporates well-validated biophysical principles, maintains the advantages of learned representations, and is more interpretable compared to fully deep learning approaches.

Chapter 7: **Outlook**

This thesis argues that inductive biases should be an essential component of generative models in the molecular sciences. The work presented here explores two means of introducing physics-inspired biases: first, by encoding thermodynamic information in the prior distribution of a diffusion model, and second, by combining physical and generative models in sequence-to-structure modeling pipelines. A few paths branch from these two initial directions.

The first is to extend Thermodynamic Maps to variables other than temperature. Temperature was the natural starting point because its effect on equilibrium fluctuations can be represented cleanly through the variance of a Gaussian prior, but the logic extends to other thermodynamic variables. Separately, exponentially tilted Thermodynamic Maps have been developed, exploring similar representations of pressure and chemical potential through structured changes in the prior [193]. Still, several questions remain open. A central challenge is how to evaluate thermodynamic generalization: benchmark systems should have known or independently computable thermodynamics and tunable phase behavior. Spin glasses, simple fluids, polymers, and well-understood molecular systems may provide useful test beds. A second challenge is determining when a heuristic representation is sufficient. The energy-dependent prior used in RNAnneal, for example, is not a direct physical temperature in the same sense as the temperature variable in Thermodynamic Maps, but it nevertheless provides a useful ranking signal for RNA conformations. Understanding the relationship between formally grounded thermodynamic encodings and practically useful heuristic encodings will be important for broadening the method while preserving its connection to thermodynamics and statistical physics.

The second is to generate latent representations of molecular data. Recent latent-variable

extensions such as latent thermodynamic flows and diffusive State Predictive Information Bottleneck suggest that representation learning and generative modeling can be mutually reinforcing, especially in the low-data regime that characterizes most molecular problems [268, 269]. A useful latent space should separate metastable states and preserve slow dynamical processes. Conversely, a generative objective can regularize the representation by forcing it to support faithful density estimation and sample generation. Integration of representation learning with thermodynamic prior distributions developed in this thesis may yield models that simultaneously identify collective variables, estimate free-energy landscapes, and generate ensembles under unseen conditions.

A third, applications-oriented direction is to use RNAnneal to explore the structural landscapes of therapeutically relevant RNAs, especially those with limited co-evolutionary information or heterogeneous structures. Because RNAnneal learns a sequence-specific score *de novo* from physics-based sampling, it does not depend on co-evolutionary information or deep phylogenetic signal in the way that popular end-to-end structure prediction models do. In principle, that should make it applicable not only to natural RNAs with limited homologous sequence information, but also to synthetic or regulatory RNAs designed for therapeutic function. Two concrete examples are VEGF-associated regulatory RNA structures, where the goal is to characterize the disordered and condition-sensitive structural landscape [270], and the MALAT1 triple helix, where structurally informed conformational ensembles could guide small-molecule screening against the unusual fold [271, 272]. These applications would test whether the proposed methodology can deliver insights for targets whose functional importance lies in conformational heterogeneity or noncanonical structure.

The long-term goal is not only to predict RNA structure ensembles more accurately, but to translate them into mechanistic insight. The road ahead will entail expanding the structural vocabulary of RNAnneal beyond the motifs emphasized in the presented benchmark, including pseudoknots, G-quadruplexes, triplexes, and other noncanonical folds; linking predicted states to ligandability, accessibility, and function; and incorporating experimental measurements as modeling constraints. If successful, such direction could push sequence-to-structure models toward becoming an integral part of the decision-making process for therapeutic discovery.

Appendix A: **Supplementary Material for Chapter 2**

A.1 Architectural Details

We now provide details of the specific flavors of models implemented in this work. For the Denoising Diffusion Probabilistic Model we selected a variance preserving (VP) diffusion framework, trained via denoising score matching, or the L_2 loss between the predicted unnoised image and true unnoised image. We chose the VP diffusion framework over others, such as the variance exploding (VE) framework, to preserve continuity with other ongoing projects in our group [66] requiring the VP construction to relate the variance seen in the model with thermodynamic fluctuations. Our diffusion process transports samples from the empirical data distribution to a normal distribution with zero mean and unit variance during training. The neural network at the model's center is the U-Net convolutional neural network architecture [117]. This architecture is widely used across diffusion models, specifically for image generation, as it emphasizes local correlations by design.

We chose to optimize the *model dimension* hyperparameter by jointly assessing the KL divergences resulting from different choices of model dimension over a test dataset for the DDPM model and the CFM model, which uses the same architecture. The same optimized value, 128, was chosen for each, and as a result, the two models have an equal number of learnable parameters. Neural Spline Flows use a different neural network architecture, and hyperparameter optimization must be done independently. Figure A.1 shows the hyperparameter tuning results for the three generative models. The 'layers' parameter of NS corresponds to the depth of the network, whereas the 'model dimension' parameter of CFM and DDPM corresponds to the width and, implicitly, the depth of the network. The base code for the DDPM implemented is available at `https:`

//github.com/tiwarylab/simple-diffusion.

For the Conditional Flow Matching model, we chose the 'Exact Optimal Transport' plan (OT-CFM) [273]. The CFM framework allows for various transport objectives that inform a conditional vector field used to transport samples from the prior to the target distribution, which a neural network learns to approximate via the L_2 loss. While OT-CFM is generalizable to any base distribution, we chose the standard normal distribution to match the base distribution of DDPM. We selected this transport plan in particular due to its claimed efficiency over other transport plans for high-dimensional data, a quality its authors attribute to the straight, easily integrable trajectories resulting from the imposition of the optimal transport objective. We chose Conditional Flow Matching over other generalized continuous flow-based frameworks due to the repository's ease of use.

Initially, we set Conditional Flow Matching to use a Multi-Layer Perceptron (MLP) to regress the learned conditional vector field, but we decided to modify the code to work with the same U-Net as DDPM. This choice made the comparison more fair, and we selected the same model dimension hyperparameter for both models based on tuning tests. The base code for the CFM model implemented is available at `https://github.com/atong01/conditional-flow-matching`.

The category of discrete normalizing flows to which Neural Spline Flows belong is perhaps the largest, containing any invertible transformation (or series thereof) between the base and target distributions. We initially implemented real-valued non-volume preserving normalizing flows (Real NVP) but opted to change to NS flows for a more recent model known for its series of expressive rational quadratic spline transformations, which serve as the affine or additive transformations in coupling layers. As a normalizing flow, NS is trained by maximizing the total log-likelihood of the training data. This objective guides learning the spline parameters at each discrete node, parameterized by the ResMADE (Residual Masked Autoencoder for Distribution Estimation) architecture. We opt to tune the 'layers' parameter, a proxy for the depth of the model. The original NS base code is available at `https://github.com/bayesiains/nsf`. However, we used the `normalizing-flows` repository, a valuable collection of several normalizing flows models, available at `https://github.com/VincentStimper/normalizing-flows`.

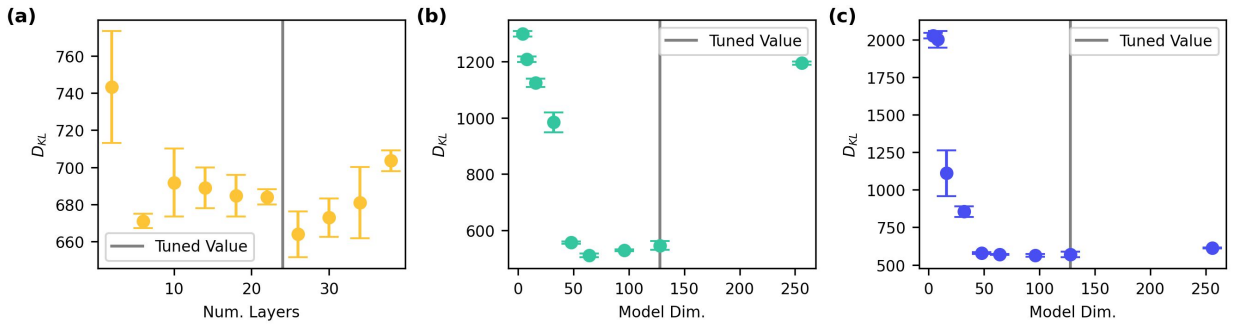


Figure A.1: **Hyperparameter tuning for each model.** a) Tuning of the 'layers' parameter of NS. b) c) Tuning of the 'model dimension' parameter of CFM and DDPM respectively.

A.2 Data Details

Details of the Aib_9 molecular dynamics simulation are provided in Table 1. The Gaussian mixture model data was generated via `torch.distributions`. This package enables sampling from Gaussian distributions of arbitrary dimensionality with given mean vectors and covariance matrices. In our usage, covariance matrices are the identity matrix, and mean vectors are drawn randomly. To generate multimodal Gaussian mixtures, we draw samples from independent Gaussian distributions and combine them. The ratio of samples from one distribution versus another can be varied, allowing a controllable degree of asymmetry for our free energy difference estimation experiment.

All code to perform data generation (GMM and Aib_9 datasets) is available at `https://github.com/tiwarylab/model-comparison`, and the datasets are additionally located on Zenodo.

Parameter	Value
Simulation engine	OpenMM
Temperature	450 K
Water model	TIP3P
Integration step	2 fs
Energy minimization	True
NVT equilibration	1 ns
NPT equilibration	1 ns
Production run	200 ns

Table A.1: Aib_9 MD parameters

Appendix B: **Supplementary Material for Chapter 3**

B.1 MD simulations setup

B.1.1 AIB_9

The simulation of AIB_9 was setup by following a previous study [124]. The PDB file was taken from the authors with permission, and the simulations are done with the CHARMM36m all atom force field [121] using TIP3P water molecules [120], a Parrinello-Rahman barostat [24], and a Nose-Hoover thermostat [21, 22] under the NPT ensemble. Simulations were performed using GROMACS 2016. [274] The structures of AIB_9 were saved every 0.2 ps and the dihedral angles were calculated using PLUMED 2.4 [275].

B.1.2 GACC

The simulations of GACC were done with the AMBER ff12 all atom force field, using TIP3P water molecules [120], a Parrinello-Rahman barostat [24], and a Bussi-Parrinello velocity rescaling thermostat [23] under the NPT ensemble. The simulations were performed using GROMACS 2016 [274]. The AMBER force filed was chosen because it exhibits more conformational variability compared with CHARMM force filed in a previous study [276]. The structures of GACC were also saved every 0.2 ps and the relevant dihedral angles were calculated using PLUMED 2.4 [275].

The PDB file of the GACC structure that served as the starting point for our simulation was generated using PyMOL. The initial structure was assumed to be at a temperature of 10K, and the systems energy was minimized with positional restraints of 25 kcal mol^{-1} $Å^{-2}$ in a two step process

- first the steepest descent algorithm was applied for 1000 steps followed by the conjugate gradient algorithm for another 1000 steps. Next, the system was equilibrated to 150K over 100ps under the NVT ensemble with positional restraints of 25 kcal mol^{-1} Å^{-2}. Then, the GACC was equilibrated from 150K to 277K at 1 atm under the NPT ensemble for 100 ps with positional restraints of 5 kcal mol^{-1} Å^{-2}. Finally, a long 5 ns equilibration was performed over 5ns at 298K under the NPT ensemble with positional restraints of 0.5 kcal mol^{-1} Å^{-2}, after which the system was copied to 48 replicas, and each replica was equilibrated to its target temperature under the NPT ensemble with positional restraints of 0.5 kcal mol^{-1} Å^{-2}.

The replica temperatures were geometrically spaced temperatures ranging from 277K to 408K, with temperature increased by 1% for each replica. The attempt of exchanging configurations was made every 10 ps, which is determined by checking the time correlation function of the potential energy.

B.2 Network structure and hyperparameters

The structure of the network is shown in Fig. 3.1. It has a U-net structure [117], where the input is down-sampled by 4 residue blocks and then up-sampled by another 4 residue blocks. The diffusion process is divided into 1000 discrete steps. The network parameters are optimized by the Adam algorithm [277] with a learning rate equal to 2^{-5}. Exponential moving average (EMA) with a decrease rate 0.995 is used to stabilize the stochastic gradient descent.

Appendix C: **Supplementary Material for Chapter 4**

C.1 Network Architecture and Parameters

Our thermodynamic map is parameterized as a score-based model using a U-net [117] with four Resnet downsampling blocks, followed by an attention layer, and four Resnet upsampling blocks. For all experiments, we use a learning rate of 10^{-3} with the Adam optimizer [277]. We also employ self-conditioning to accelerate training convergence [278]. The thermodynamic map is trained for 250 epochs with 100 diffusion timesteps for all studies performed.

C.2 Training and Sampling

Thermodynamic maps are trained as a score-based diffusion model where the objective is to estimate the score $\mathbf{s}(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ so that the forward diffusion may be reversed. Examples of the score are generated by randomly selecting samples $\mathbf{u}_0 = (\mathbf{x}_0, \boldsymbol{\beta}_0)^\top \in \mathcal{D}$ that are noised by evaluating the forward diffusion up to a timestep $t \sim \mathcal{U}(0, 1)$. In practice, the continuous-time forward diffusion process is approximated by a discrete Markov chain with transitions

$$\mathbf{u}_i = \sqrt{1 - \lambda_i}\mathbf{u}_{i-1} + \sqrt{\lambda_i}\mathbf{z}, \tag{C.1}$$

where $\mathbf{z}$ is sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{\beta}_0^{-1})$ for the dimensions of $\mathbf{u}$ corresponding to $\mathbf{x}$, and $\mathcal{N}(\mathbf{0}, \mathbf{1})$ for the dimensions corresponding to $\boldsymbol{\beta}$. The continuous noise schedule $\lambda(t)$ is discretized as $\{\lambda_i\}_{i=1}^N$, where N is the number of diffusion timesteps.

The Markov chain defined by Eq. C.1 linearly interpolates between data $\mathbf{u}_0$ and noise $\mathbf{z}$. As

such, Eq. C.1 can be evaluated up to any $\mathbf{u}_i$ starting from $\mathbf{u}_0$ in a single step by

$$\mathbf{u}_i = \sqrt{1-\alpha_i}\mathbf{u}_0 + \sqrt{\alpha_i}\mathbf{z}, \tag{C.2}$$

where α is a cumulative noise schedule defined as

$$\alpha(t) = e^{-\int_0^t \lambda(s)ds} \tag{C.3}$$

and has been discretized the same as $\lambda(t)$. In practice, we specify $\alpha(t)$ as a monotonically decreasing function

$$\alpha(t) = (1-2a)\cdot(1-t^2) + a \text{ where } a = 10^{-5}, \tag{C.4}$$

which is the same parameterization as in [251].

The U-net backbone of the model, denoted as f_θ, takes noisy samples $\mathbf{u}_i$ as input, and is trained to return denoised samples by minimizing the loss

$$\mathcal{L} = \mathbb{E}_{\mathbf{u}_0\in\mathcal{D}}\mathbb{E}_{t\sim\mathcal{U}(0,1)}\left[||\mathbf{u}_0 - f_\theta\left(\mathbf{u}_i, t\right)||^2\right]. \tag{C.5}$$

Eq. C.5 is equivalent to score matching and has the form of the simple loss defined in [279]. Predicting $\mathbf{u}_0$ from $\mathbf{u}_i$ enables one to estimate the noise added using Eq. C.3, which in turn is equivalent to the score up to a multiplicative factor determined by the cumulative noise schedule [280].

The observations $f_\theta(\mathbf{u}_i, t)$ from the denoising network can be used to evaluate the reverse diffusion process via the DDIM [281] update equation

$$\mathbf{u}_{i-1} = \sqrt{\alpha_{i-1}}f_\theta(\mathbf{u}_i, t) + \sqrt{1-\alpha_{i-1}-\eta_i^2}\left(\frac{\mathbf{u}_i - \sqrt{1-\alpha_i}f_\theta(\mathbf{u}_i, t)}{\sqrt{\alpha_i}}\right) + \eta_i^2\mathbf{z}. \tag{C.6}$$

DDIM treats the variance of the noise in the reverse diffusion, η_i^2, as a hyperparameter. When $\eta_i = \sqrt{(1-\alpha_i)/(1-\alpha_i)}\sqrt{1-\alpha_i/\alpha_{i-1}}$ the generative process approximates the reverse diffusion

SDE. When $\eta_i = 0$, the generative process approximates the probability flow ODE [282]. The generative process is evaluated with $\eta_i = 0$ for all studies performed.

C.3 Ising Model

The Ising Model presented in the main text is simulated on a 32×32 square lattice with periodic boundary conditions. Lattice configurations $\mathbf{x}_0 \in \mathbb{R}^{32\times32}$ are recorded every 1000 lattice sweeps and are uniformly dequantized prior to training the network. The Monte-Carlo simulation scheme ensures that samples at each temperature are always drawn from the equilibrium distribution. Therefore, the entries of $\boldsymbol{\beta}_0 \in \mathbb{R}^{32\times32}$ are identically β_{bath}. The outputs of the sampling procedure are binarized to ±1 based on $\text{sgn}(\mathbf{x})$.

C.4 G-vectors

G-vectors are an internal coordinate system that represents RNA structures as an orientational, equivariant contact map between all pairs of nucleobases [173]. The G-vector between nucleobases i and j is computed from the displacement vectors between the centers of the six-membered rings of the bases $\mathbf{r}_{ij} \in \mathbb{R}^3$ as

$$\mathbf{G}_{ij}(\mathbf{r}_{ij}) = \begin{pmatrix} \sin(\gamma|\mathbf{r}_{ij}|)\frac{r_{ij,x}}{|\mathbf{r}_{ij}|} \\ \sin(\gamma|\mathbf{r}_{ij}|)\frac{r_{ij,y}}{|\mathbf{r}_{ij}|} \\ \sin(\gamma|\mathbf{r}_{ij}|)\frac{r_{ij,z}}{|\mathbf{r}_{ij}|} \\ 1 + \cos(\gamma|\mathbf{r}_{ij}|) \end{pmatrix} \times \frac{\Theta(|\mathbf{r}_{ij}| - r_c)}{\gamma}, \tag{C.7}$$

where $|\mathbf{r}_{ij}|$ is the distance between bases i and j, Θ is the Heaviside step function, r_c is a cutoff of 2.4 Å, and $\gamma = \pi/r_c$. The internal representation of an RNA with N bases then takes the form of a $N \times N \times 4$ dimensional tensor, making the G-vector representation convenient for use with convolutional networks.

C.5 Ensemble Weighted Observables

Consistent with an ensemble perspective of biomolecular structure, we compute ensemble-weighted fraction folded curves for the RNA systems studied in the main text. In principle, any observable $A(\mathbf{x})$ computed from microscopic coordinates $\mathbf{x}$ can be weighted by the conformational ensemble at temperature β^{-1}, denoted as $p(\mathbf{x}, \beta)$, through

$$\langle A(\mathbf{x}, \beta) \rangle_p = \langle A(\mathbf{x}) | p(\mathbf{x}, \beta) \rangle, \tag{C.8}$$

where $\langle \cdot | \cdot \rangle$ denotes the inner product. Estimating $p(\mathbf{x}, \beta)$ in Eq. C.8 through binning is usually intractable for high-dimensional $\mathbf{x}$. For this reason, it is often more effective to describe the system in terms of some lower-dimensional collective variables $\xi = f(\mathbf{x})$. For example, in the GCAA tetraloop and HIV-TAR RNA experiments, the ensemble-weighted number of base pairs is computed along the first two principal components of the G-vectors for RNA structures, with measurements of $A(\xi)$ gathered from MD simulations, and $p(\xi, \beta)$ generated by the thermodynamic map.

C.6 Rosetta, MD Simulations, and Adaptive Sampling

Initially, 10^4 structures are generated through MC sampling on the Rosetta potential [169]. G-vectors are computed for all generated structures, and candidate structures are sampled via regular space clustering on the first two principal components of the G-vectors. Before clustering, we compute the distribution energies of the generated structures in vacuum using the MD force field, and we discard any structures with energy greater than the mean plus two standard deviations.

These initial structures are then simulated using the OpenMM simulation engine [283] with the D.E. Shaw RNA force field [156] at 1M KCl concentration and 1 atm with the TIP-4P water model [284]. Simulation temperatures are uniformly spaced at 10K intervals between 310K and 450K, and are assigned to Rosetta structures according to increasing energy.

Structures are equilibrated for 10ns before the first production simulation of 50ns. Between

rounds of adaptive sampling, structures are equilibrated for 1ns. Simulation structures are saved at 1ns intervals during the production simulation, and these structures are represented as G-vectors to form the training data for the thermodynamic map.

Beyond the initial round of simulation beginning from Rosetta structures, simulations are launched from new structures which are adaptively sampled via regular space clustering on 10^5 samples generated by the thermodynamic map for each temperature on the simulation ladder (uniformly spaced at 10K intervals between 310K and 450K). Only samples with free energy less than 2kcal/mol relative to the most abundant conformation are considered for adaptive sampling at each temperature.

C.7 Local Equilibrium and Linear Fluctuations

C.7.1 Motivation

Generating statistically independent samples from the global equilibrium distribution of a complex system is computationally expensive, and for sufficiently complex systems may be completely infeasible. An alternative way to explore the energy landscape is to run many simulations in parallel, each of which exists in a local equilibrium that differs from the global equilibrium distribution by a constant shift in the free energy. One must be careful in concluding the parallel simulations since the aggregate of the simulations in local equilibrium may not represent the global equilibrium. For example, if the initial conditions of the parallel simulations are not sampled from the global equilibrium distribution, then the statistics of the obtained data will likely not reflect equilibrium.

Methods such as Replica Exchange Molecular Dynamics (REMD) ensure that simulations remain in equilibrium throughout by enforcing exchanges between replicas to follow the detailed balance criteria, but suffer from poor scalability with system size [65]. Maintaining frequent exchanges while enforcing the energy-based detailed balance condition requires replicas to be more closely spaced on the temperature ladder as system size increases to maintain computational

efficiency. For this reason, it is computationally infeasible to study large systems through replica exchange.

We develop an alternative method of inferring equilibrium from TM-aMD that assumes the scale of fluctuations should increase linearly with temperature. The temperature of a system can be defined in terms of coordinate fluctuations, but correlations between the fluctuating degrees of freedom and non-linear dependence of fluctuations with temperature make relating fluctuations to temperatures in this manner infeasible [285]. However, for the simple harmonic prior, each coordinate of the simple system fluctuates independently of the others, and the relationship between coordinate fluctuations and temperature is independently linear for each degree of freedom. Based on this observation, we define the inverse-temperature variables $\boldsymbol{\beta}$ in the thermodynamic map as the simulation's fluctuations along each degree of freedom. The fluctuations $\boldsymbol{\beta}$ in turn define the temperature of the prior system, where a linear relationship holds between $\boldsymbol{\beta}$ and the thermodynamic temperature T.

In other words, we rely on the ability of the thermodynamic map to transform a complex system whose fluctuations along all degrees of freedom are highly correlated with one another into a simpler, correlation-free system where the correspondence between fluctuations and temperature is tractable. We emphasize that this protocol is not needed when sampling from the global equilibrium distribution. For example, in the Ising model experiment, each entry of $\boldsymbol{\beta}$ may be set to β_{bath} since Monte Carlo sampling guarantees global equilibrium. Below we detail the linear fluctuation protocol.

C.7.2 Protocol

Suppose that N simulations are performed independently and over a short timescale such that large-scale conformational transitions do not occur. In this case, each replica explores a single free energy minimum, generating a local equilibrium distribution of conformations $\{p_n(\mathbf{x})\}_{n=1}^N$ where $\mathbf{x} \in \mathbb{R}^M$. The fluctuations of each replica are quantified by $\beta_{nm}^{-1} = \text{Variance}(p_n(x_m))$. That is, the fluctuations are computed coordinate-wise and separately for each replica, leading to N

inverse-temperature variables $\boldsymbol{\beta}_n^{-1} \in \mathbb{R}^M$.

During the training of the thermodynamic map, samples $\mathbf{x} \sim p_n(\mathbf{x})$ are paired with their corresponding $\boldsymbol{\beta}_n^{-1}$, so that samples from $p_n(x_m)$ are mapped to a harmonic oscillator (the prior system) where each degree of freedom is at temperature β_{nm}^{-1}.

To generate samples at a given T, we fit a linear model, f_m, to each $\{(T_n, \beta_{nm}^{-1})\}_{n=1}^N$ resulting in a multilinear function $\mathbf{f} = (f_1, f_2, \ldots, f_M)$ that relates temperatures to fluctuations in a coordinate-wise fashion. Generating samples at temperature T amounts to sampling from the prior at (statistical) temperature $\boldsymbol{\beta}^{-1} = \mathbf{f}(T)$.

In the main text, we assume linear fluctuations to infer the global equilibrium for the GCAA tetraloop and HIV-TAR RNAs. The coordinates of a N nucleobase RNA are represented as a vector $\mathbf{G} \in \mathbb{R}^{N \times N \times 4}$ computed from the pairwise displacement vectors between bases $\mathbf{r} \in \mathbb{R}^{N \times N \times 3}$ (see Appendix C.4). The fluctuation variables $\boldsymbol{\beta}$ are computed directly from $\mathbf{r}$ and averaged over the spatial dimensions to obtain $\boldsymbol{\beta} \in \mathbb{R}^{N \times N}$, with a single β_{ij} shared between all coordinates of $\mathbf{G}_{ij} \in \mathbb{R}^4$.

Initially, we expect the assumed linear relationship between fluctuations and temperatures will be inaccurate. However, after many iterations of TM-aMD with Fluctuation Mapping, we expect that the linear dependence will gradually be imposed on the simulated system. We verify this intuition in Fig. C.1 by characterizing the accuracy of the multilinear fit $\mathbf{f}$ for each $\boldsymbol{\beta}$ for the HIV-TAR RNA. Fig. C.1 shows the Pearson r value for each $\beta_{ij}^{-1} = f_{ij}(T)$ across each iteration of the TM-aMD algorithm. Initially, the linear assumption is inaccurate, as evidenced by the large number of fits with small r values. However, as the algorithm progresses, the quality of fit improves indicating that the linear relationship between fluctuations and temperature in the prior has been transferred to the RNA system.

Algorithm 1 TM-aMD Protocol

Require: Seed structures $\{\mathbf{x}_i^{(0)}\}_{i=1}^N$, temperatures $\{T_k\}_{k=1}^K$, $j = 0$.
$\{\mathbf{x}_i^{(0)}\}_{i=1}^N$ *are seeds for TM-aMD iteration* 0.
Assign each structure $\mathbf{x}_i^{(0)}$ a $T \in \{T_k\}_{k=1}^K$ to obtain $\{(\mathbf{x}_i^{(0)}, T_i)\}_{i=1}^N$
while not converged **do**
  See Appendix C.8.
  for each simulation seed $(\mathbf{x}^{(j)}, T)$ **do**
    Simulate $\mathbf{x}^{(j)}$ at temperature T and record M samples $\{\mathbf{x}_m\}_{m=1}^M$
    if linear fluctuations **then**
      $\boldsymbol{\beta}^{-1}$ = Variance$(\{\mathbf{x}_m\}_{m=1}^M)$ (e.g. MD simulations)
      Compute coordinate-wise per replica (see Appendix C.7).
    else
      $\boldsymbol{\beta}^{-1} = T$ (e.g. Monte-Carlo)
    end if
    Pair each $\mathbf{x} \in \{\mathbf{x}_m\}_{m=1}^M$ with $\boldsymbol{\beta}^{-1}$ to form state vectors $\{(\mathbf{x}, \boldsymbol{\beta})_m^\top\}_{m=1}^M$
  end for
  Form dataset consisting of all simulation state vectors $\{(\mathbf{x}, \boldsymbol{\beta})_l^\top\}_{l=1}^{MN}$
  Multilinear fit $\mathbf{f}$ to $\{(T_i, \boldsymbol{\beta}_i^{-1})\}_{i=1}^N$
  See Appendix C.7.
  Train thermodynamic map using $\{(\mathbf{x}, \boldsymbol{\beta})_l^\top\}_{l=1}^{MN}$
  See Appendix C.2.
  for each temperature T in $\{T_k\}_{k=1}^K$ **do**
    Generate samples $\{\hat{\mathbf{x}}\}$ from the thermodynamic map with prior $\boldsymbol{\beta}^{-1} = \mathbf{f}(T)$.
    See Appendix C.2.
    Project samples $\{\hat{\mathbf{x}}\}$ into low-dimensional space $\{\hat{\xi}\}$ ($\hat{\xi}_i = g(\hat{\mathbf{x}}_i)$).
    For example, g is PCA.
    Obtain new seeds $\{(\mathbf{x}_i^{(j+1)}, T)\}_{i=1}^{N_k}$ by clustering on $\{\hat{\xi}\}$
    N_k *replicas for each* T_k.
  end for
  $N \leftarrow \sum_k N_k$
  Obtain $\{(\mathbf{x}_i^{(j+1)}, T_i)\}_{i=1}^N$ consisting of all seeds
  N *replicas for TM-aMD iteration* $j + 1$.
  $j = j + 1$
end while

C.8 Convergence to Equilibrium

A central unanswered question is how to determine convergence to the equilibrium distribution for TM-aMD. Other methods, such as those based on Markov state models, rely on enforcing the detailed balance condition for transitions between states [286]. However, the samples produced by

thermodynamic maps are statistically independent, and checking detailed balance in this manner requires correlated samples. Rather than checking the detailed balance conditions, we argue that if the thermodynamic map has learned the equilibrium distribution, then the MD and TM-generated distributions should be in agreement and should not change between iterations. As stated in the main text, this condition is necessary, although not completely sufficient to identify equilibrium. For simplicity, instead of monitoring agreement between the MD and TM distributions, we monitor agreement between MD and TM predictions of observables.

More concretely, for the RNA systems studied we monitor agreement between melting temperature predictions, as shown in Fig. C.2. For the GCAA tetraloop (Fig. C.2A), we observe disagreement between MD and TM-predicted melting temperatures until $\sim 45\mu$s of MD simulation, after which the melting temperatures show agreement and are unchanging between iterations. The final three iterations satisfy our heuristic equilibrium condition and are used to compute the melting temperature reported in the main text. On the other hand, the melting temperature predictions for HIV-TAR do not show the same convergence. Fig. C.2B shows that for four of the six iterations, the MD and TM-predicted melting temperatures are in agreement, but are clearly not converged. Nonetheless, we use the four iterations showing agreement (3,5,6,7) to compute the melting temperature reported in the main text.

Another striking feature of the convergence plots (Fig. C.2) is the instability of melting temperature predictions for simpler GCAA tetraloop, which occur over a range of $500°$C. Examining the TM-generated conformational landscape at each iteration explains these highly variable predictions. Fig. C.3 shows the TM-predicted conformational landscape at each iteration of Fig. C.2 for the GCAA tetraloop. The iterations with extreme melting temperature predictions (4-6) correspond to the discovery of the unfolded state. All of the Rosetta-generated structures correspond to the folded state, leading to gradual exploration of the unfolded state over the first three iterations. Iterations 4-6 learn the relative populations of the folded and unfolded states across temperature, while the final three iterations show convergence in terms of the relative populations of the folded and unfolded states.

In contrast to the GCAA tetraloop, where Rosetta only predicts the most stable structure, the Rosetta predictions for HIV-TAR constitute an ensemble (Fig. C.4). While TM-aMD explores new states not present in the initial ensemble for the GCAA tetraloop, HIV-TAR requires less exploration. Overall, the findings outlined in our convergence analysis show that the quality of the initial ensemble has a dramatic impact on the convergence of the TM-aMD algorithm.

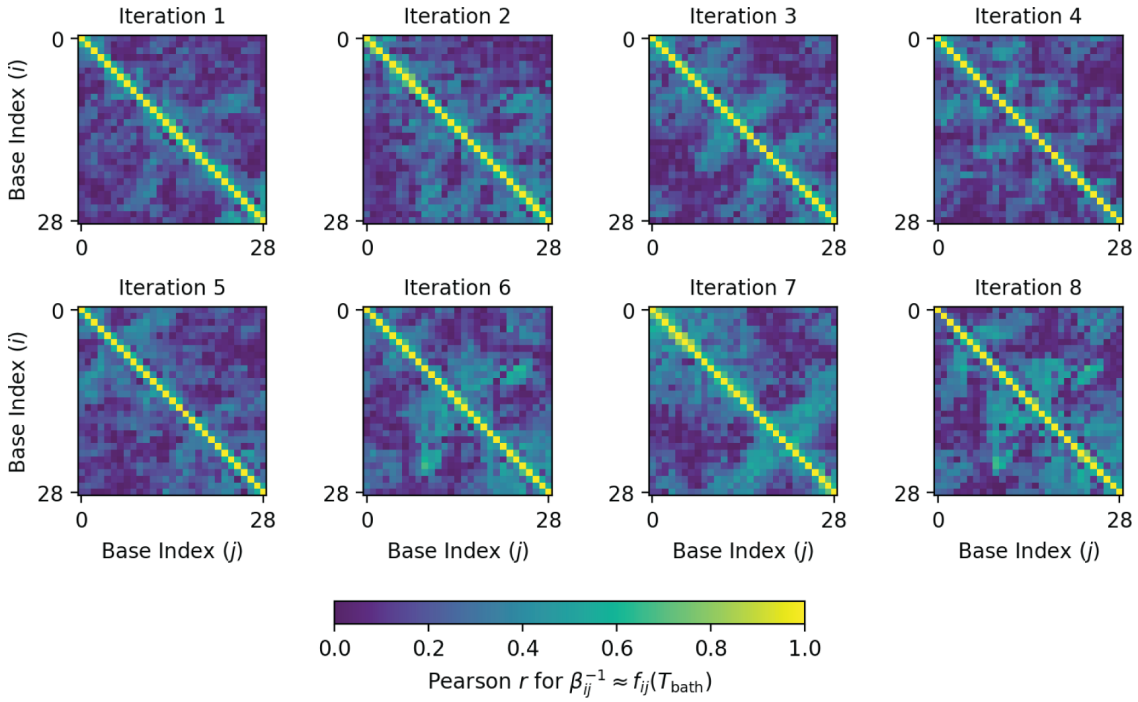


Figure C.1: **Quality of linear fluctuation fit for the HIV-TAR RNA.** For each iteration of TM-aMD applied to the HIV-TAR RNA in Section 4.6 the accuracy of the Fluctuation Mapping method is assessed. Since the dependence of fluctuations with temperature is linear in the prior system we expect that as TM-aMD progresses a similar dependence should be observed in the simulated RNA system. For each iteration of TM-aMD we compute fluctuations from the pairwise displacement vectors $\mathbf{r}$ as $\{\beta_{ij}^{(n)}\}_{n=1}^{N} = \{\langle \text{Variance}(\mathbf{r}_{ijk}^{(n)})\rangle_k\}_{n=1}^{N}$, where i and j are nucleobase indices, k denotes the spatial coordinates, and n denotes the replica index. We then evaluate the multilinear fit $\mathbf{f}(T)$, where each f_{ij} is a linear fit to $\{(T^{(n)}, \beta_{ij}^{(n)^{-1}})\}_{n=1}^{N}$. The quality of the fit, and therefore the validity of the Fluctuation Mapping assumption, is characterized by the Pearson r values for each f_{ij}, as depicted in the figure. Initially, the assumption is inaccurate, as evidenced by the number of entries with small r values. However, as more iterations of TM-aMD are performed the quality of fit improves, indicating the thermodynamics of the replica simulations more closely matches that of the prior, validating our use of Fluctuation Mapping.

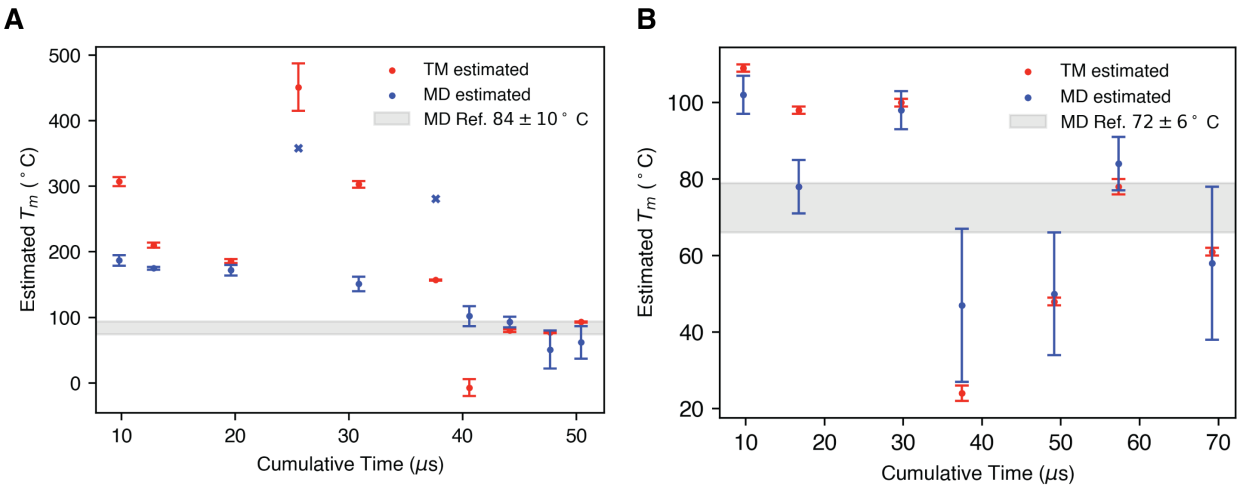


Figure C.2: **Convergence of Melting Temperature Calculations. A** Melting temperatures for the GCAA tetraloop computed independently for each iteration of TM-aMD with one standard deviation uncertainties on the prediction. Convergence occurs after 45μs of MD simulation. For visual clarity, predictions with uncertainties greater than 100°C are represented as x's. **B** Melting temperatures for HIV-TAR are computed independently for each iteration of TM-aMD with one standard deviation uncertainties on the prediction. In both panels, the TM predictions show smaller uncertainties relative to the MD predictions.

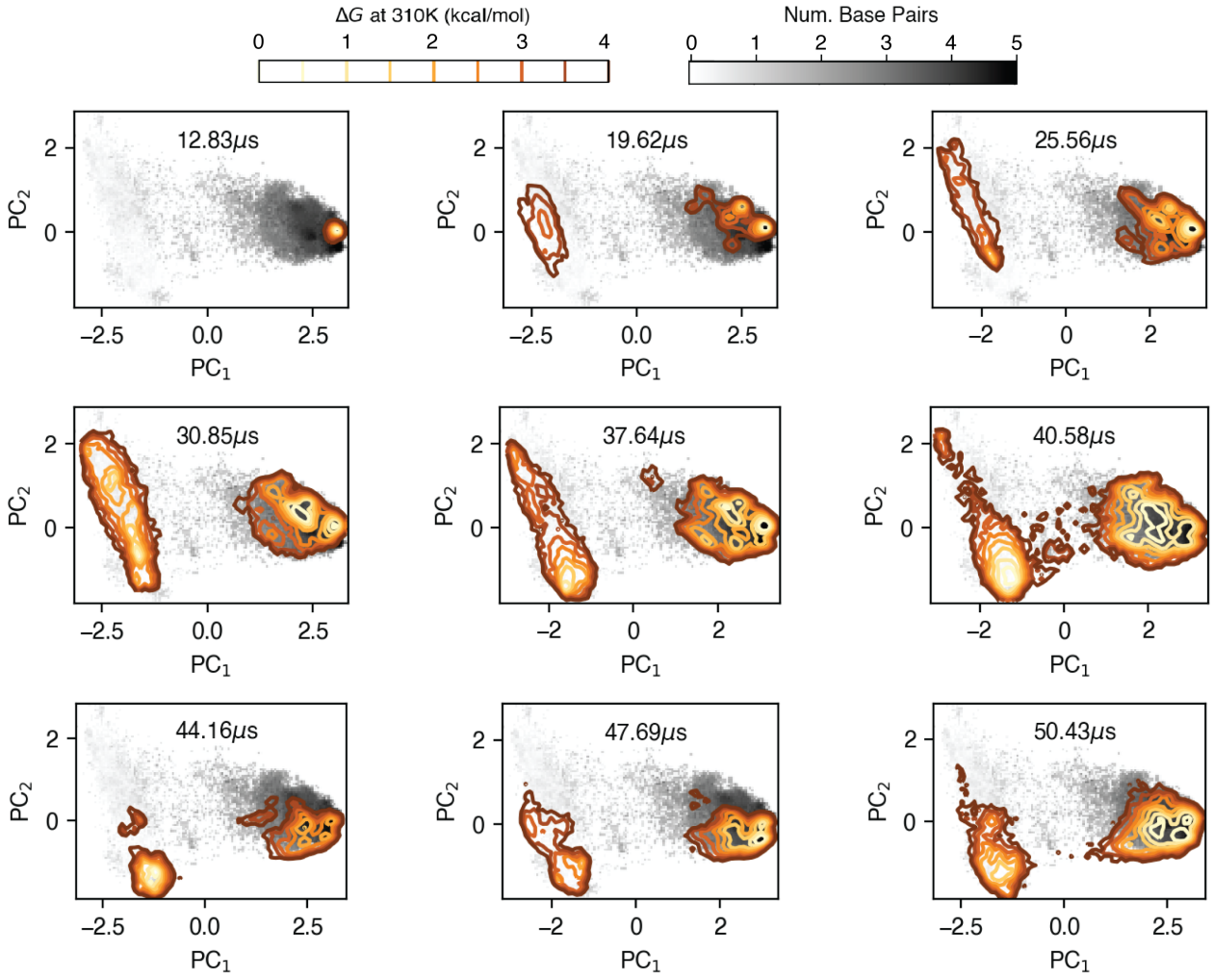


Figure C.3: **Convergence Towards Equilibrium for GCAA.** The TM-predicted conformational landscape at each iteration is superimposed on the distribution of all MD structures, shaded by the number of base pairs. The cumulative simulation time associated with iteration is listed.

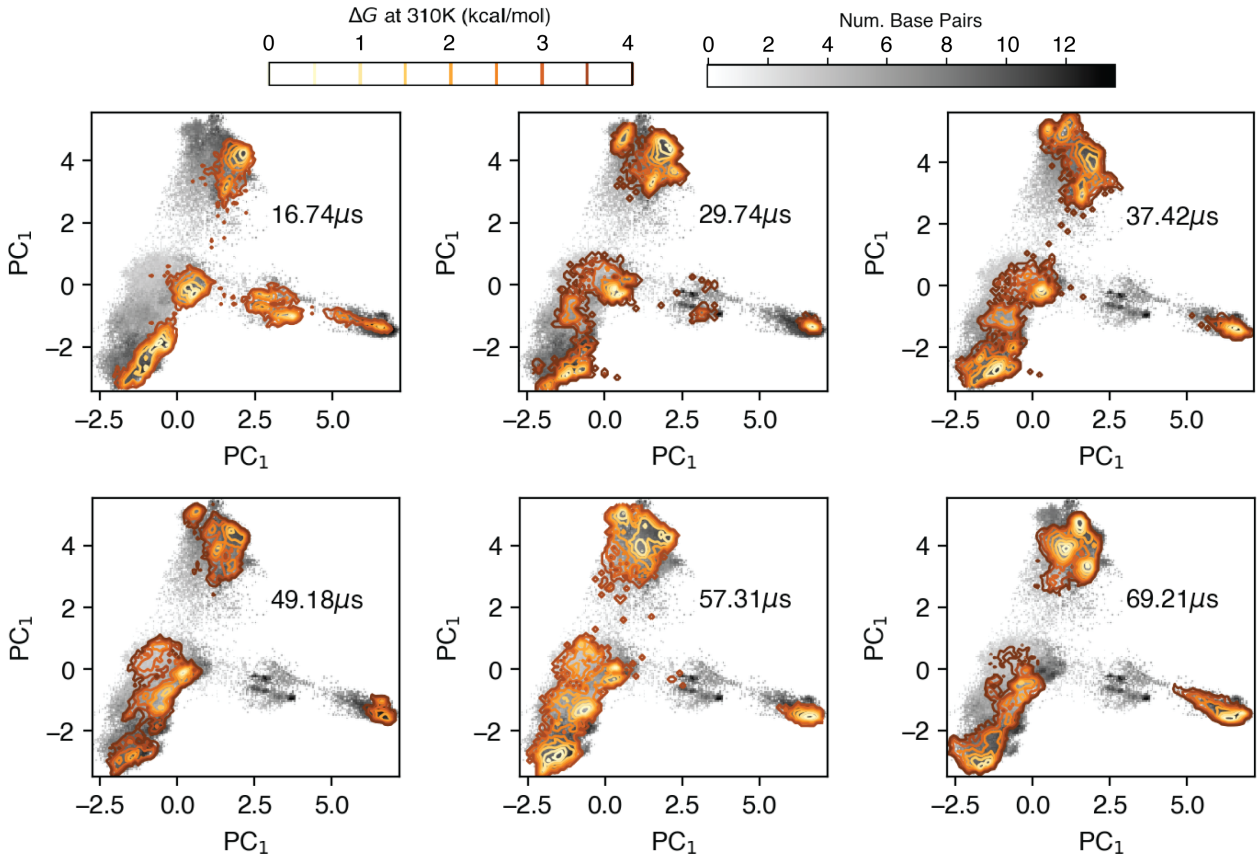


Figure C.4: **Convergence Towards Equilibrium for HIV-TAR.** The TM-predicted conformational landscape at each iteration is superimposed on the distribution of all MD structures, shaded by the number of base pairs. The cumulative simulation time associated with iteration is listed.

Appendix D: **Supplementary Material for Chapter 5**

D.1 Secondary Structure Prediction

The secondary structure of an RNA is a simplified representation that specifies Watson-Crick base pairing. There are numerous, long-standing approaches available to predict secondary structures from the primary sequence. [287] RNAnneal relies on three programs to predict RNA secondary structures: EternaFold [195], LinearFold [196], and LinearAliFold [196]. EternaFold is a probabilistic method for secondary structure prediction that has been benchmarked against several widely used secondary structure prediction algorithms (Algorithm 2). LinearFold is an algorithm for predicting secondary structures that has linear runtime with sequence length (Algorithm 3). LinearAliFold extends the functionality of LinearFold by implementing stochastic sampling (Algorithm 4). After a large number of secondary structures have been predicted by the above methods, the structures are aggregated and redundant ones are filtered out. Fig. SD.3 measures the fraction of correctly predicted bases for the predicted distributions of secondary structures for each riboswitch.

D.2 FARFAR2

Rosetta's Fragment Assembly of RNA with Full-Atom Refinement (FARFAR) is a hybrid knowledge- and physics-based approach to sampling models of 3D RNA structure. [288] FARFAR samples 3D models by performing Monte Carlo sampling on a low-resolution potential using a library of experimentally validated structure fragments, and then relaxes the structure in a higher-resolution all-atom potential. The quality of the conformations is evaluated by the Rosetta score.

FARFAR2 further improves the performance of FARFAR with updates to the sampling algorithms, scoring weights, and fragment libraries. [198] The Rosetta score referred to in the main text and Fig. 5.2g-h corresponds to the `rna_hires.wts` file in Rosetta v2021.16.61629. Algorithm 5 describes how RNAnneal uses FARFAR2 to sample a candidate model of the 3D structure with the primary sequence and a secondary structure as input.

D.3 Clustering

A crucial step in RNAnneal is selecting representative conformations to simulate (see Molecular Dynamics) from the set of structures predicted by FARFAR2. To sample diverse conformations, we first compute G-vectors for the FARFAR2 conformations (see R- and G-vectors), and then project the G-vectors onto $d = 2, 3, \ldots, 26$ principal components. For each d value, the data is clustered using the advanced density peaks algorithm [289] implemented in `dadapy` [290] with $Z = 3$. Representatives are selected based on the the d value that yields the maximum number of clusters (Fig. SD.4). To select, e.g., 200 candidate structures, we first sort the conformations within each cluster based on their Rosetta score and then sort the clusters based on the top-scoring conformation in each cluster. The clusters are then traversed in order, and the top-scoring conformation is selected as a representative and subsequently removed from the cluster. If there are fewer than 200 candidate structures after all clusters have been traversed, then they are reordered based on the new top-scoring conformation and the selection procedure is repeated until 200 conformations are obtained. The entire procedure is detailed in Algorithm 7.

D.4 Molecular Dynamics

RNAnneal simulates models of 3D RNA structure with implicit solvent MD in order to introduce variations in structure caused by the thermal environment. The simulations are conducted under the NVT ensemble using the OpenMM simulation engine and the DES-AMBER RNA force field. [156, 291] The conformations are equilibrated by gradually increasing the temperature from

1K to 300K and then increasing the integration timestep from 1fs to the production integration timestep of 5fs (Table D.2). The production simulation is 100ns and frames are saved at 1ns intervals; additional simulation parameters like ion concentration and solvent model are provided in Table D.3. The DES-AMBER score is the potential energy of the MD simulations (i.e., the energy values of the force field and solvent model) standardized by first aggregating the energy across all simulations (for a given RNA) and then subtracting the minimum value and dividing the standard deviation.

D.5 R- and G-vectors

RNAnneal featurizes conformations with two nucleobase-centric representations constructed from R- and G-vectors. [173] An R-vector is a displacement vector $\mathbf{r}_{ij}$ directed between the six-membered aromatic ring of nucleobases i and j. The R-representation of a conformation with N nucleobases is thus a tensor of shape $N \times N \times 3$. Similarly, the G-representation is a $N \times N \times 4$ tensor consisting of G-vectors that capture the geometry of localized base-pairing and stacking interactions. The G-vector, $\mathbf{G}_{ij}$, is computed as

$$\mathbf{G}_{ij}(\mathbf{r}_{ij}) = \begin{pmatrix} \sin(\gamma|\mathbf{r}_{ij}|)\frac{r_{ij,x}}{|\mathbf{r}_{ij}|} \\ \sin(\gamma|\mathbf{r}_{ij}|)\frac{r_{ij,y}}{|\mathbf{r}_{ij}|} \\ \sin(\gamma|\mathbf{r}_{ij}|)\frac{r_{ij,z}}{|\mathbf{r}_{ij}|} \\ 1+\cos(\gamma|\mathbf{r}_{ij}|) \end{pmatrix} \times \frac{\Theta(|\mathbf{r}_{ij}|-r_c)}{\gamma} \tag{D.1}$$

where Θ is a step function, r_c is a cutoff of 2.4Å, and $\gamma = \pi/r_c$.

D.6 SPIB

The State Predictive Information Bottleneck (SPIB) framework allows one to embed a high-dimensional dynamical system into a lower-dimensional latent space while preserving the slowest

degrees of freedom of the system. [199] SPIB takes the form of a time-lagged autoencoder, where high dimensional coordinates at time t are passed to the encoder and embedded in a latent space; the decoder must then predict the state label at time $t + \Delta t$ solely from the latent embedding. SPIB has two main hyperparameters: the time lag, Δt and a regularizer, β. The time lag corresponds to the lifetime of the predicted states, while β regularizes the latent space to follow a mixture of Gaussians prior. We train two sets of SPIB models on the first 8 principal components of the R- and G-vectors. For each feature type, we train seven models with the number of IB components, d, ranging from 2 to 8. For all models, we use a linear encoder, non-linear decoder, a 2ns lag time, and set $\beta = 0.05$. The SPIB implementation is provided by the `af2rave` python package. [292]

D.7 Thermodynamic Maps

A thermodynamic map (TM) is a diffusion model variant that models the dependence of a structural ensemble on physical parameters like temperature, pressure, or chemical potential by representing their statistical effect on the prior distribution. [66, 193, 194] Temperature, for example, is represented by tempering the prior (i.e., adjusting the variance) on a per-sample basis, while pressure and chemical potential correspond to exponential tilting operations that shift the mean of the prior. [66, 193] Once trained, the TM model can then be used to generate new samples or evaluate the likelihood of existing samples at a given set of thermodynamic conditions. In this work we modify the construction so that the prior is tempered based on the energy of the conformations. Concretely, for a given representation $\mathbf{x}_0$, we set the variance of the prior equal to the DES-AMBER score, U_0, of the conformation corresponding to $\mathbf{x}_0$ (see Molecular Dynamics; Fig. 5.1c). The RNAnneal Score for a given conformation is computed from the log-likelihoods of the latent representations under the trained models. The log-likelihood calculation is summarized in Algorithm 8 and described in detail in Ref. [293] Our TM implementation infers the energy-dependence of the conformational ensemble, and as a result one must choose a variance (i.e., a temperature) for the prior when computing the score. All the results in the main text employ a unit

variance prior, but other values sometimes lead to better performance (Fig. SD.14-SD.16).

D.8 RNAnneal Score

RNAnneal scores conformations using an ensemble of Thermodynamic Maps trained on geometric and energetic features extracted from a set of 3D conformations (Fig. 5.1b-d). A total of 28 TMs are trained for each RNA, covering all four combinations of representations (R- and G-vectors) and projections (PCA and SPIB). For each feature-projection combination, the number of projection components d varies from 2 to 8 inclusive. After training, the RNAnneal score is computed from the log-likelihoods of the training data under the TM ensemble (Algorithm 8 and 9).

D.9 Reference Structures

RNAnneal's structure prediction performance was evaluated on a set of riboswitch sequences with structures that have been solved by high-resolution experimental methods (Table D.1 and Fig. SD.1). The reference ERCs were prepared by downloading the associated Protein Data Bank (PDB) entry, removing any non-nucleic residues, filling in missing nucleobases using FARFAR2, and subsequently minimizing the conformation with the Rosetta energy function (see Algorithm 6).

D.10 RMSD

Root Mean Squared Deviation (RMSD) is a metric commonly used to evaluate discrepancies between structure models. We evaluate the heavy-atom RMSD of each model after alignment to the native structure. The RMSD between a model $\mathbf{x}$ and reference $\mathbf{y}$ conformation containing N heavy atoms is defined as

$$\mathrm{RMSD}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d(\mathbf{x}_i, \mathbf{y}_i)^2}, \tag{D.2}$$

where $d(\mathbf{x}_i, \mathbf{y}_i)$ measures the distance between the i-th atom in both models. The RMSD values for RNAnneal conformations and molecular dynamics simulations of the reference structure are reported in Fig. SD.6 and Fig. SD.5.

D.11 INF

Interaction Network Fidelity (INF) quantifies the similarity between two conformations based on their Leontis-Westhof base pairing and stacking classifications. [204, 294] Let S_m and S_r denote the sets of nucleobase interactions (base pairs and base stacks) in the model and reference conformations, respectively. The INF is defined as the Fowlkes-Mallows Index between these interaction sets:

$$\mathrm{INF} = \sqrt{\frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}|} \times \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FN}|}} \tag{D.3}$$

where $\mathrm{TP} = S_m \cap S_r$ are interactions present in both structures, $\mathrm{FP} = S_m \; S_r$ are interactions present only in the model, and $\mathrm{FN} = S_r \; S_m$ are interactions present only in the reference. Minor, localized structural differences (e.g., at a helix junction) may yield large RMSD values due to poor global alignment. Because the INF compares interaction networks directly without requiring alignment, it is more robust to such variations and to thermal fluctuations. We compute the INF using the ClaRNA [295] implementation in `rna-tools` [296]. INF values are reported in Fig. SD.2 and SD.8-D.9.

D.12 AUAC

The Area Under the Accumulation Curve (AUAC) is a metric suited for evaluating classification task performance (Fig. 5.1e and 5.2g). [212] Let $x \in [0, 1]$ be the normalized rank and $f : r \to \{0, 1\}$ be an associated set of labels, with 1 indicating the positive (near-ERC) class and 0

indicating the decoy class. The empirical CDF is then

$$A(r) = \int_0^r f(s)ds \tag{D.4}$$

and the area under the accumulation curve is

$$\mathrm{AUAC} = \int_0^1 A(r)dr. \tag{D.5}$$

When the AUAC is 1, the classification is perfect, i.e., members of the positive class are always ranked above decoys. The accumulation curves for the 15 riboswitches investigated in the main text are provided in Fig. SD.15 and SD.16.

D.13 BEDROC

The Boltzmann-Enhanced Discrimination of Receiver Operator Characteristic (BEDROC) modifies the Area Under the Accumulation Curve (AUAC) to focus on early recognition of the positive (near-ERC) class (Fig. 5.1e and 5.2g). [212] The AUAC is skewed towards late recognition since contributions from the lowest ranked samples dominate the area calculation in Eq. D.5. One may introduce an exponentially decreasing weight on the rank in the AUAC calculation to focus on early recognition, i.e.,

$$w\mathrm{AUAC} = \int_0^1 A(r)e^{-\alpha r}dr, \tag{D.6}$$

which is normalized to obtain

$$\mathrm{BEDROC} = \frac{w\mathrm{AUAC} - w\mathrm{AUAC}_{\mathrm{min}}}{w\mathrm{AUAC}_{\mathrm{max}} - w\mathrm{AUAC}_{\mathrm{min}}}. \tag{D.7}$$

The tunable parameter α controls the focus on early recognition. Our choice of $\alpha = 8$ is such that the top 20% of the normalized rank contributes 80% of the $w\mathrm{AUAC}$. Early recognition is a suitable evaluation metric when there is a large class imbalance and accurate classification among

the highest ranks is essential.

D.14 Interaction Entropy

Interaction entropy measures how heterogeneous each nucleobase's interaction patterns are across an ensemble of conformations. First, each nucleobase i is assigned a Leontis–Westhof (LW) base-base contact classification for every structure in the ensemble. [294] The distribution of these classifications defines probabilities $p_i(c_j)$ for each contact c with partner nucleotide j. The interaction entropy is then the Shannon entropy of this distribution,

$$H_i = -\sum_j \sum_{c \in \mathcal{C}} p_i(c_j)\, \log_2 p_i(c_j),$$

where $\mathcal{C}$ denotes the set of LW classes. Expressing the Shannon entropy in bits lends a direct interpretation: a nucleotide with interaction entropy H participates in 2^H distinct base-base interactions across the ensemble (Fig. 5.3b). The probabilities may be computed as unweighted frequencies or with prior weights associated to each structure. In this work we use state probabilities (see State Ranking) computed from the RNAnneal score as prior weights.

D.15 State Ranking

We briefly describe the procedure by which a conformational ensemble, i.e., a small number of representative conformations and their weights, is predicted by RNAnneal. We project the RNAnneal conformations onto the $d = 2$ information-bottleneck (IB) space learned by SPIB to obtain latent representations $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N\}$. We further assign the predicted state label, $y_i \in \{y_0, y_1, \ldots, y_M\}$ and RNAnneal Score, s_i, to each conformation. Next, we histogram the conformations on a $K \times L$ grid in the IB space (Fig. 5.2j), using the mean s in each bin (denoted by $\overline{s}_{kl}$) to define a Boltzmann-like weight $w_{kl} = \exp[-\overline{s}_{kl}]$. We assign each populated bin a state label, y_{kl}, based on the most frequent state label of the conformations in the bin, and define a total

weight, W_i, for state y_i by $W_i = \sum_{y_{kl}=y_i} w_{kl}$. Finally, normalizing the state weights yields state probabilities p_{yi}, and the frame with the minimal s_i within each state is taken as its representative conformation. The entire procedure is described in Algorithm 11. The states predicted in the main text (Fig. 5.3) and supplement (Fig. SD.17-SD.20) were predicted based on a $d = 2$ SPIB model trained on the first 2 PCA components of the G-vectors.

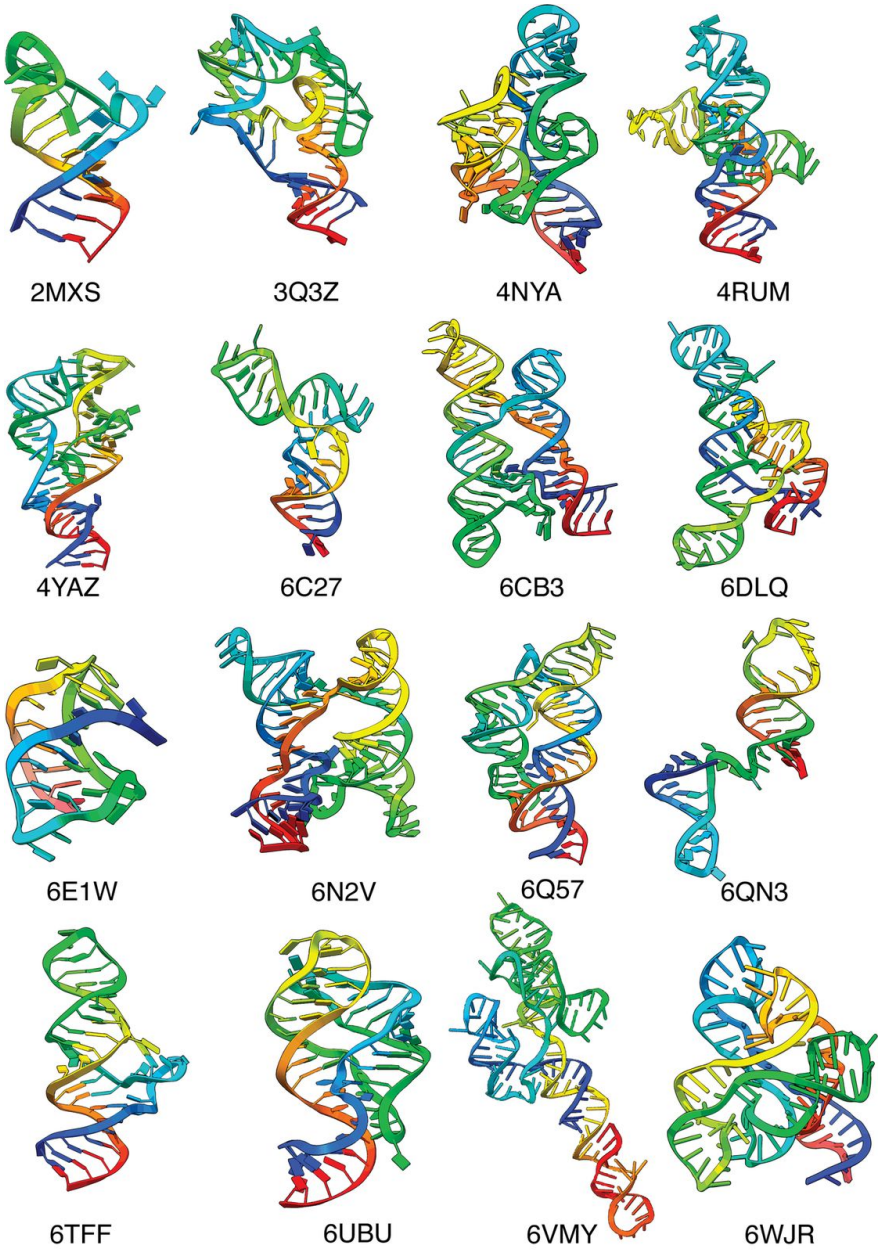


Figure D.1: **Riboswitch reference ERCs labeled with PDB accession code.**

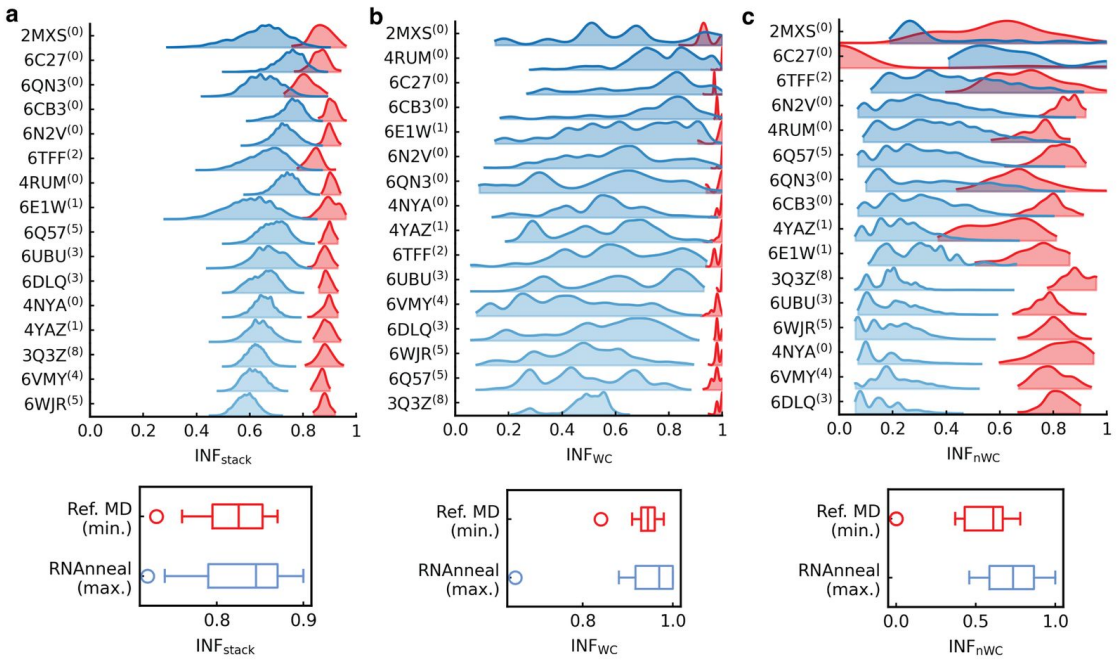


Figure D.2: **Interaction Network Fidelity (INF) by interaction type.** The INF is reported for **a**. stacking interactions (Median [IQR]; Ref. MD: 0.82 $[0.79 - 0.86]$; RNAnneal: 0.85 $[0.80 - 0.87]$), **b**. Watson-Crick (WC) interactions (Ref. MD: 0.95 $[0.93 - 0.96]$; RNAnneal: 0.98 $[0.92 - 1.0]$), **c**. non-Watson-Crick (nWC) interactions (Ref. MD: 0.62 $[0.48 - 0.67]$; RNAnneal: 0.80 $[0.59 - 0.87]$). It is worth noting that the conformation with, e.g., the maximum $\mathrm{INF}_{\mathrm{WC}}$, does not necessarily have the maximum $\mathrm{INF}_{\mathrm{stack}}$ or $\mathrm{INF}_{\mathrm{nWC}}$.

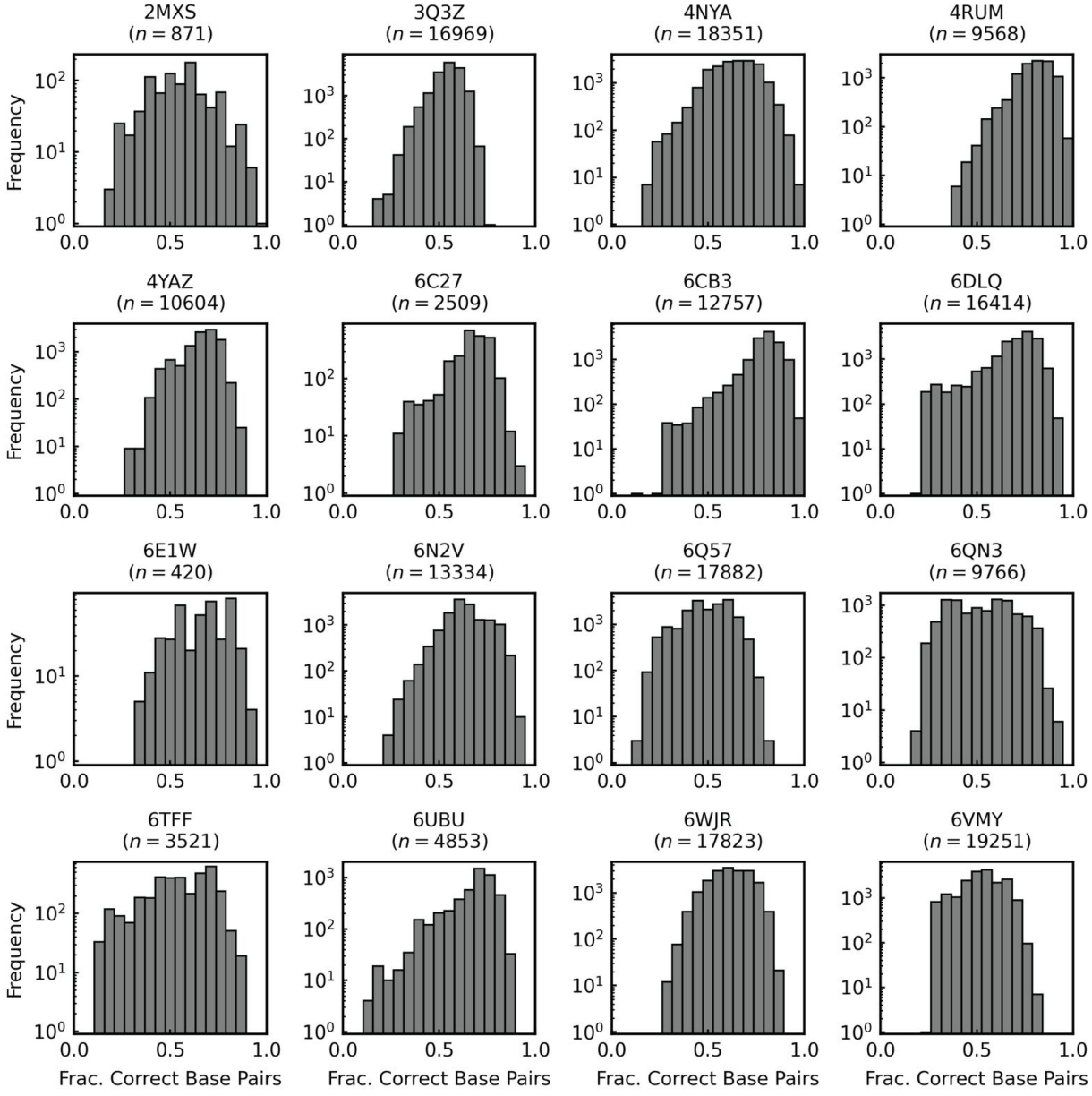


Figure D.3: **Evaluation of secondary structures predicted by EternaFold, LinearFold, and LinearAlifold.** Secondary structures sampled by Eternafold, LinearFold, and LinearAilfold are compared to the reference secondary structure extracted from the listed PDB entry; n denotes the total number of secondary structures sampled.

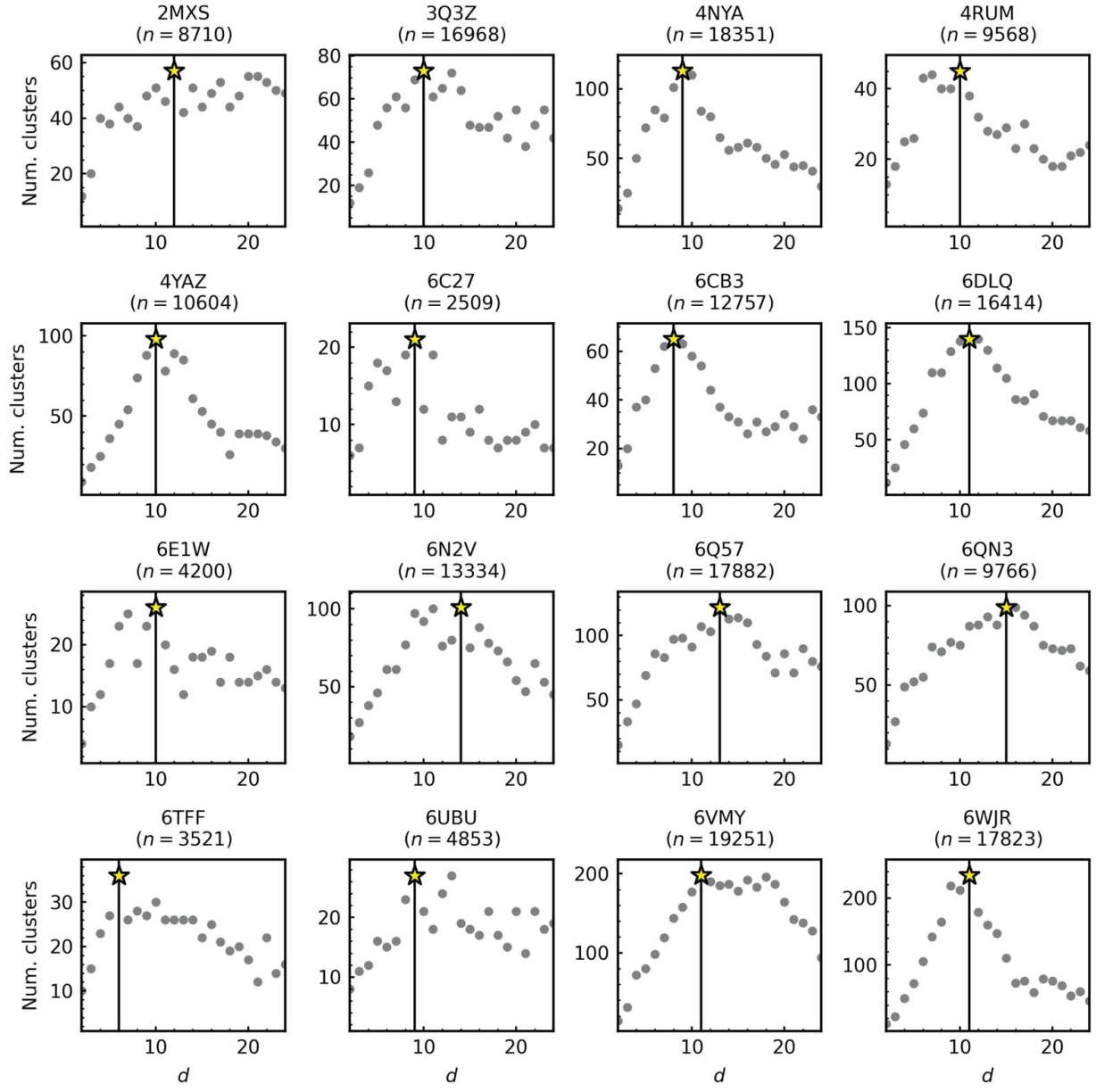


Figure D.4: **Number of clusters vs d.** For each RNA, the G-vectors are projected onto d components, and clustering is performed via the Advanced Density Peaks algorithm implemented in `dadapy`. The number of clusters, N_c is recorded as a function of d. The star values and vertical line indicate the d where N_c is maximized; n denotes the total number of models used in the calculation.

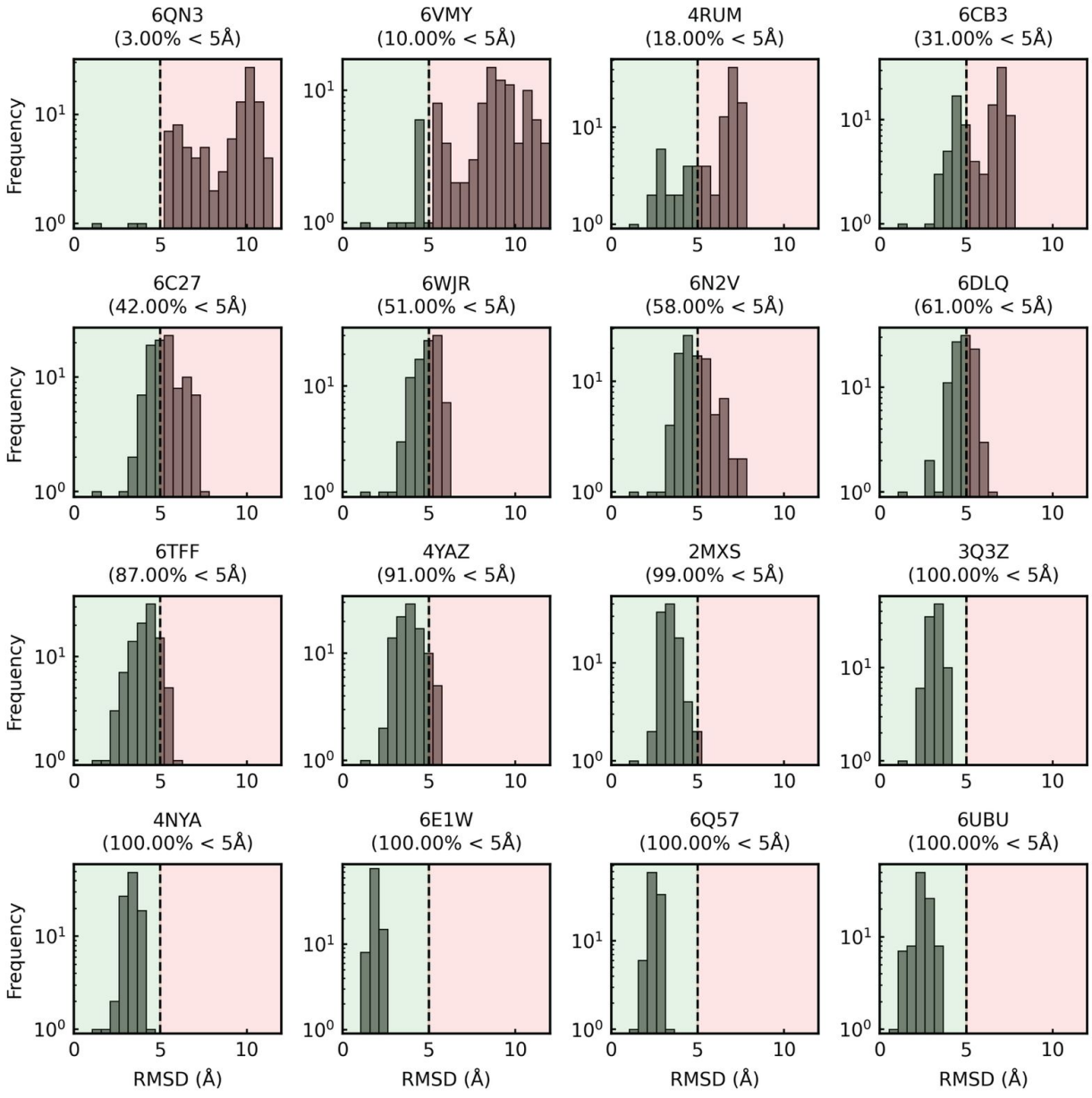


Figure D.5: **RMSD variation over a 100ns MD simulation launched from the reference conformation.** The dashed line placed at 5Å indicates the cutoff used to classify ERCs (RMSD≤ 5Å) and decoys (RMSD > 5Å) conformations. The proportion of ERCs is provided underneath the PDB accession code.

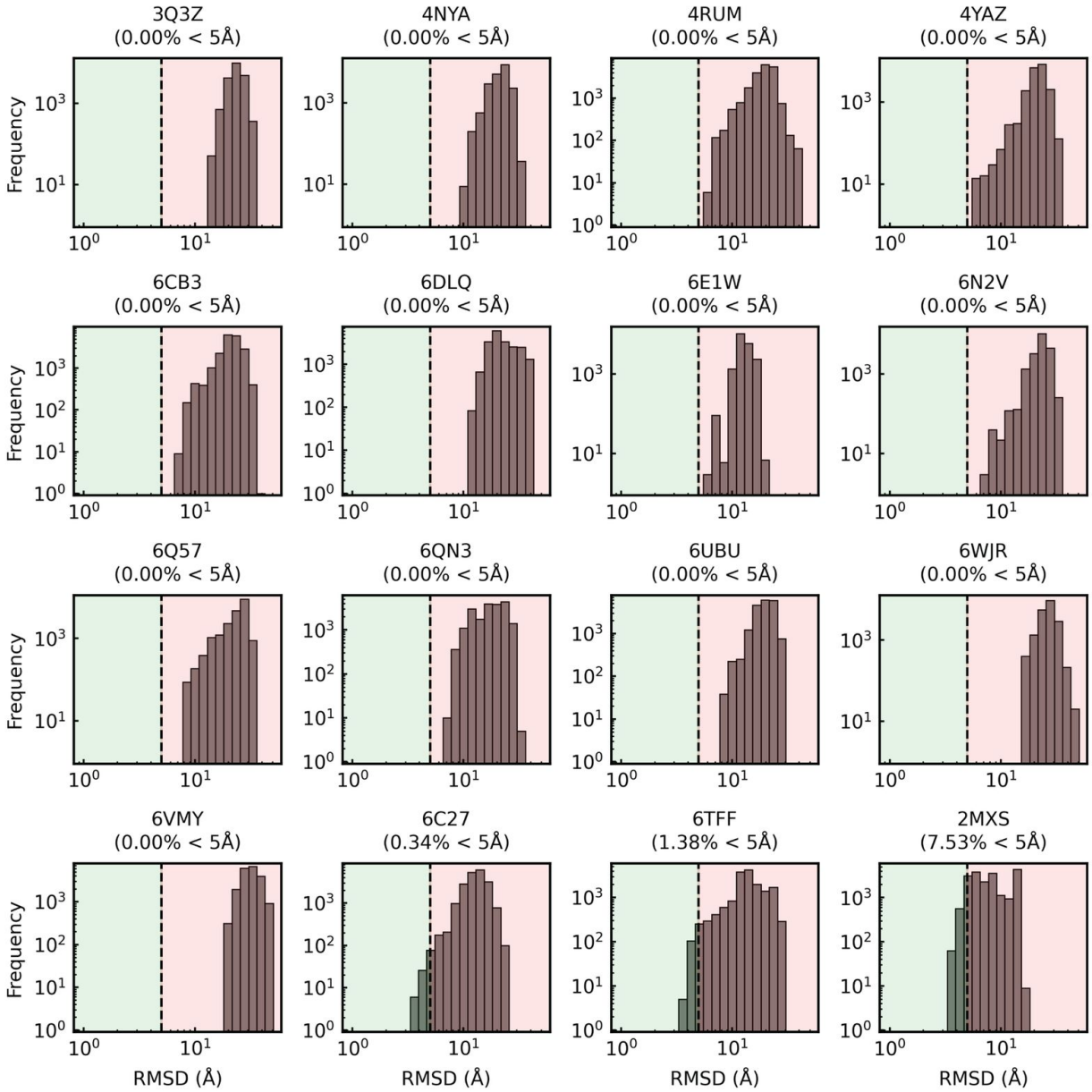


Figure D.6: **RMSD distribution of conformations sampled by RNAnneal.** The dashed line placed at 5Å indicates the cutoff used to classify ERCs (RMSD ≤ 5Å) and decoys (RMSD > 5Å) conformations. The proportion of near-ERCs is provided underneath the PDB accession code.

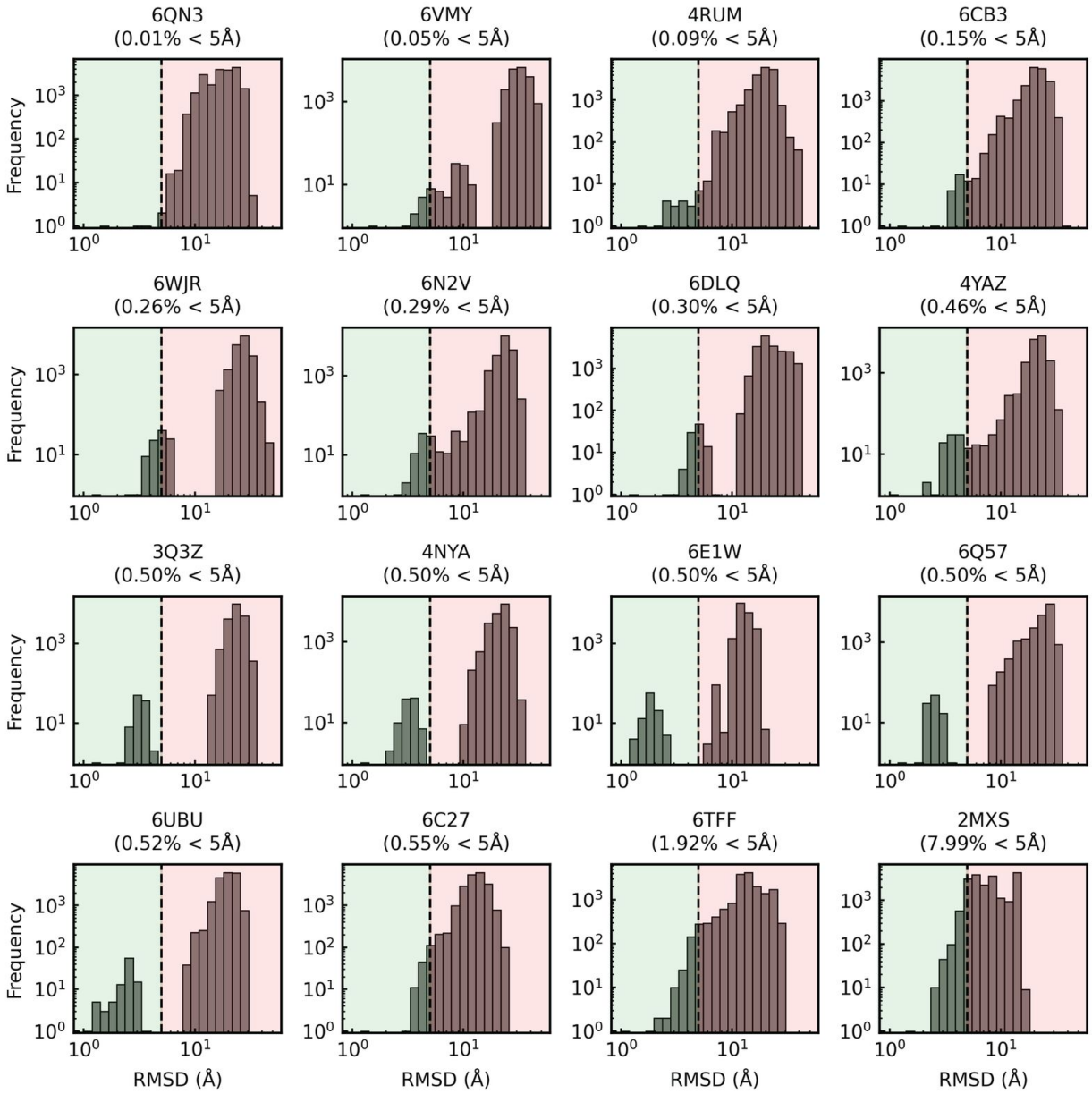


Figure D.7: **RMSD distribution of conformations for the classification task.** The classification task merges structures sampled by RNAnneal with frames from simulations launched from the reference conformations (i.e., merges Fig. SD.5 and SD.6). The dashed line placed at 5Å indicates the cutoff used to classify ERC (RMSD $\leq$ 5Å) and decoy (RMSD$>$ 5Å) conformations. The proportion of near-ERCs is provided underneath the PDB accession code.

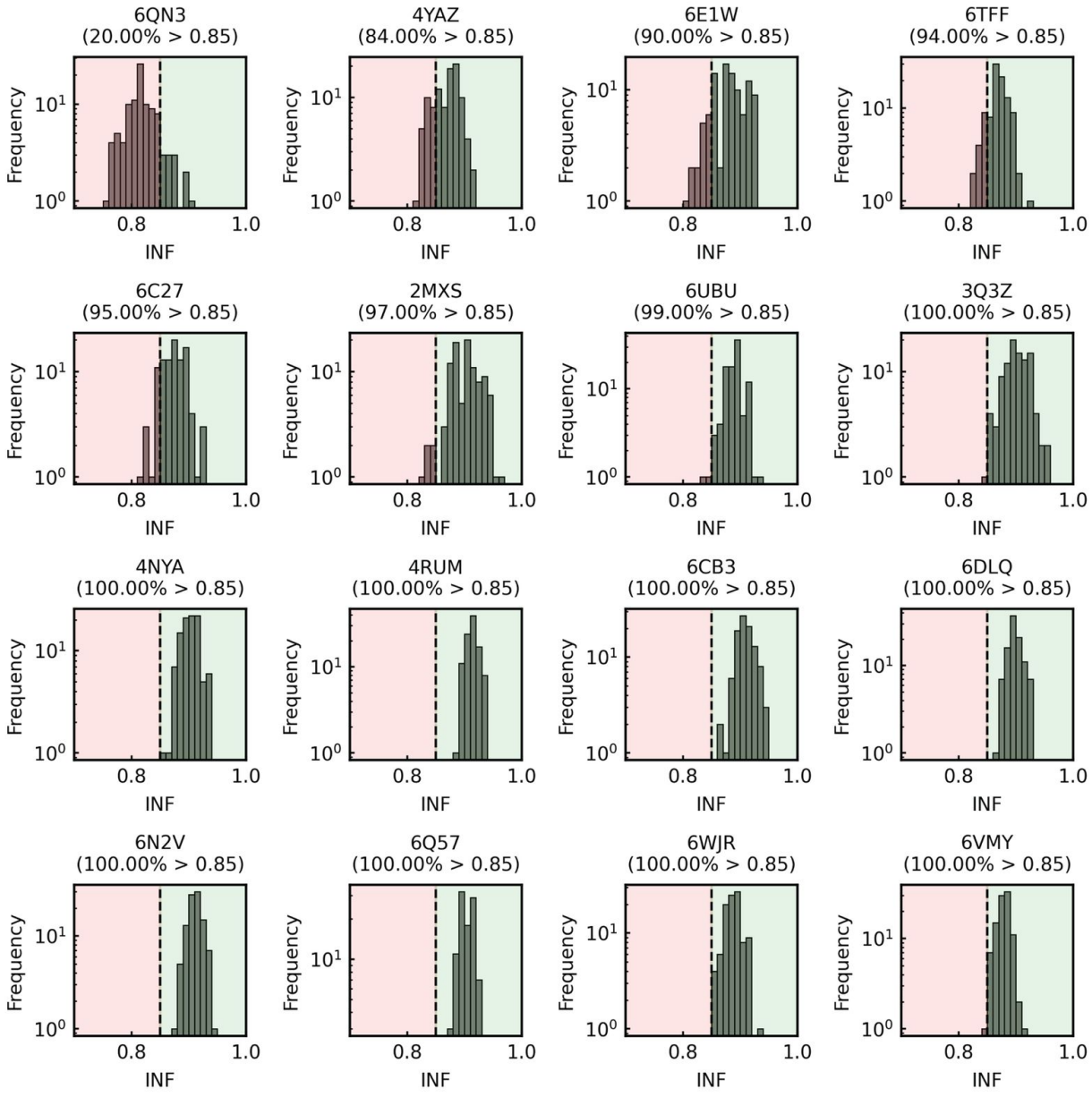


Figure D.8: **INF variation over a 100ns MD simulation launched from the reference conformation.** The dashed line placed at 0.85 indicates the cutoff used to classify ERC (INF$\geq$0.85) and decoy (INF$<$0.85) conformations. The proportion of near-ERCs is provided underneath the PDB accession code.

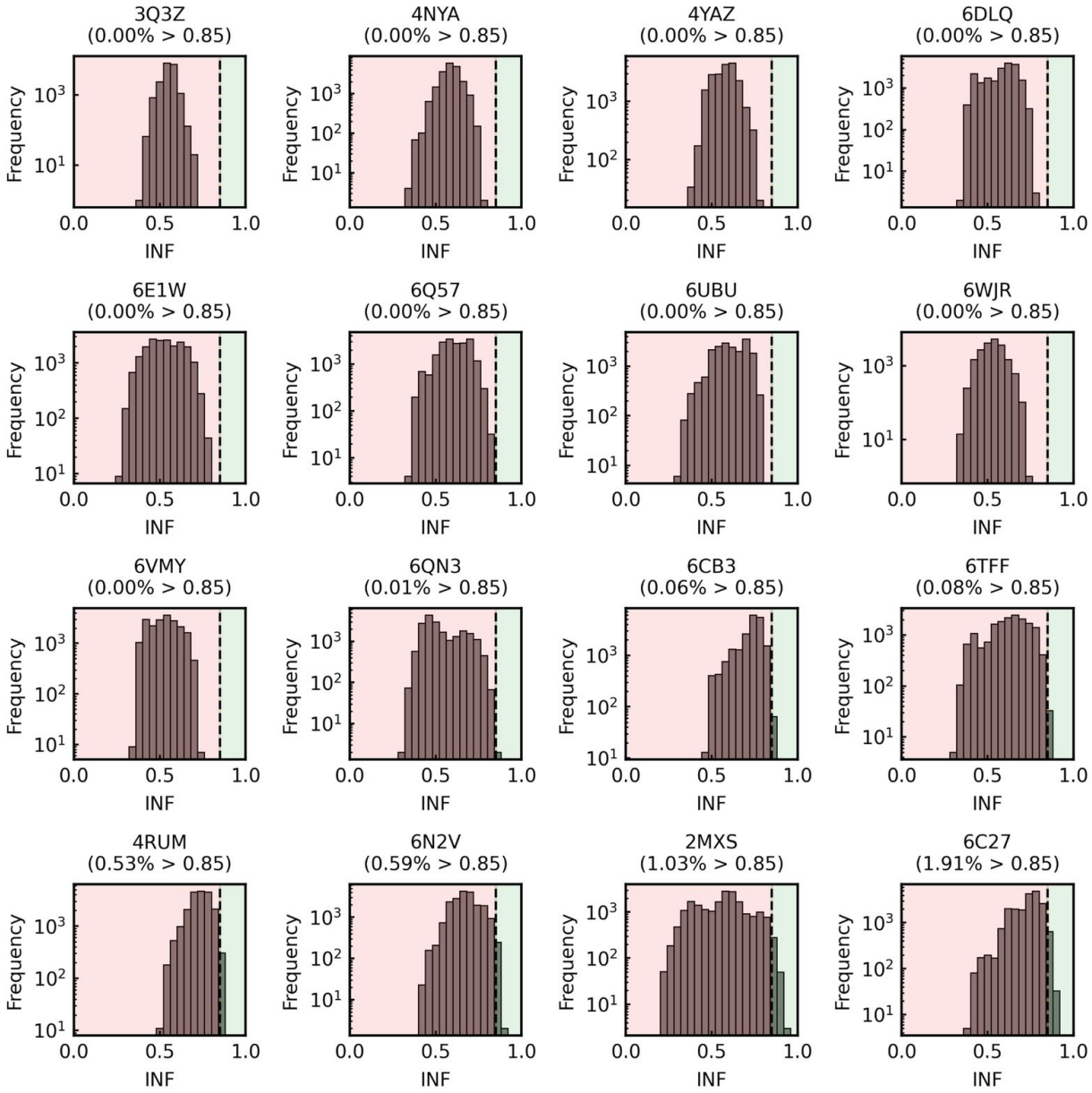


Figure D.9: **INF distribution of conformations generated by RNAnneal.** The dashed line placed at 0.85 indicates the cutoff used to classify ERC (INF$\geq$0.85) and decoy (INF$<$0.85) conformations. The proportion of near-ERCs is provided underneath the PDB accession code.

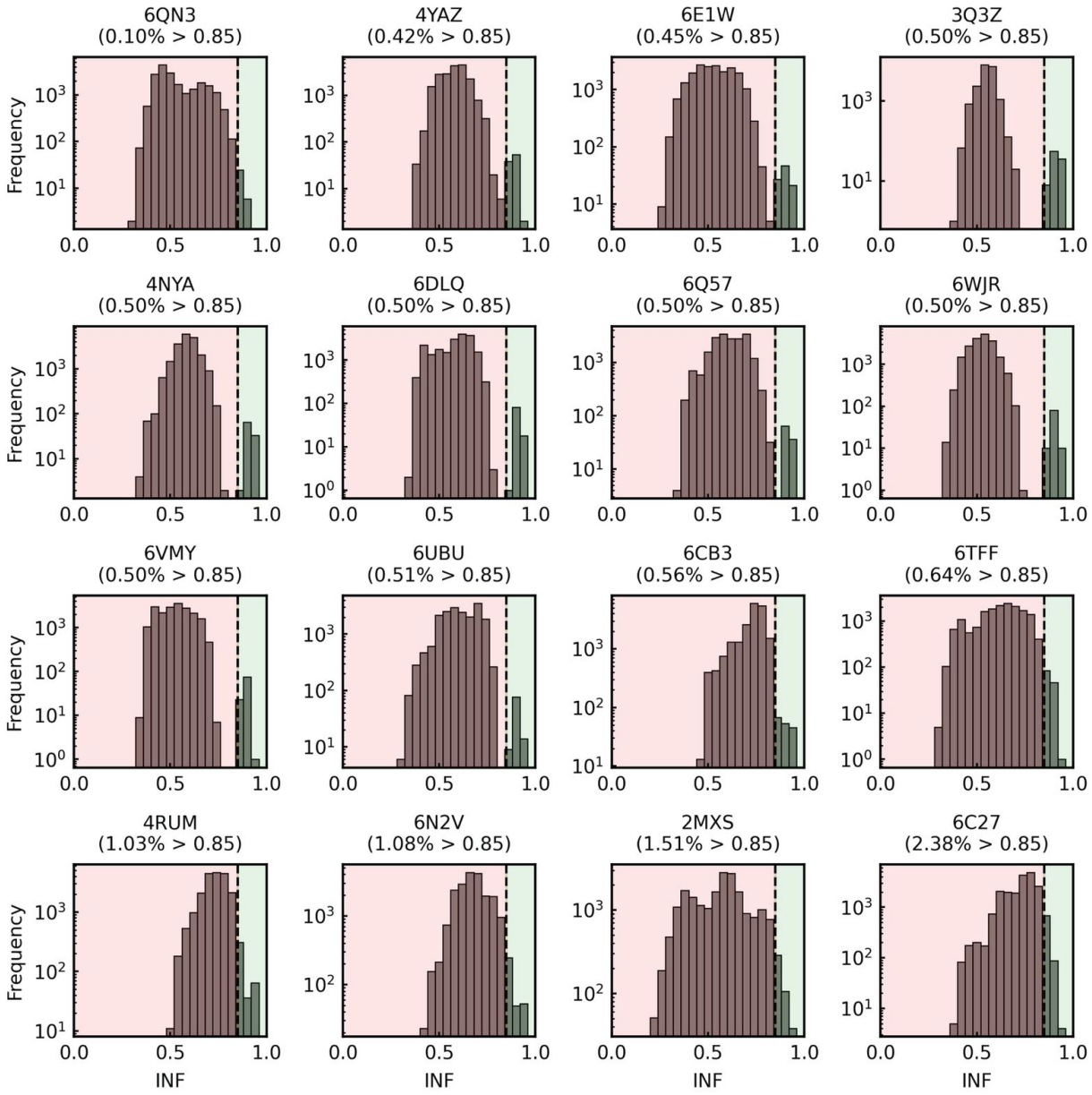


Figure D.10: **INF distribution of conformations for the classification task.** The classification task merges structures sampled by RNAnneal with frames from simulations launched from the reference conformations (i.e., merges Fig. SD.8 and SD.9). The dashed line placed at 0.85 indicates the cutoff used to classify ERC (INF$\geq$0.85) and decoy (INF$<$0.85) conformations. The proportion of near-ERCs is provided underneath the PDB accession code.

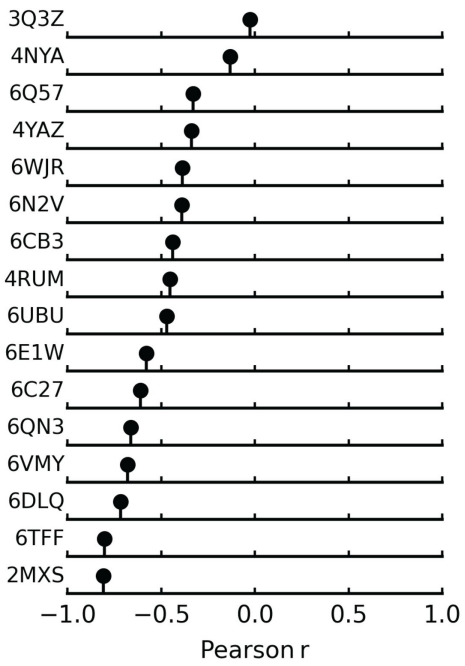


Figure D.11: **Correlation between RMSD and INF for RNAnneal conformations.**

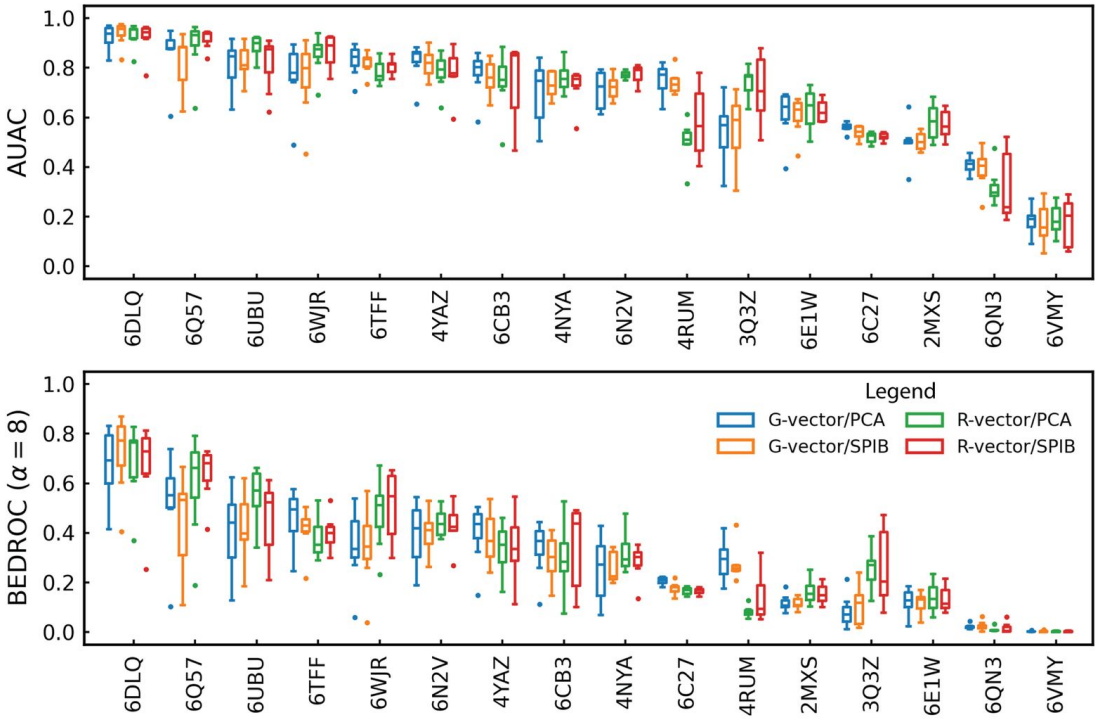


Figure D.12: **AUAC and BEDROC for each combination of training features.**

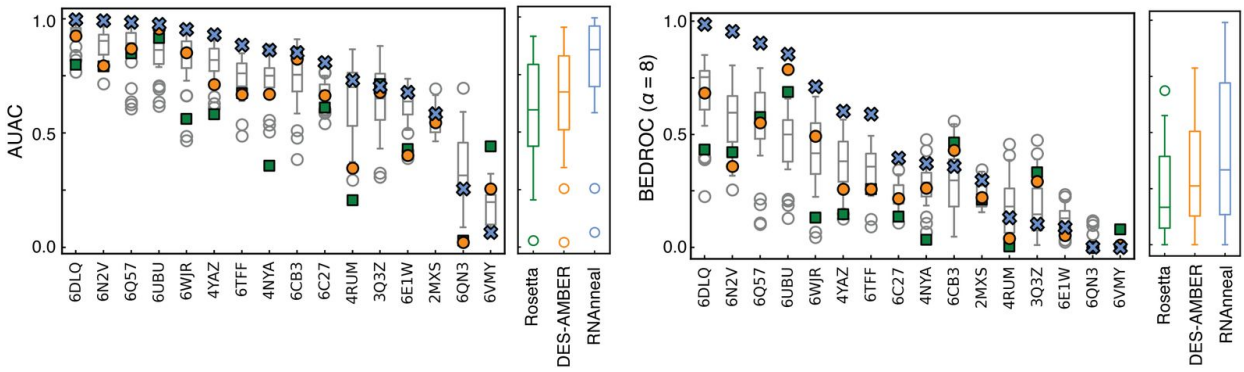


Figure D.13: **AUAC and BEDROC for RMSD classification criteria.** Classification power of the Rosetta (blue), DES-AMBER (orange), and RNAnneal (blue) scores when ERCs are defined as conformations < 5Å in RMSD from the reference conformation. The gray boxplots show the distribution of AUAC and BEDROC values for each model in the TM ensemble.

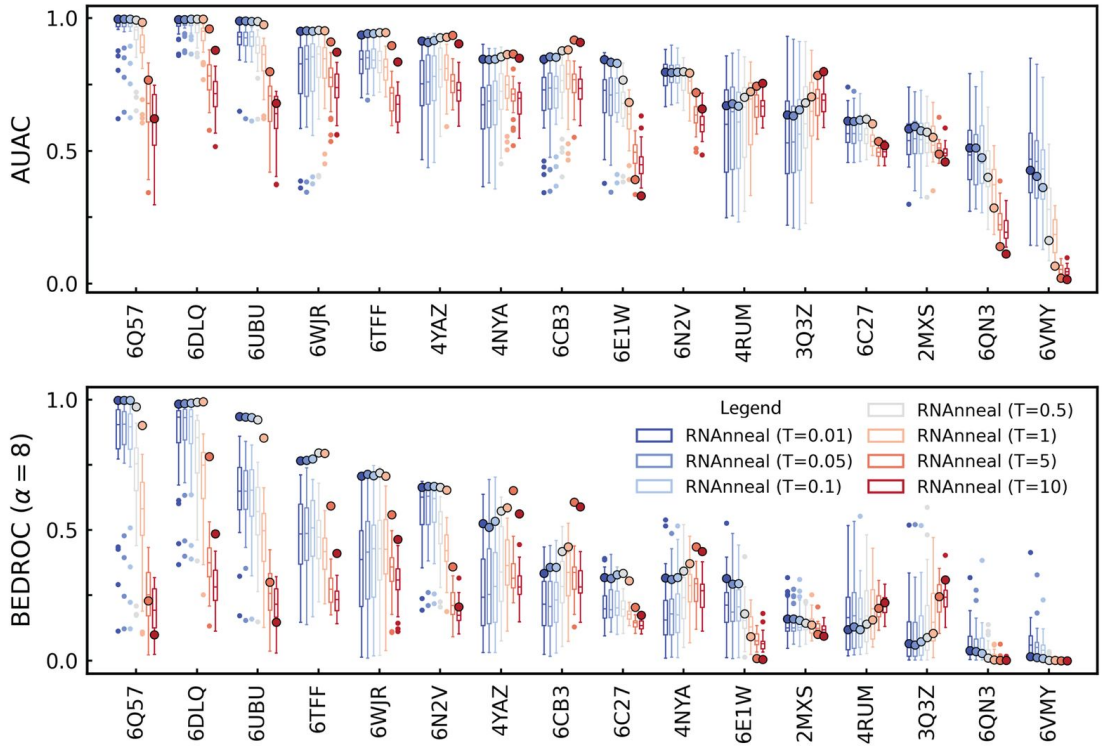


Figure D.14: **Variation in AUAC and BEDROC with scoring temperature.**

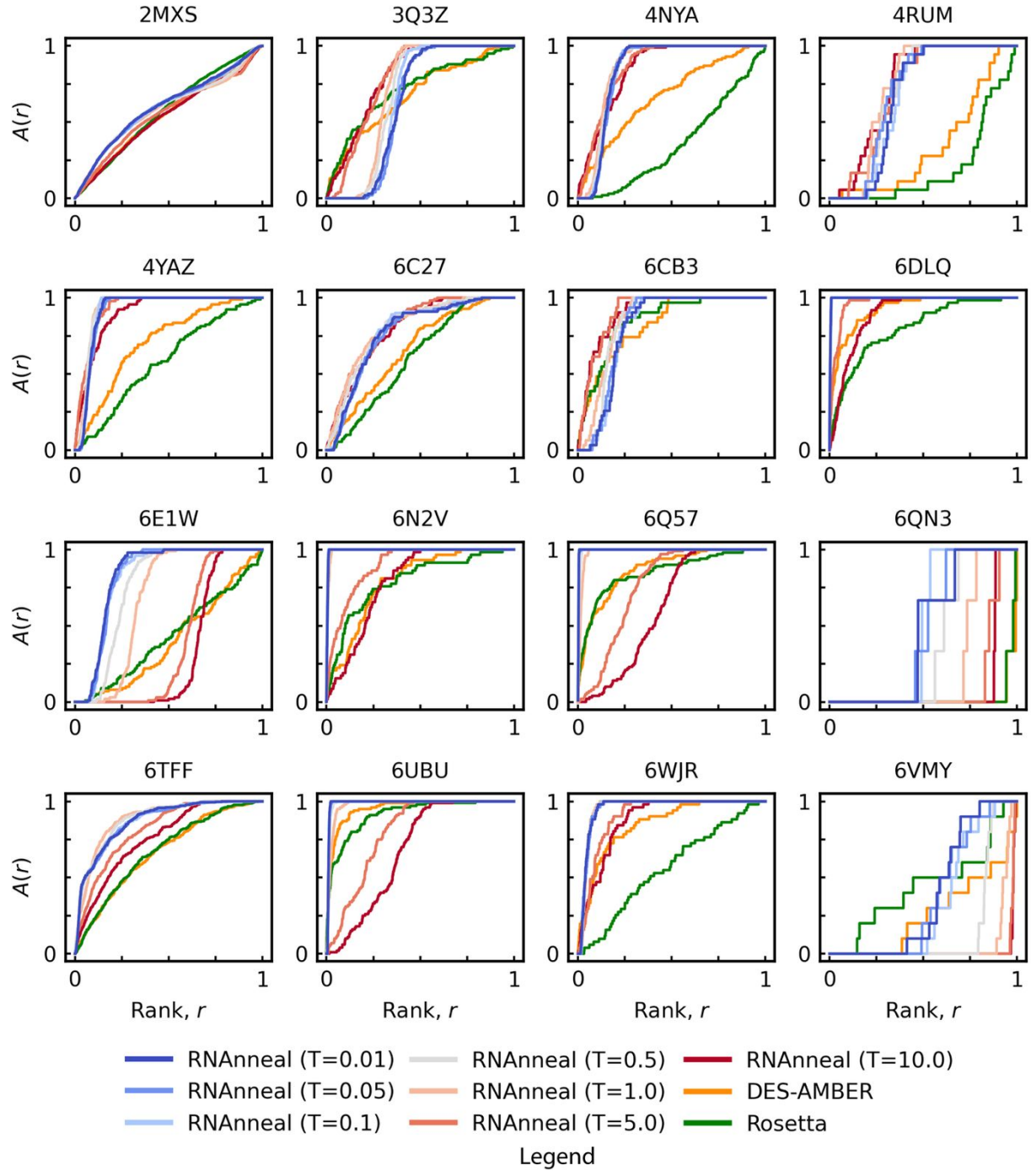


Figure D.15: **Accumulation Curves.** Accumulation curves, $A(r)$ are reported for each riboswitch, identified by the PDB entry of the reference conformation.

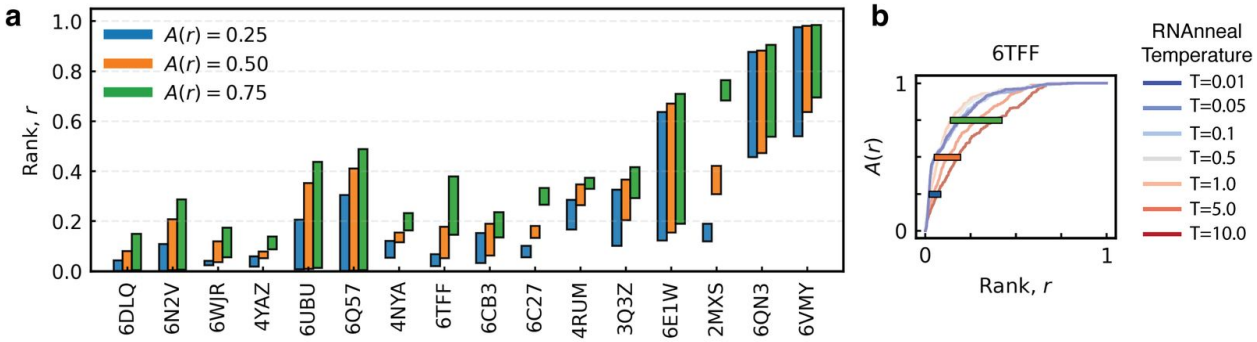


Figure D.16: **Summary of Accumulation Curves a.** The set of accumulation curves for each riboswitch PDB is shown with the range of rank, r, for which $A(r) = 0.25$, $A(r) = 0.50$, and $A(r) = 0.75$. **b.** The range indicated by the bars shows the variation of r with temperature at fixed values of $A(r) = 0.25$, $A(r) = 0.50$, and $A(r) = 0.75$. The accumulation curve for the NAD^+ riboswitch (6E1W [211]) is shown as an example (see Fig. SD.15 for all riboswitches).

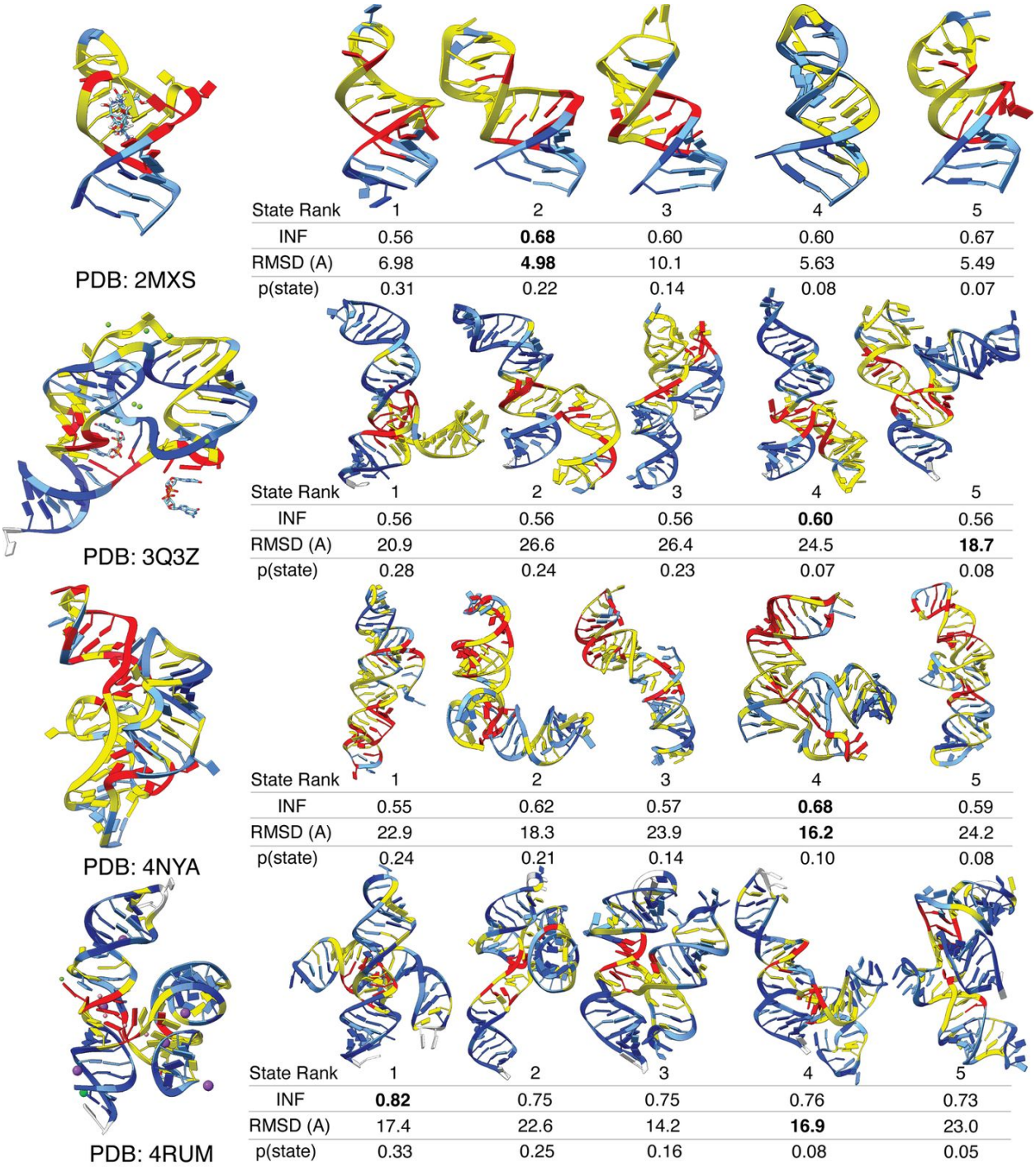


Figure D.17: **Conformational ensembles predicted by RNAnneal (2MXS, 3Q3Z [221], 4NYA [219], 4RUM [208]).** The reference ERC for each RNA are depicted to the left. The five most probable states are reported alongside their probabilities. The best scoring model from each state is shown as a representative conformation, and the INF and RMSD values compared to the ERC are reported. Nucleotides are colored based on the interaction entropy (IE) predicted by RNAnneal according to the ranges marked in Fig. 5.3b (dark blue: $0 \leq \text{IE} < 0.5$; light blue: $0.5 \leq \text{IE} < 1.5$; yellow: $1.5 \leq \text{IE} < 2.5$; red: $\text{IE} \geq 2.5$)

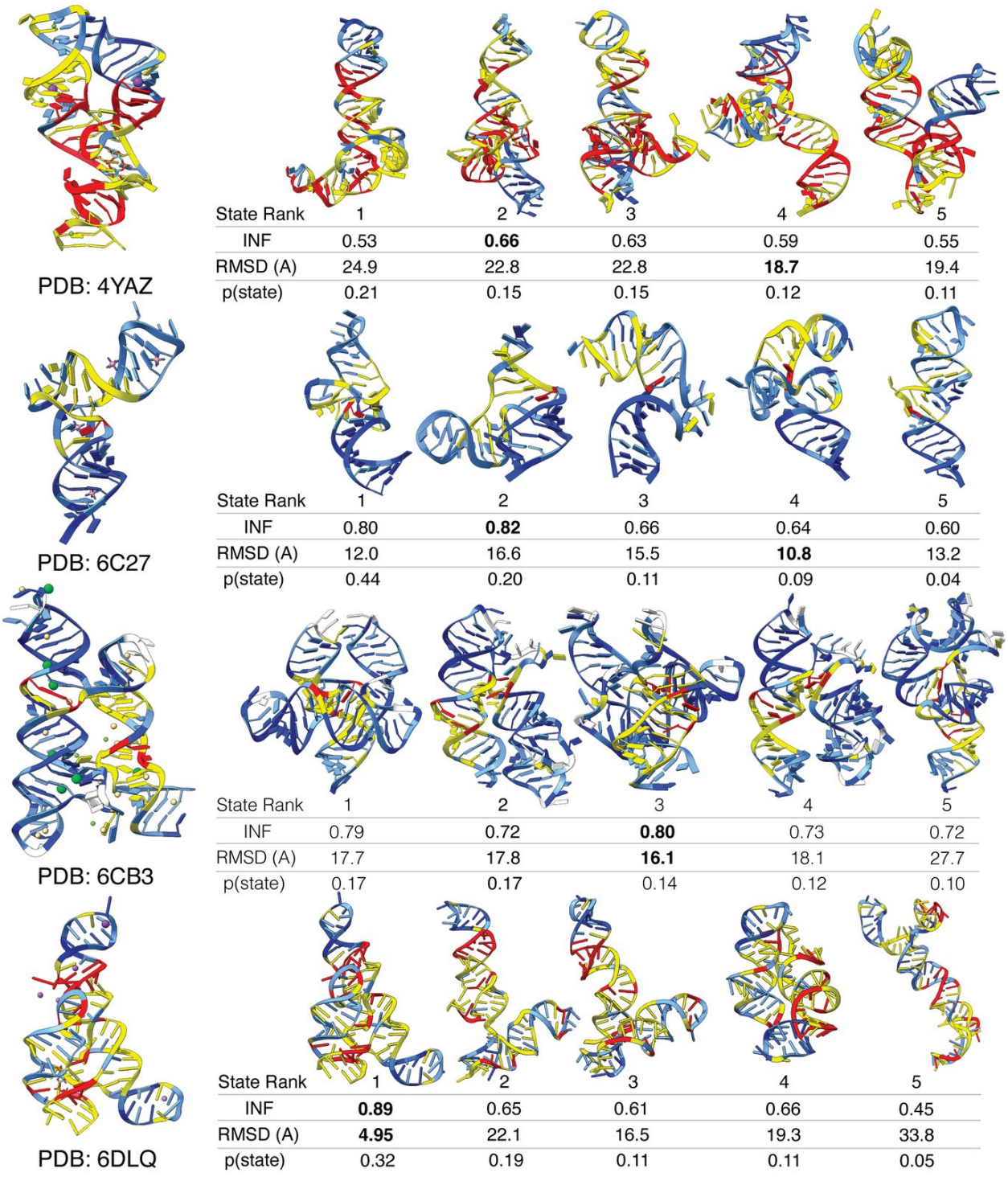


Figure D.18: **Conformational ensembles predicted by RNAnneal (4YAZ [220], 6C27 [217], 6CB3 [209], 6DLQ [213])**. The reference ERC for each RNA are depicted to the left. The five most probable states are reported alongside their probabilities. The best scoring model from each state is shown as a representative conformation, and the INF and RMSD values compared to the ERC are reported. Nucleotides are colored based on the interaction entropy (IE) predicted by RNAnneal according to the ranges marked in Fig. 5.3b (dark blue: $0 \leq \text{IE} < 0.5$; light blue: $0.5 \leq \text{IE} < 1.5$; yellow: $1.5 \leq \text{IE} < 2.5$; red: $\text{IE} \geq 2.5$)

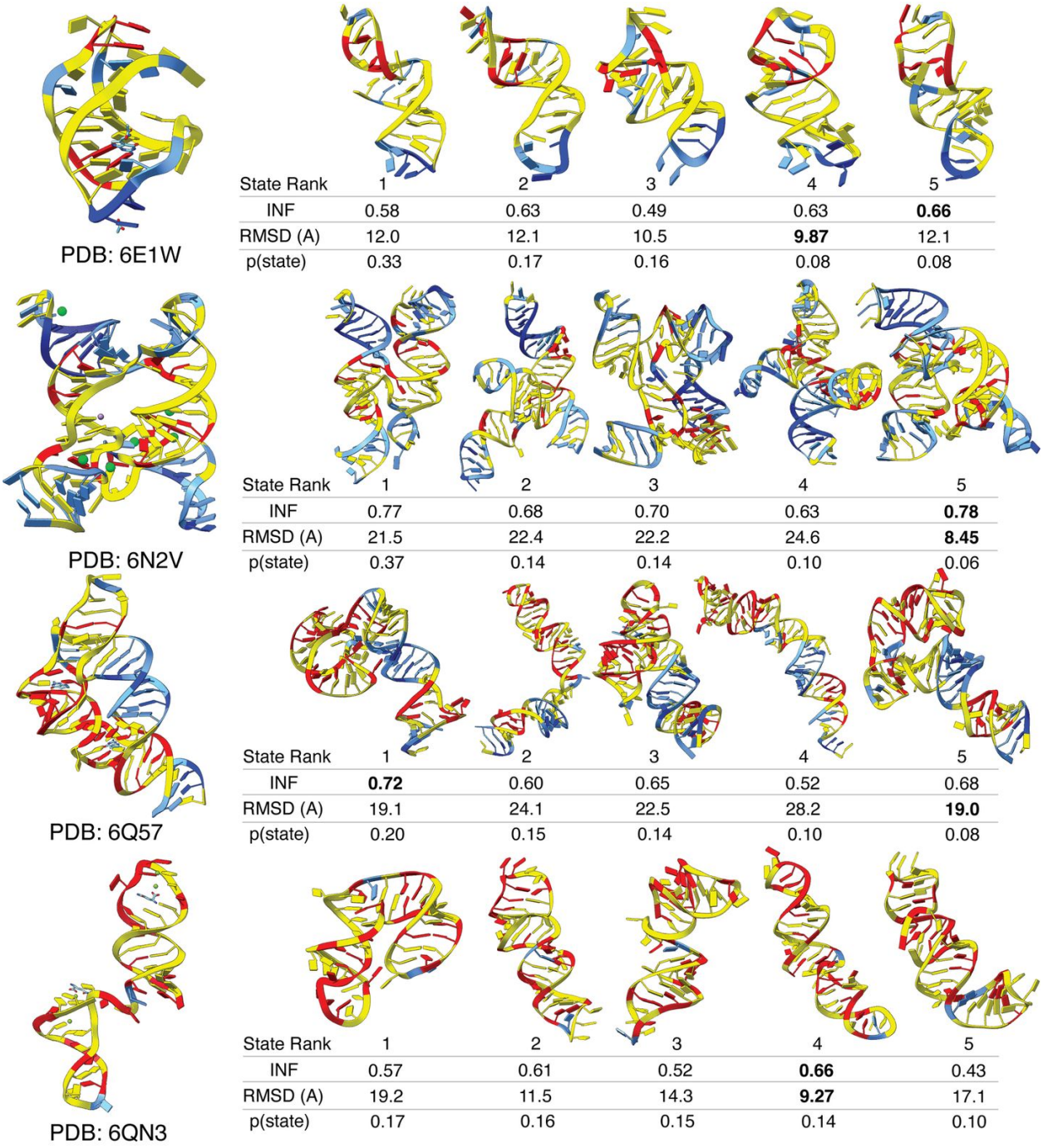


Figure D.19: **Conformational ensembles predicted by RNAnneal (6E1W [211], 6N2V [207], 6Q57 [214], 6QN3 [210])**. The reference ERC for each RNA are depicted to the left. The five most probable states are reported alongside their probabilities. The best scoring model from each state is shown as a representative conformation, and the INF and RMSD values compared to the ERC are reported. Nucleotides are colored based on the interaction entropy (IE) predicted by RNAnneal according to the ranges marked in Fig. 5.3b (dark blue: $0 \leq \text{IE} < 0.5$; light blue: $0.5 \leq \text{IE} < 1.5$; yellow: $1.5 \leq \text{IE} < 2.5$; red: $\text{IE} \geq 2.5$)

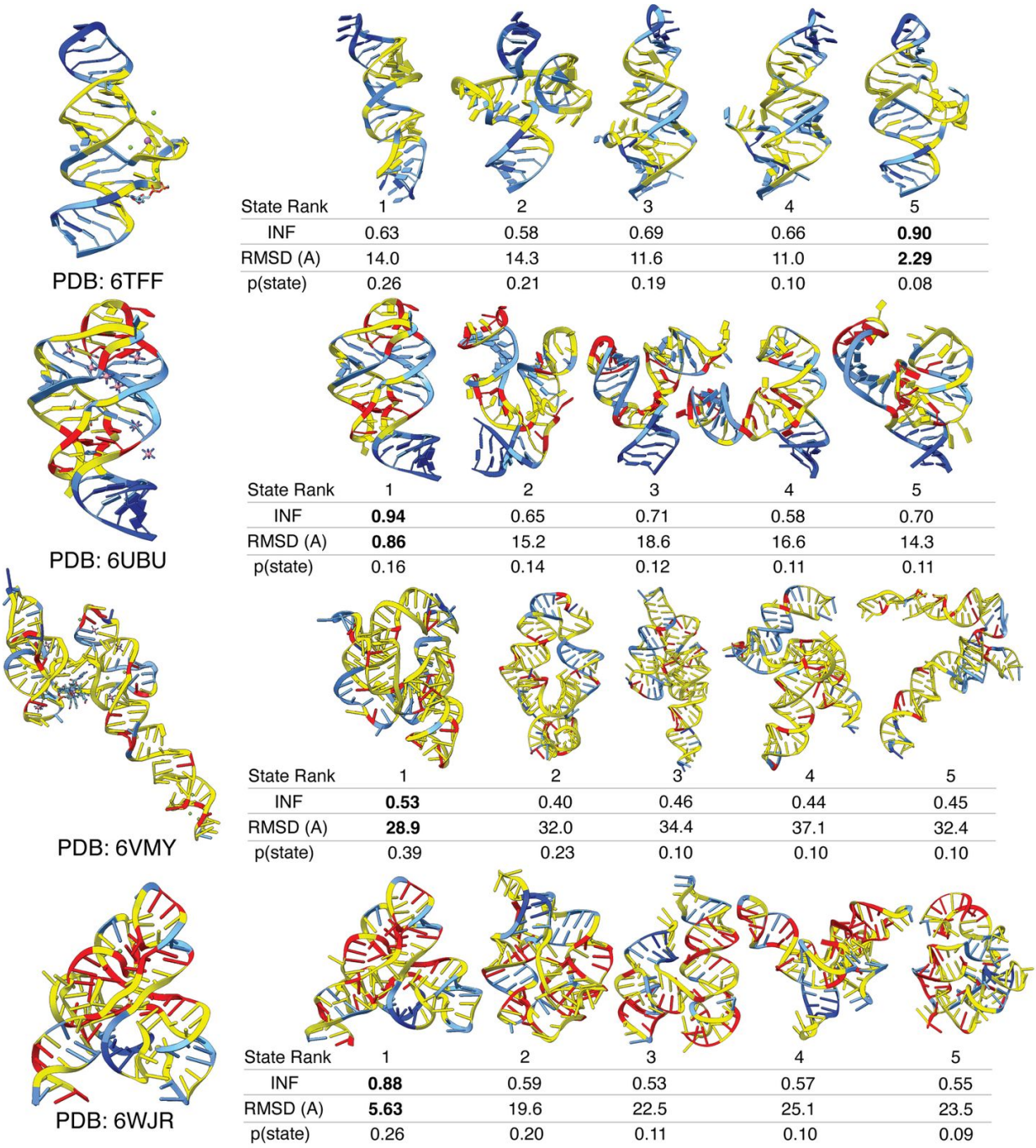


Figure D.20: **Conformational ensembles predicted by RNAnneal (6TFF [205], 6UBU [215], 6WJR [216], and 6VMY [218]).** The reference ERC for each RNA are depicted to the left. The five most probable states are reported alongside their probabilities. The best scoring model from each state is shown as a representative conformation, and the INF and RMSD values compared to the ERC are reported. Nucleotides are colored based on the interaction entropy (IE) predicted by RNAnneal according to the ranges marked in Fig. 5.3b (dark blue: $0 \leq \text{IE} < 0.5$; light blue: $0.5 \leq \text{IE} < 1.5$; yellow: $1.5 \leq \text{IE} < 2.5$; red: $\text{IE} \geq 2.5$)

Table D.1: Details of riboswitch sequences and reference conformations

RNA	PDB	Monomer	State	Length	Sequence	Secondary Structure
Neomycin sensing riboswitch	2MXS [206]	Y	Holo	28	ggcugcuuguccuuuaaug guccaguc	(.(((...(.((.....)).)))).)
c-di-GMP-II riboswitch	3Q3Z [221]	Y	Holo	76	gcgcggaaacaaugaugaa uggguuuaaauugggcacu ugacucauuuugaguuagu agugcaaccgaccgugcu	((.(((..((.......((((.((......[{{[[[[[..)).)))).....))....]]]]]].}}..))).)).
TPP riboswitch	4NYA [219]	Y	Holo	81	ggacucggggugcccuucu gcgugaaggcugagaaaua cccguaucaccugaucugg auaaugccagcguagggaa guuc	(.(((((((.((((.(((. ....)))))).....))..))))).....(((...((((......)))).)))..))).)
NiCo transition -metal riboswitch	4RUM [219]	Y	Holo	93	gggaacugagcaggcaaug accagagcggucaugcagc cgggcugcgaaagcggcaa cagaugauacacgcacauc uguggggacaguuccca	(((((((((..(.(((.(((((.....)))))).)..))).((.((....)).)).((((((.........))))))).).)))))))).
3',3'-cGAMP sensing riboswitch	4YAZ [220]	Y	Holo	84	gguacacgacaauacuaaa ccauccgcgaggauggggc ggaaagccuaagggucucc cugagacagccgggcugcc gaaauauc	(.((..((.....(.... ((((((....)))))))..[.)...(((...(((((((. ..)))))).])).)))...))...)).)
SAM-III riboswitch	6C27 [217]	Y	Apo	48	gggacaaguucccgaaagg auggcggaaacgccagaug ccuugucc	.(((((((.(.(.....).)(.((.....)).).... .)))))))).
YkoY riboswitch	6CB3 [209]	N	Holo	100	aaaggggaguagcgucggg aaaccgaaacaaagucguc aauucgugaggaaacucac cggcuuuguugacauacga aaguauguuuagcaagacc uuuc	.((((.....((.((((. ...)))).(((((((.((.. .....(((((....))))))).)))))))..((((((.. ..))))))...))....))))..
PRPP riboswitch	6DLQ [213]	Y	Holo	108	ggaaaguguguctaggguu ccgcgugcuucggcacgga cugguccaagugacacaga cgcauucgugcguuacacc ggagggauagaagcccagg cggguagguuuc	..((...(((((..((... ((.(((((....))))).. ..)).))....)))))..(((((....)))))..(.[[[..(((.......))).).]]].....)).
PreQ1 riboswitch	6E1W [211]	Y	Holo	31	cugggucgcaguuucccca guuaacaaaac	(((((.....[....))))).........]
Mn^{2+}-sensing riboswitch	6N2V [207]	Y	Holo	99	gcuuggggaguagccugcu uucggaaacgaaagcgccu guaucaacauacucggcga aagccguggugcaggaccg aaaggucuggcgagaccag gcc	.(.(((.....(((..(. ((((....)))).)..(((.(.((.......((((.. ..))))))).).)))).(((. ...)))..)))....))).)..
Tetrahydro-folate riboswitch	6Q57 [214]	Y	Holo	90	ggagaguagaugauucgcg uuaaguguguugaauggg augucgucacacaacgaag cgagagcgcggugaaucau ugcauccgcucca	.(((.....((((((.((.......(((((((...[[[[[...)))))))....((....)).)).)..)))))).. .]]]]].))))..

RNA	PDB	Monomer	State	Length	Sequence	Secondary Structure
Glutamine II riboswitch	6QN3 [210]	N	Holo	51	cguucacccuucggggcgc augaaaugggaguagggaa cgggauucucau	......(((....))).... .....(.(.((.(..... ....).)).).)
NAD$^+$ riboswitch	6TFF [205]	Y	Holo	53	ggcuucaacaaccccguag guugggccgaaaggcagcg aaucuacuggagcc	((((.((.....[[.((((.(((.(((....))).]]))).)))))).))))
Guanine riboswitch	6UBU [215]	Y	Holo	68	ggacauauaaucgcgugga uauggcacgcaaguuucua ccgggcaccguaaaugucc gacuaugucc	(((((((..(.(((((.. ...[[)))))).)[.....)](((.((]].....)).)))..))))))
FMN riboswitch	6WJR [216]	Y	Apo	112	ggaucuucggggcagggug aaauucccgaccgguggua uaguccacgaaaguauuug cuuugauuuggugaaauuc caaaaccgacaguagaguc uggaugagagaagauu	..((((((.......((.. .....)).((((((((.. ..[[)))))..(((.....)))...(((((.......)))))]]))).[[[....))]]].......)))))).
Cobalamin riboswitch	6VMY [218]	Y	Holo	149	ggucaaauaggugccgguc cgugaacaacagccggcuu aaaagggaaaccgguaaaa gccggugcggucccgccac uguaauuggccaagcgcca agagccaggauaccugccu guuugaucagcacgaauuc ugcgaggacagauga	(.((.((.(((.(((((.. [((.....)).)))))).. ...(((....((((..... .)))).((((.....)).[[[....((((.....)))) ...))]]]]...))).))) .)).)).)..(.....((. ....).)...).

Stage	Duration (ps)	Timestep (fs)	Temperature (K)
Equilibration 1	1	0.1	1
Equilibration 2.1	1	1	10
Equilibration 2.2	1	1	20
Equilibration 2.3	1	1	30
Equilibration 2.4	1	1	40
Equilibration 2.5	1	1	50
Equilibration 3.1	2	1	100
Equilibration 3.2	2	1	150
Equilibration 3.3	2	1	200
Equilibration 3.4	2	1	250
Equilibration 3.5	2	1	300
Equilibration 4.1	20	1	300
Equilibration 4.2	20	2	300
Equilibration 4.3	20	3	300
Equilibration 4.4	20	4	300
Equilibration 4.5	20	5	300
Production	100,000	5	300

Table D.2: Molecular Dynamics Equilibration Procedure

Parameter	Value
Simulation Engine	OpenMM
Ion Type	NaCl
Ion Concentation	1 M
Box Shape	Dodecahedron
Box Padding	0.5 nm
Force Field (Solute)	RNA.Shaw_charmm22-ions.xml
Force Field (Solvent)	implicit/gbn2.xml
Pressure	1 atm
Sampling Frequency	1 ns

Table D.3: Molecular Dynamics Simulation Parameters.

Algorithm 2 Sampling RNA secondary structures with `EternaFold`

Require: `example.fa` (input sequence);
`N` (maximum number of samples, `5000`)

```
mkdir -p eternafold
EternaFold/src/contrafold sample example.fa
 -params EternaFold/parameters/EternaFoldParams.v1
 -nsamples N > eternafold/ef.ss
```

Algorithm 3 Sampling RNA secondary structures with `LinearFold`

Require: `example.fa` (input sequence)

```
mkdir -p linearfold-v {Vienna parameters}
mkdir -p linearfold-c {CONTRAfold parameters}
cat example.fa | linearfold -zuker -delta 1000 -V | tail -n +5
 | awk '{print $1}' > linearfold-v/zuker.ss
cat example.fa | linearfold -zuker -delta 1000 | tail -n +5 |
 awk '{print $1}' > linearfold-c/zuker.ss
```

Algorithm 4 Sampling RNA secondary structures with `LinearAliFold`

Require: `example.fa` (input sequence);
`N` (maximum number of samples, `20000`)

```
mkdir -p linearalifold-v {Vienna parameters}
mkdir -p linearalifold-bl {BL* parameters}
cat example.fa | linearalifold.py -s N -e 1 >
linearalifold-v/stochastic.ss
cat example.fa | linearalifold.py -s N -e 2 >
linearalifold-bl/stochastic.ss
cat example.fa | linearalifold.py -p -B linearalifold-v/bpp.txt
-M linearalifold-v/mea.ss -T linearalifold-v/threshknot.ss -C
linearalifold-v/centroid.ss -e 1
cat example.fa | linearalifold.py -p -B linearalifold-bl/bpp.txt
-M linearalifold-bl/mea.ss -T linearalifold-bl/threshknot.ss
-C linearalifold-bl/centroid.ss -e 2
```

Algorithm 5 Conformational sampling with FARFAR2

Require: `SEQ` (input sequence);
`DB` (secondary structure in dot-bracket notation)

```
rna_denovo.default.linuxgccrelease
 -sequence SEQ {From example.fa in Algorithms 2-4}
 -secstruct DB {cat */*.ss | uniq from Algorithms 2-4}
 -nstruct 1
 -cycles 20000
 -minimize_rna true
 -rna:denovo:rounds 10
 -rna:denovo:minimize:minimize_bps true
 -rna:denovo:bps_moves true
```

Algorithm 6 Reference conformation preparation with FARFAR2

Require: `SEQ` (input sequence);
`REF` (reference PDB file)

```
rna_denovo.default.linuxgccrelease
 -sequence $(echo $SEQ) {lower-case sequence}
 -secstruct '.....' {all bases unpaired; equal to length of $SEQ}
 -nstruct 1
 -out:file:silent $(dirname $REF)/silent.out
 -s $REF {experimental model}
 -minimize_rna true
 -rna:denovo:rounds 1
 -out:overwrite true
 -cycles 10000
 -rna:denovo:minimize:minimize_bps true
```

Algorithm 7 Unsupervised selection of diverse candidate conformations

Require: FARFAR2 models $\{\mathbf{x}_i\}_{i=1}^{N}$;
Density-peaks parameter $Z{=}3$;
Principal component range $d{=}2, 3 \ldots 26$;
Target count $K{=}200$

Featurize each model $\mathbf{x}_i$ into a G-vector
for $d \in \{2, 3, \ldots, 26\}$ **do**
  Project all G-vectors onto the first d principal components
  Cluster the projection with `dadapy` (advanced density peaks, $Z{=}3$)
  Record the number of clusters obtained
end for
Let d^* be the dimension giving the maximum cluster count {see Fig. SD.4}
Re-cluster the d^*-dimensional data to obtain $\{\mathcal{C}_j\}_{j=1}^{M}$
for all clusters $\mathcal{C}_j$ **do**
  Sort conformations in $\mathcal{C}_j$ by ascending Rosetta score
end for
Sort the clusters by the best Rosetta score among all conformations in each cluster
$\mathcal{S} \leftarrow \varnothing$ {set of selected candidates}
while $|\mathcal{S}| < K$ **do**
  for $j = 1$ **to** M **do**
    if $\mathcal{C}_j \neq \varnothing$ **then**
      Remove the top-scoring conformation c from $\mathcal{C}_j$ and add c to $\mathcal{S}$
      if $|\mathcal{S}| = K$ **then**
        break
      end if
    end if
  end for
  Re-sort clusters by their current top-scoring conformation
end while
return $\mathcal{S}$ {200 candidate conformations}

Algorithm 8 Likelihood estimation for a thermodynamic map (diffusion model) using the Hutchinson trace estimator

Require: Observed frame $\mathbf{u}_0 \equiv (\mathbf{x}_0, U_0)^\top$ (data at $t_0 \equiv t = 0$);
Diffusion process $\mathcal{D}$ with time grid $\{t_0 < t_1 < \ldots < t_K\}$;
One–step map $\mathrm{STEP}_{t_k \to t_{k+1}} : \mathbf{u} \mapsto \mathbf{u}'$; {Neural network}
Prior temperature T; {$T = 1$ unless otherwise stated}
Temperature-dependent prior density $q_T(\mathbf{u}_{t_K}) \equiv \mathcal{N}(0, T) \times \mathcal{N}(0, 1)$; {Different priors for $\mathbf{x}$ and U}
Schedule of noise scales $\{\beta_k\}_{k=1}^K$;
Number of Hutchinson probe vectors M

$\mathbf{u} \leftarrow \mathbf{u}_0$ {state at the current time t_k}
$\mathcal{G} \leftarrow [\,]$ {to store divergence g_k}
for $k = 0$ **to** $K - 1$ **do**
  Draw M probe vectors $\{\mathbf{v}^{(m)}\}_{m=1}^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for all probes $m = 1, \ldots, M$ **do**
    $(\mathbf{u}'^{(m)}, \mathbf{J}\mathbf{v}^{(m)}) \leftarrow \mathrm{JVP}\Big(\mathrm{STEP}_{t_k \to t_{k+1}}, \mathbf{u}, \mathbf{v}^{(m)}\Big)$ {Jacobian-vector product}
  end for
  $g_k \leftarrow \dfrac{1}{M} \displaystyle\sum_{m=1}^{M} \big\langle \mathbf{J}\mathbf{v}^{(m)}, \mathbf{v}^{(m)} \big\rangle$ {Hutchinson trace estimate of $\nabla_{\mathbf{u}} \cdot \mathrm{STEP}$}
  Append g_k to $\mathcal{G}$
  $\mathbf{u} \leftarrow \mathbf{u}'^{(1)}$ {all probes have the same output}
end for
$\Delta\ell_{pq} \leftarrow \mathrm{trapz}\big(\mathcal{G}, \{\beta_k\}_{k=1}^K\big)$ {trapezoidal integration on (g_k, β_k)}
$\log q \leftarrow \log q_T\big(\mathbf{u}\big)$ {prior log-density at t_K at temperature T}
$\log p_\theta\big(\mathbf{u}_0\big) \leftarrow \log q + \Delta_{pq}$
return $\log p_\theta(\mathbf{u}_0)$

Algorithm 9 RNAnneal Score

Require: Log-likelihoods $\{L_i^{(\mathcal{F})}\}$;
Feature sets $\mathcal{F} = \{\text{G-vector}, \text{R-vector}\} \times \{\text{PCA}, \text{SPIB}\}$;
Component range $\mathcal{D} = 2, 3, \ldots, 8$

$L_{\min} \leftarrow \min_i L_i, \quad \sigma_L \leftarrow \sqrt{\text{Var}\{L_i\}}$
for all frames i **do**
  $l_i \leftarrow \big(L_i - L_{\min}\big) \,/\, \sigma_L$ {subtract min and divide by SD}
end for
Sort $\{l_i\}$ in ascending order
Initialize lists $\mathcal{S} = [\,]$ and $\mathcal{L} = [\,]$
for all confrmers i **do**
  $\mathcal{L}(i) \leftarrow [\,]$
  for all feature-projection pairs $(v, f) \in \mathcal{F}$ **do**
    for all dimensions $d \in \mathcal{D}$ **do**
      Append $l_i^{(v,f,d)}$ to $\mathcal{L}(i)$
    end for
  end for
  Append $\bar{s}_i = \dfrac{1}{|\mathcal{L}(i)|} \displaystyle\sum_{l \in \mathcal{L}(i)} l$ to $\mathcal{S}(i)$ {RNAnneal Score}
end for
return $\mathcal{S}(i)$

Algorithm 10 Interaction Entropy calculation

Require: Ensemble of conformations $\{X_1, \ldots, X_M\}$;
Nucleotides $i = 1, \ldots, N$;
Prior weights $\{w_1, \ldots, w_M\}$ with $\sum_j w_j = 1$;
LW classification function $\text{LW}(i, X_j) \to c \in \mathcal{C}$.

for all nucleotides $i = 1, \ldots, N$ **do**
  Initialize frequency table $f_i(c) \leftarrow 0$ for all $c \in \mathcal{C}$
  for all conformations $j = 1, \ldots, M$ **do**
    $c \leftarrow \text{LW}(i, X_j)$
    $f_i(c) \leftarrow f_i(c) + w_j$
  end for
  $p_i(c) \leftarrow f_i(c) / \sum_{c' \in \mathcal{C}} f_i(c')$ {Normalize}
  $H_i \leftarrow -\sum_{c \in \mathcal{C}} p_i(c) \log_2 p_i(c)$
end for
return $\{H_i\}_{i=1}^N$ {interaction entropies for all nucleotides}

Algorithm 11 Boltzmann-weighted state ranking and representative selection

Require: SPIB projections $\{\mathbf{x}_i\}$;
State labels $\{y_i\}$;
RNAnneal scores $\{s_i\}$;
Frame keys $\{f_i\}$;
Histogram grid size $(K = 25, L = 25)$;
Minimum population $N_{\min} = 10$;
Pre-filter count $k_{\text{top}} = 10$;
Report count $k_{\text{rep}} = 10$

$\mathcal{F} \leftarrow \{(y_i, \mathbf{x}_i, s_i, f_i)\}$
Group $\mathcal{F}$ by state label to obtain $\{\mathcal{F}_y\}$
Discard states with $|\mathcal{F}_y| < N_{\min}$
Compute the average score $\bar{s}_y$ for every surviving state
Keep the k_{top} states with the lowest $\bar{s}_y$
Initialise histogram H on an $K \times L$ grid over 2D IB space
for all frames $(y, \mathbf{x}, s, f) \in \bigcup \mathcal{F}_y$ **do**
    Increment H at the bin containing $\mathbf{x}$ with (s, y)
end for
for all occupied bins (k, l) **do**
    $\bar{s}_{kl} \leftarrow$ mean score in bin (k, l)
    $w_{kl} \leftarrow \exp(-\bar{s}_{kl})$ {Boltzmann weight in bin h_{kl}}
    $y^* \leftarrow \arg\max_y$ {count of y in (k, l)} {State assigned to bin h_{kl}}
    $W_{y^*} \leftarrow W_{y^*} + w_{kl}$ {Each bin h_{kl} with state y^* contributes w_{kl} to the total state weight W_{y^*}}
end for
$p_y \leftarrow W_y \,/\, \sum_{y'} W_{y'}$ {Normalize weights}
Order states by p_y (descending) to obtain $(y_{(1)}, \ldots, y_{(k_{\text{rep}})})$
for $r = 1$ **to** k_{rep} **do**
    $y \leftarrow y_{(r)}$
    $\mathcal{F}_y \leftarrow \{(s, f) \in \mathcal{F}_y\}$
    $f^* \leftarrow \arg\min_{(s,f)\in\mathcal{F}_y} \{s\}$ {representative frame}
    Record $\left(r, y, p_y, f^*\right)$
end for
return Table of the k_{rep} most-probable states with their representatives

Appendix E: **Supplementary Material for Chapter 6**

E.1 Software and Data

All Glide calculations were performed using Schrödinger Suite (2023-2 release). The protein protonation state is assigned with PROPKA33,34 at pH 7.4. DUD-E actives and decoys were processed using the LigPrep program of Schrödinger Suite starting from a SMILES representation of each molecule, using the following settings: (1) Tautomers were generated using Schrödinger's Epik with a target pH of 7.0 +/- 1.0 and (2) a maximum of 32 stereoisomers were generated for each molecule. The Posebusters (https://zenodo.org/records/8278563) dataset was used for molecular docking, and the DUD-E (http://dude.docking.org) dataset and AF2 database (https://alphafold.ebi.ac.uk) for virtual screening.

Bibliography

[1] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[2] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666, 1958.

[3] S. H. Kim, J. L. Sussman, F. L. Suddath, G. J. Quigley, A. McPherson, A. H. J. Wang, N. C. Seeman, and A. Rich. The general structure of transfer RNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 71(12):4970–4974, 1974.

[4] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6309–6313, 1980.

[5] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

[6] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 84(21):7524–7528, 1987.

[7] David J. Wales. Exploring energy landscapes. *Annual Review of Physical Chemistry*, 69:401–425, 2018.

[8] Andrea Pagnani, Giorgio Parisi, and Federico Ricci-Tersenghi. Glassy transition in a disordered model for the RNA secondary structure. *Physical Review Letters*, 84(9):2026–2029, 2000.

[9] Edward M. Purcell, Henry C. Torrey, and Robert V. Pound. Resonance absorption by nuclear magnetic moments in a solid. *Physical Review*, 69(1–2):37–38, 1946.

[10] Rahul Roy, Sungchul Hohng, and Taekjip Ha. A practical guide to single molecule FRET. *Nature Methods*, 5(6):507–516, 2008.

[11] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[12] Donald A. McQuarrie. *Statistical Mechanics*. Harper & Row, New York, 1976.

[13] Cyrus Levinthal. How to fold graciously. In P. Debrunner, J. C. M. Tsibris, and E. Münck, editors, *Mössbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello, Illinois*, pages 22–24. University of Illinois Press, Urbana, IL, 1969.

[14] Giorgio Parisi. On the replica approach to glasses. *arXiv preprint cond-mat/9701068*, 1997.

[15] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general AMBER force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[16] Kenno Vanommeslaeghe, Elisabeth Hatcher, Christopher Acharya, Somdeb Kundu, Shatakjyoti Zhong, Jörg Shim, Emmanuelle Darian, Olgun Guvench, Isabelle Lopes, and Alexander D. Jr. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, 2010.

[17] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[18] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[19] Michael P. Allen and Dominic J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 2 edition, 2017.

[20] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159:98–103, 1967.

[21] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[22] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[23] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.

[24] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[25] Scott Kirkpatrick, C. Daniel Gelatt, and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[26] E Marinari and G Parisi. Simulated tempering: A new monte carlo scheme. *Europhysics Letters (EPL)*, 19(6):451–458, July 1992.

[27] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

[28] Glenn M. Torrie and John P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[29] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Ann. Rev. Phys. Chem.*, 67:159–184, 2016.

[30] Cameron Abrams and Giovanni Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2014.

[31] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comp. Chem.*, 13(8):1011–1021, 1992.

[32] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12), 2008.

[33] Emilio Gallicchio, Michael Andrec, Anthony K Felts, and Ronald M Levy. Temperature reweighting and histogram reweighting methods for replica exchange simulations. *The Journal of Physical Chemistry B*, 109(14):6722–6731, 2005.

[34] Robert W Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22(8):1420–1426, 1954.

[35] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *J. Comp. Phy.*, 22(2):245–268, 1976.

[36] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, pages 163–185, 1998.

[37] Christopher Jarzynski. Targeted free energy perturbation. *Phys. Rev. E*, 65(4):046122, 2002.

[38] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *arXiv preprint arXiv:1605.08803*, 2017.

[39] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

[40] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2021.

[43] Philippe Brion and Eric Westhof. Hierarchy and dynamics of rna folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

[44] Leven M. Wadley, Kevin S. Keating, Carlos M. Duarte, and Anna Marie Pyle. Evaluating and learning from RNA pseudotorsional space: Quantitative validation of a reduced representation for RNA structure. *Journal of Molecular Biology*, 372(4):942–957, 2007.

[45] 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.

[46] Gregory R Bowman. Alphafold and protein folding: Not dead yet! the frontier is conformational ensembles. *Annual review of biomedical data science*, 7(1):51–57, 2024.

[47] Ryota Yamagami, Jacob P Sieg, and Philip C Bevilacqua. Functional roles of chelated magnesium ions in rna folding and function. *Biochemistry*, 60(31):2374–2386, 2021.

[48] Lin Huang and David MJ Lilley. Some general principles of riboswitch structure and interactions with small-molecule ligands. *Quarterly Reviews of Biophysics*, 58:e13, 2025.

[49] Sergio M. Villordo, Claudia V. Filomatori, Irma Sánchez-Vargas, Carol D. Blair, and Andrea V. Gamarnik. Dengue virus RNA structure specialization facilitates host adaptation. *PLOS Pathogens*, 11(1):e1004604, 2015.

[50] 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.

[51] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Charles Hillier, Francis Langston, Razvan Pascanu, Caglar Gulcehre, Sergio Gomez Colmenarejo, Suman Ravindranath, Tim Lillicrap, and Matthew Botvinick. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[52] Fabian B. Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[53] Salvatore Cuomo, Vincenzo Schiano di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics-informed neural networks: Where we are and what's next. *arXiv preprint arXiv:2201.05624*, 2022.

[54] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[55] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[56] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[57] Pratyush Tiwary, Lukas Herron, Richard John, Suemin Lee, Disha Sanwal, and Ruiyu Wang. Generative artificial intelligence for computational chemistry: a roadmap to predicting emergent phenomena. *arXiv preprint arXiv:2409.03118*, 2024.

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[59] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[60] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[61] Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, pages 1–10, 2024.

[62] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

[63] Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F Miller III, and Animashree Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6(2):195–208, 2024.

[64] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking, 2022.

[65] Yihang Wang, Lukas Herron, and Pratyush Tiwary. From data to noise to data for mixing physics across temperatures with generative artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.*, 119(32):e2203656119, 2022.

[66] Lukas Herron, Kinjal Mondal, John S Schneekloth Jr, and Pratyush Tiwary. Inferring phase transitions and critical exponents from limited observations with thermodynamic maps. *Proceedings of the National Academy of Sciences*, 121(52):e2321971121, 2024.

[67] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[68] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

[69] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022.

[70] Xenofon Karakonstantis, Efren Fernandez-Grande, and Peter Gerstoft. Efficient sound field reconstruction with conditional invertible neural networks, 2024.

[71] Sungwon Kim et al. P-flow: A fast and data-efficient zero-shot tts through speech prompting. *Advances in Neural Information Processing Systems*, 36, 2024.

[72] Xutao Guo, Yanwu Yang, Chenfei Ye, Shang Lu, Bo Peng, Hua Huang, Yang Xiang, and Ting Ma. Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2023.

[73] Shams Mehdi, Dedi Wang, Shashank Pant, and Pratyush Tiwary. Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck. *Journal of chemical theory and computation*, 18(5):3231–3238, 2022.

[74] Virgiliu Botan, Ellen HG Backus, Rolf Pfister, Alessandro Moretto, Marco Crisma, Claudio Toniolo, Phuong H Nguyen, Gerhard Stock, and Peter Hamm. Energy transport in peptide helices. *Proceedings of the National Academy of Sciences*, 104(31):12749–12754, 2007.

[75] Andrea Rizzi, Paolo Carloni, and Michele Parrinello. Targeted free energy perturbation revisited: Accurate free energies from mapped reference potentials. *The journal of physical chemistry letters*, 12(39):9449–9454, 2021.

[76] Michele Invernizzi, Andreas Kramer, Cecilia Clementi, and Frank Noé. Skipping the replica exchange ladder with normalizing flows. *The Journal of Physical Chemistry Letters*, 13(50):11643–11649, 2022.

[77] Aljoscha M Hahn and H Then. Using bijective maps to improve free-energy estimates. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 79(1):011113, 2009.

[78] Peter Wirnsberger, Andrew J Ballard, George Papamakarios, Stuart Abercrombie, Sebastien Racaniere, Alexander Pritzel, Danilo Jimenez Rezende, and Charles Blundell. Targeted free energy estimation via learned mappings. *The Journal of Chemical Physics*, 153(14), 2020.

[79] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[80] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[81] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[82] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[83] Hao Wu, Jonas Köhler, and Frank Noé. Stochastic normalizing flows, 2020.

[84] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Adv. Neurips.*, 31, 2018.

[85] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[86] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[87] Nicholas M Boffi and Eric Vanden-Eijnden. Probability flow solution of the fokker–planck equation. *Machine Learning: Science and Technology*, 4(3):035012, 2023.

[88] Aapo Hyvarinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[89] Erwin Schrodinger. Sur la theorie relativiste de lelectron et l'interpretation de la mecanique quantique. In *Annales de linstitut Henri Poincare*, volume 2, pages 269–310, 1932.

[90] Christian Leonard. A survey of the schr\" odinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[91] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrodinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

[92] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal transport in systems and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1):89–113, 2021.

[93] Hans Follmer. Random fields and diffusion processes. *Lect. Notes Math*, 1362:101–204, 1988.

[94] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[95] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[96] Michael S Albergo, Mark Goldstein, Nicholas M Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. *arXiv preprint arXiv:2310.03725*, 2023.

[97] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2(3), 2023.

[98] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*, 2023.

[99] Sebastian Buchenberg, Norbert Schaudinnus, and Gerhard Stock. Hierarchical biomolecular dynamics: Picosecond hydrogen bonding regulates microsecond conformational transitions. *Journal of Chemical Theory and Computation*, 11(3):1330–1336, February 2015.

[100] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[101] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[102] Keith J Laidler and M Catherine King. The development of transition-state theory. *The Journal of physical chemistry*, 87(15):2657–2664, 1983.

[103] Margaret L Scalley and David Baker. Protein folding kinetics exhibit an arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proceedings of the National Academy of Sciences*, 94(20):10636–10640, 1997.

[104] Hue Sun Chan and Ken A Dill. Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins: Structure, Function, and Bioinformatics*, 30(1):2–33, 1998.

[105] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[106] Ulrich HE Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, 1997.

[107] Robert Abel, Lingle Wang, Edward D Harder, BJ Berne, and Richard A Friesner. Advancing drug discovery through enhanced free energy calculations. *Accounts of chemical research*, 50(7):1625–1632, 2017.

[108] Pu Liu, Byungchan Kim, Richard A Friesner, and Bruce J Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences*, 102(39):13749–13754, 2005.

[109] Lingle Wang, Richard A Friesner, and Bruce 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.

[110] Andrew J Ballard and Christopher Jarzynski. Replica exchange with nonequilibrium switches. *Proceedings of the National Academy of Sciences*, 106(30):12224–12229, 2009.

[111] Simon Trebst, Matthias Troyer, and Ulrich HE Hansmann. Optimized parallel tempering simulations of proteins. *The Journal of chemical physics*, 124(17):174903, 2006.

[112] Walter Nadler and Ulrich HE Hansmann. Optimizing replica exchange moves for molecular dynamics. *Physical Review E*, 76(5):057102, 2007.

[113] Jaegil Kim, Thomas Keyes, and John E Straub. Generalized replica exchange method. *The Journal of chemical physics*, 132(22):224107, 2010.

[114] John D Chodera and Michael R Shirts. Replica exchange and expanded ensemble simulations as gibbs sampling: Simple improvements for enhanced mixing. *The Journal of chemical physics*, 135(19):194110, 2011.

[115] Alejandro Gil-Ley and Giovanni Bussi. Enhanced conformational sampling using replica exchange with collective-variable tempering. *Journal of chemical theory and computation*, 11(3):1077–1085, 2015.

[116] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[117] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 234–241. Springer, 2015.

[118] Yuxin Wu and Kaiming He. Group normalization. *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[119] Ashish Vaswani et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[120] 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.

[121] Jing Huang et al. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods*, 14(1):71–73, 2017.

[122] V Botan et al. Energy landscape of a helical peptide. *Proceedings of the National Academy of Sciences*, 104:12749–12754, 2007.

[123] Mithun Biswas, Benjamin Lickert, and Gerhard Stock. Metadynamics enhanced markov modeling of protein dynamics. *The Journal of Physical Chemistry B*, 122(21):5508–5514, 2018.

[124] Shams Mehdi, Dedi Wang, Shashank Pant, and Pratyush Tiwary. Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck. *Journal of Chemical Theory and Computation*, 18(5):3231–3238, 2022.

[125] Sylvie Bannwarth and Anne Gatignol. HIV-1 TAR RNA: The target of molecular interactions between the virus and its host. *Current HIV Research*, 3(1):61–71, 2005.

[126] Ursula Schulze-Gahmen et al. Insights into hiv-1 proviral transcription from integrative structure and dynamics of the tat:aff4:p-tefb:tar complex. *eLife*, 5:e15910, 2016.

[127] Laura R Ganser et al. High-performance virtual screening by targeting a high-resolution rna dynamic ensemble. *Nature structural & molecular biology*, 25:425–434, 2018.

[128] Jiří Šponer et al. Rna structural dynamics as captured by molecular simulations: A comprehensive overview. *Chemical reviews*, 118:4177–4338, 2018.

[129] Christina Bergonzo, Niel M. Henriksen, Daniel R. Roe, and Thomas E. III Cheatham. Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. *Journal of Chemical Theory and Computation*, 10(1):492–499, 2014.

[130] Manuel Dibak, Leon Klein, and Frank Noé. Temperature-steerable flows. *arXiv preprint arXiv:2012.00429*, 2020.

[131] Christopher Jarzynski. Rare events and the convergence of exponentially averaged work values. *Physical Review E*, 73(4):046105, 2006.

[132] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International conference on machine learning*, pages 1530–1538, 2015.

[133] Hao Wu, Jonas Köhler, and Frank Noé. Stochastic normalizing flows. *Advances in Neural Information Processing Systems*, 33:5933–5944, 2020.

[134] David J Wales. Exploring energy landscapes. *Ann. Rev. Phys. Chem.*, 69:401–425, 2018.

[135] Frederick James. Monte carlo theory and practice. *Rep. Prog. Phy.*, 43(9):1145, 1980.

[136] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond*, volume 9. World Scientific Publishing Company, 1987.

[137] Fiona M Jucker, Hans A Heus, Ping F Yip, Ellen HM Moors, and Arthur Pardi. A network of heterogeneous hydrogen bonds in gnra tetraloops. *J. Mol. Bio.*, 264(5):968–980, 1996.

[138] Fareed Aboul-ela, Jonathan Karn, and Gabriele Varani. Structure of hiv-1 tar rna in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nuc. Aci. Res.*, 24(20):3974–3981, 1996.

[139] Laura R Ganser, Chia-Chieh Chu, Hal P Bogerd, Megan L Kelly, Bryan R Cullen, and Hashim M Al-Hashimi. Probing rna conformational equilibria within the functional cellular context. *Cell Rep.*, 30(8):2472–2480, 2020.

[140] Richard Phillips Feynman. Slow electrons in a polar crystal. *Phys. Rev.*, 97(3):660, 1955.

[141] Peter Wirnsberger, George Papamakarios, Borja Ibarz, Sébastien Racanière, Andrew J Ballard, Alexander Pritzel, and Charles Blundell. Normalizing flows for atomic solids. *Mach. Lear. Sci. Tech.*, 3(2):025009, 2022.

[142] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Trans.*, 43(11):3964–3979, 2020.

[143] Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael R Andersen, Jonathan Huggins, and Aki Vehtari. Challenges and opportunities in high dimensional variational inference. *Adv. Neurips.*, 34:7787–7798, 2021.

[144] Jonas Köhler, Andreas Krämer, and Frank Noé. Smooth normalizing flows. *Adv. Neurips.*, 34:2796–2809, 2021.

[145] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Adv. Neurips.*, 32, 2019.

[146] Laurence Illing Midgley, Vincent Stimper, Gregor NC Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. *arXiv preprint arXiv:2208.01893*, 2022.

[147] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Adv. Neurips.*, 32, 2019.

[148] Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of fokker–planck equations through gradient–log–density estimation. *Entropy*, 22(8):802, July 2020.

[149] Hannes Risken and Hannes Risken. *Fokker-planck equation*. Springer, 1996.

[150] Brian DO Anderson. Reverse-time diffusion equation models. *Stoc. Proc. Appl.*, 12(3):313–326, 1982.

[151] Calvin Luo. Understanding diffusion models: A unified perspective, 2022.

[152] Ken Dill and Sarina Bromberg. *Molecular driving forces: statistical thermodynamics*. Garland Science, 2010.

[153] Somendra M Bhattacharjee and Avinash Khare. Fifty years of the exact solution of the two-dimensional ising model by onsager. *Curr. Sci.*, 69(10):816–821, 1995.

[154] John Cardy. *Scaling and renormalization in statistical physics*, volume 5. Cambridge university press, 1996.

[155] Ulli Wolff. Critical slowing down. *Nuclear Physics B-Proceedings Supplements*, 17:93–102, 1990.

[156] 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. *Proc. Natl. Acad. Sci. U.S.A.*, 115(7):E1346–E1355, 2018.

[157] Petra Kuhrova, Robert B Best, Sandro Bottaro, Giovanni Bussi, Jiri Sponer, Michal Otyepka, and Pavel Banas. Computer folding of rna tetraloops: identification of key force field deficiencies. *J. Chem. Theor. Comp.*, 12(9):4534–4548, 2016.

[158] Hsiao-Ping Hsu and Kurt Kremer. Glass transition temperature of (ultra-) thin polymer films. *J. Chem. Phys.*, 159(7), 2023.

[159] Andrea Pagnani, Giorgio Parisi, and Federico Ricci-Tersenghi. Glassy transition in a disordered model for the rna secondary structure. *Phys. Rev. Lett.*, 84(9):2026, 2000.

[160] Debayan Chakraborty, Rosana Collepardo-Guevara, and David J Wales. Energy landscapes, folding mechanisms, and kinetics of rna tetraloop hairpins. *J. Amer. Chem. Soc.*, 136(52):18052–18061, 2014.

[161] Aaron T Frank, Qi Zhang, Hashim M Al-Hashimi, and Ioan Andricioaei. Slowdown of interhelical motions induces a glass transition in rna. *Bioph. Jour.*, 108(12):2876–2885, 2015.

[162] Atreyee Banerjee, Aysenur Iscen, Kurt Kremer, and Oleksandra Kukharenko. Determining glass transition in all-atom acrylic polymeric melt simulations using machine learning. *J. Chem. Phys.*, 159(7), 2023.

[163] Janghyun Lee, Elizabeth A Dethoff, and Hashim M Al-Hashimi. Invisible rna state dynamically couples distant motifs. *Proc. Natl. Acad. Sci. U.S.A.*, 111(26):9485–9490, 2014.

[164] Andrew C Stelzer, Aaron T Frank, Jeremy D Kratz, Michael D Swanson, Marta J Gonzalez-Hernandez, Janghyun Lee, Ioan Andricioaei, David M Markovitz, and Hashim M Al-Hashimi. Discovery of selective bioactive small molecules by targeting an rna dynamic ensemble. *Nat. Chem. Bio.*, 7(8):553–559, 2011.

[165] Michael E Fisher and Walter Selke. Infinitely many commensurate phases in a simple ising model. *Phys. Rev. Lett.*, 44(23):1502, 1980.

[166] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.*, 79(8):2554–2558, 1982.

[167] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55(14):1530, 1985.

[168] Quentin Vicens and Jeffrey S Kieft. Thoughts on how to think (and talk) about rna structure. *Proc. Natl. Acad. Sci. U.S.A.*, 119(17):e2112677119, 2022.

[169] Rhiju Das and David Baker. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[170] Andreas Oxenfarth, Felix Kummerer, Sandro Bottaro, Robbin Schnieders, Gyorgy Pinter, Hendrik RA Jonker, Boris Furtig, Christian Richter, Martin Blackledge, Kresten Lindorff-Larsen, et al. Integrated nmr/molecular dynamics determination of the ensemble conformation of a thermodynamically stable cuug rna tetraloop. *J. Amer. Chem. Soc.*, 2023.

[171] Gul H Zerze, Pablo M Piaggi, and Pablo G Debenedetti. A computational study of rna tetraloop thermodynamics, including misfolded states. *The Journal of Physical Chemistry B*, 125(50):13685–13695, 2021.

[172] Justin P Sheehy, Amber R Davis, and Brent M Znosko. Thermodynamic characterization of naturally occurring rna tetraloops. *Rna*, 16(2):417–429, 2010.

[173] Sandro Bottaro, Francesco Di Palma, and Giovanni Bussi. The role of nucleobase interactions in rna structure and dynamics. *Nuc. Aci. Res.*, 42(21):13306–13314, 2014.

[174] Sai Shashank Chavali, Sachitanand M Mali, Jermaine L Jenkins, Rudi Fasan, and Joseph E Wedekind. Co-crystal structures of hiv tar rna bound to lab-evolved proteins show key roles for arginine relevant to the design of cyclic peptide tar inhibitors. *J. Bio. Chem.*, 295(49):16470–16486, 2020.

[175] Amy L Smith, Joseph Kassman, Khalid J Srour, and Ana Maria Soto. Effect of salt concentration on the conformation of tar rna and its association with aminoglycoside antibiotics. *Biochemistry*, 50(44):9434–9445, 2011.

[176] Shi-Jie Chen and Ken A Dill. Rna folding energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 97(2):646–651, 2000.

[177] Felix Ritort. Static chaos and scaling behavior in the spin-glass phase. *Phys. Rev. B*, 50(10):6844, 1994.

[178] Marco Baity-Jesi, Enrico Calore, Andrés Cruz, Luis Antonio Fernandez, José Miguel Gil-Narvion, Isidoro Gonzalez-Adalid Pemartin, Antonio Gordillo-Guerrero, David Iniguez, Andrea Maiorano, Enzo Marinari, et al. Temperature chaos is present in off-equilibrium spin-glass dynamics. *Commun. Phys.*, 4(1):74, 2021.

[179] Adam L Beberg, Daniel L Ensign, Guha Jayachandran, Siraj Khaliq, and Vijay S Pande. Folding@ home: Lessons from eight years of volunteer distributed computing. In *2009 IEEE International Symposium on Parallel & Distributed Processing*, pages 1–8. IEEE, 2009.

[180] Maxwell I Zimmerman and Gregory R Bowman. Fast conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theor. Comp.*, 11(12):5747–5757, 2015.

[181] Zahra Shamsi, Kevin J Cheng, and Diwakar Shukla. Reinforcement learning based adaptive sampling: Reaping rewards by exploring protein conformational landscapes. *J. Phys. Chem. B*, 122(35):8386–8395, 2018.

[182] Bruce J Berne, Michal Borkovec, and John E Straub. Classical and modern methods in reaction rate theory. *J. Phys. Chem.*, 92(13):3711–3725, 1988.

[183] Pratyush Tiwary and Michele Parrinello. From metadynamics to dynamics. *Phys. Rev. Lett.*, 111(23):230602, 2013.

[184] Sydney Brenner, François Jacob, and Matthew Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581, 1961.

[185] Yue Wan, Michael Kertesz, Robert C. Spitale, Eran Segal, and Howard Y. Chang. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655, 2011.

[186] Phei Er Saw and Erwei Song. Advancements in clinical rna therapeutics: Present developments and prospective outlooks. *Cell Reports Medicine*, 5(5), 2024.

[187] Jessica L Childs-Disney, Xueyi Yang, Quentin MR Gibaut, Yuquan Tong, Robert T Batey, and Matthew D Disney. Targeting rna structures with small molecules. *Nature Reviews Drug Discovery*, 21(10):736–762, 2022.

[188] Hasane Ratni, Martin Ebeling, John Baird, Stefanie Bendels, Johan Bylund, Karen S Chen, Nora Denk, Zhihua Feng, Luke Green, Melanie Guerard, et al. Discovery of risdiplam, a selective survival of motor neuron-2 (smn2) gene splicing modifier for the treatment of spinal muscular atrophy (sma), 2018.

[189] Norbert Pardi, Michael J Hogan, Frederick W Porter, and Drew Weissman. mrna vaccines—a new era in vaccinology. *Nature reviews Drug discovery*, 17(4):261–279, 2018.

[190] Lisa A Jackson, Evan J Anderson, Nadine G Rouphael, Paul C Roberts, Mamodikoe Makhene, Rhea N Coler, Michele P McCullough, James D Chappell, Mark R Denison, Laura J Stevens, et al. An mrna vaccine against sars-cov-2—preliminary report. *New England journal of medicine*, 383(20):1920–1931, 2020.

[191] Edoardo Morandi, Ilaria Manfredonia, Lisa M Simon, Francesca Anselmi, Martijn J van Hemert, Salvatore Oliviero, and Danny Incarnato. Genome-scale deconvolution of rna structure ensembles. *Nature Methods*, 18(3):249–252, 2021.

[192] Jienyu Ding, Yun-Tzai Lee, Yuba Bhandari, Charles D Schwieters, Lixin Fan, Ping Yu, Sergey G Tarosov, Jason R Stagno, Buyong Ma, Ruth Nussinov, et al. Visualizing rna conformational and architectural heterogeneity in solution. *Nature communications*, 14(1):714, 2023.

[193] Suemin Lee, Ruiyu Wang, Lukas Herron, and Pratyush Tiwary. Exponentially tilted thermodynamic maps (exptm): Predicting phase transitions across temperature, pressure, and chemical potential, 2025.

[194] Eric R. Beyerle and Pratyush Tiwary. Inferring the isotropic-nematic phase transition with generative machine learning. *Phys. Rev. Lett.*, pages –, Jul 2025.

[195] Hannah K Wayment-Steele, Wipapat Kladwang, Alexandra I Strom, Jeehyung Lee, Adrien Treuille, Alex Becka, Eterna Participants, and Rhiju Das. Rna secondary structure packages evaluated and improved by high-throughput experiments. *Nature methods*, 19(10):1234–1242, 2022.

[196] Liang Huang, He Zhang, Dezhong Deng, Kai Zhao, Kaibo Liu, David A Hendrix, and David H Mathews. Linearfold: linear-time approximate rna folding by 5'-to-3'dynamic programming and beam search. *Bioinformatics*, 35(14):i295–i304, 2019.

[197] Apoorv Malik, Liang Zhang, Milan Gautam, Ning Dai, Sizhen Li, He Zhang, David H Mathews, and Liang Huang. Linearalifold: linear-time consensus structure prediction for rna alignments. *Journal of Molecular Biology*, 436(17):168694, 2024.

[198] Andrew Martin Watkins, Ramya Rangan, and Rhiju Das. Farfar2: improved de novo rosetta prediction of complex global rna folds. *Structure*, 28(8):963–976, 2020.

[199] Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154(13), 2021.

[200] Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Masha Karelina, Rhiju Das, and Ron O Dror. Geometric deep learning of rna structure. *Science*, 373(6558):1047–1051, 2021.

[201] Rhiju Das, Rachael C Kretsch, Adam J Simpkin, Thomas Mulvaney, Phillip Pham, Ramya Rangan, Fan Bu, Ronan M Keegan, Maya Topf, Daniel J Rigden, et al. Assessment of three-dimensional rna structure prediction in casp15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1747–1770, 2023.

[202] Fan Bu, Yagoub Adam, Ryszard W Adamiak, Maciej Antczak, Belisa Rebeca H de Aquino, Nagendar Goud Badepally, Robert T Batey, Eugene F Baulin, Pawel Boinski, Michal J Boniecki, et al. Rna-puzzles round v: blind predictions of 23 rna structures. *Nature methods*, 22(2):399–411, 2025.

[203] Rachael C Kretsch, Alissa M Hummer, Shujun He, Rongqing Yuan, Jing Zhang, Thomas Karagianes, Qian Cong, Andriy Kryshtafovych, and Rhiju Das. Assessment of nucleic acid structure prediction in casp16. *bioRxiv*, pages 2025–05, 2025.

[204] Marc Parisien, José Almeida Cruz, Eric Westhof, and François Major. New metrics for comparing and assessing discrepancies between rna 3d structures and models. *Rna*, 15(10):1875–1885, 2009.

[205] L. Huang and D.M.J. Lilley. Crystal structure of the adp-binding domain of the nad+ riboswitch with nicotinamide adenine dinucleotide (nad+), September 2020.

[206] S. Schmidtke, E. Duchardt-Ferner, O. Ohlenschlaeger, D. Gottstein, and J. Wohnert. Solution nmr-structure of the neomycin sensing riboswitch rna bound to paromomycin, December 2015.

[207] I.R. Price and A. Ke. Manganese riboswitch from xanthmonas oryzae bound to mn(ii): 6n2v, November 2018.

[208] A. Ramesh and W.C. Winkler. Crystal structure of the nico transition-metal riboswitch bound to cobalt: 4rum, November 2014.

[209] S. Bachas and A.R. Ferre-D'amare. Crystal structure of the l.lactis ykoy riboswitch bound to cadmium: 6cb3, February 2018.

[210] L. Huang and D.M.J. Lilley. Structure of the glutamine ii riboswitch, June 2019.

[211] T. Numata, C.M. Connelly, J.S. Schneekloth, and A.R. Ferre-D'Amare. Crystal structure of a class i preq1 riboswitch complexed with preq1, April 2019.

[212] Jean-François Truchon and Christopher I Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling*, 47(2):488–508, 2007.

[213] A. Peselis and A. Serganov. Prpp riboswitch bound to prpp, manganese chloride soaked structure: 6dlq, June 2018.

[214] M.S. Dunstan. X-ray crystal structure of the tetrahydrofolate riboswitch aptamer bound to 5-deazatetrahydropterin: 6q57, December 2018.

[215] M.M. Matyjasik and R.T. Batey. 1.60 a resolution structure of the guanine riboswitch bound to guanine: 6ubu, September 2019.

[216] H.M. Wilt, Y.-X. Wang, and J.R. Stagno. Apo structure of the fmn riboswitch aptamer domain in the presence of sulfate: 6wjr, April 2020.

[217] J.C. Grigg, I.R. Price, and A. Ke. Sam-iii riboswitch on-state: 6c27, January 2018.

[218] C.W. Chan and A. Mondragon. Structure of the b. subtilis cobalamin riboswitch, June 2020.

[219] K.D. Warner, P. Homan, K.M. Weeks, A.G. Smith, C. Abell, and A.R. Ferre-D'Amare. Crystal structure of the e. coli thim riboswitch in complex with 5-(azidomethyl)-2-methylpyrimidin-4-amine, June 2014.

[220] Aiming Ren, Xin C Wang, Colleen A Kellenberger, Kanagalaghatta R Rajashankar, Roger A Jones, Ming C Hammond, and Dinshaw J Patel. Structural basis for molecular discrimination by a 3', 3'-cgamp sensing riboswitch. *Cell reports*, 11(1):1–12, 2015.

[221] K.D. Smith, C.A. Shanahan, E.L. Moore, A.C. Simon, and S.A. Strobel. Structure of a c-di-gmp-ii riboswitch from c. acetobutylicum bound to c-di-gmp: 3q3z, December 2010.

[222] Michael Sarver, Craig L Zirbel, Jesse Stombaugh, Ali Mokdad, and Neocles B Leontis. Fr3d: finding local and composite recurrent structural motifs in rna 3d structures. *Journal of mathematical biology*, 56(1):215–252, 2008.

[223] Christian Laing and Tamar Schlick. Analysis of four-way junctions in rna structures. *Journal of molecular biology*, 390(3):547–559, 2009.

[224] James Roll, Craig L Zirbel, Blake Sweeney, Anton I Petrov, and Neocles Leontis. Jar3d webserver: Scoring and aligning rna loop sequences to known 3d motifs. *Nucleic acids research*, 44(W1):W320–W327, 2016.

[225] Xiaojun Xu, Peinan Zhao, and Shi-Jie Chen. Vfold: a web server for rna structure and folding thermodynamics prediction. *PloS one*, 9(9):e107504, 2014.

[226] Xiaojun Xu and Shi-Jie Chen. Physics-based rna structure prediction. *Biophysics reports*, 1(1):2–13, 2015.

[227] Namhee Kim, Mai Zahran, and Tamar Schlick. Computational prediction of riboswitch tertiary structures including pseudoknots by ragtop: a hierarchical graph sampling approach. In *Methods in enzymology*, volume 553, pages 115–135. Elsevier, 2015.

[228] Peng Xiong, Ruibo Wu, Jian Zhan, and Yaoqi Zhou. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving rna model refinement. *Nature communications*, 12(1):2777, 2021.

[229] Jun Li and Shi-Jie Chen. Rnajp: enhanced rna 3d structure predictions with non-canonical interactions and global topology sampling. *Nucleic acids research*, 51(7):3341–3356, 2023.

[230] Jun Li, Sicheng Zhang, Dong Zhang, and Shi-Jie Chen. Vfold-pipeline: a web server for rna 3d structure prediction from sequences. *Bioinformatics*, 38(16):4042–4043, 2022.

[231] Andrew M Watkins, Caleb Geniesse, Wipapat Kladwang, Paul Zakrevsky, Luc Jaeger, and Rhiju Das. Blind prediction of noncanonical rna structure at atomic accuracy. *Science Advances*, 4(5):eaar5316, 2018.

[232] Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pages 2024–10, 2024.

[233] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11, 2025.

[234] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature methods*, 21(1):117–121, 2024.

[235] Amy C Anderson. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003.

[236] Brian K Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

[237] Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical reviews*, 9:91–102, 2017.

[238] Kathryn Klarich, Brian Goldman, Trevor Kramer, Patrick Riley, and W Patrick Walters. Thompson sampling– an efficient method for searching ultralarge synthesis on demand databases. *Journal of Chemical Information and Modeling*, 64(4):1158–1171, 2024.

[239] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[240] Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003.

[241] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

[242] William J Allen, Trent E Balius, Sudipto Mukherjee, Scott R Brozell, Demetri T Moustakas, P Therese Lang, David A Case, Irwin D Kuntz, and Robert C Rizzo. Dock 6: Impact of new features and current docking performance. *Journal of computational chemistry*, 36(15):1132–1156, 2015.

[243] Robert B Murphy, Matthew P Repasky, Jeremy R Greenwood, Ivan Tubert-Brohman, Steven Jerome, Ramakrishna Annabhimoju, Nicholas A Boyles, Christopher D Schmitz, Robert Abel, Ramy Farid, et al. Wscore: a flexible and accurate treatment of explicit water molecules in ligand–receptor docking. *Journal of medicinal chemistry*, 59(9):4364–4384, 2016.

[244] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

[245] David Errington, Constantin Schneider, Cédric Bouysset, and Frédéric A Dreyer. Assessing interaction recovery of predicted protein-ligand poses. *arXiv preprint arXiv:2409.20227*, 2024.

[246] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[247] T. Yoon, J. Y. Choi, S. Kwon, and E. K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference*, 2023. and Generative Modeling, 2023.

[248] Tom M Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University, 1980.

[249] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

[250] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

[251] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

[252] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *Advances in Neural Information Processing Systems*, 35:24240–24253, 2022.

[253] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

[254] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023.

[255] Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

[256] Gabriele Corso, Arthur Deng, Benjamin Fry, Nicholas Polizza, Regina Brazilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. *International Conference on Learning Representations*, 2024.

[257] Ruth Nussinov and Chung-Jung Tsai. The different ways through which specificity works in orthosteric and allosteric drugs. *Current pharmaceutical design*, 18(9):1311–1316, 2012.

[258] Xin Xie, Tingting Yu, Xiang Li, Nan Zhang, Leonard J Foster, Cheng Peng, Wei Huang, and Gu He. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. *Signal transduction and targeted therapy*, 8(1):335, 2023.

[259] Ruth Nussinov, Mingzhen Zhang, Yonglan Liu, and Hyunbum Jang. Alphafold, allosteric, and orthosteric drug discovery: Ways forward. *Drug discovery today*, 28(6):103551, 2023.

[260] Xiangyu Liu, Ali Masoudi, Alem W Kahsai, Li-Yin Huang, Biswaranjan Pani, Dean P Staus, Paul J Shim, Kunio Hirata, Rishabh K Simhal, Allison M Schwalb, et al. Mechanism of β2ar regulation by an intracellular positive allosteric modulator. *Science*, 364(6447):1283–1287, 2019.

[261] Shunsuke Imai, Tomoki Yokomizo, Yutaka Kofuku, Yutaro Shiraishi, Takumi Ueda, and Ichio Shimada. Structural equilibrium underlying ligand-dependent activation of β2-adrenoreceptor. *Nature Chemical Biology*, 16(4):430–439, 2020.

[262] Mariusz Butkiewicz, Edward W. Lowe, Ralf Mueller, Jeffrey L. Mendenhall, Pedro L. Teixeira, C. David Weaver, and Jens Meiler. Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules*, 18(1):735–756, 2013.

[263] Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

[264] P.C. Agu, C.A. Afiukwa, and O.U. Orji. Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Sci Rep*, 13, 2023.

[265] Yuqi Zhang, Marton Vass, Da Shi, Esam Abualrous, Jennifer M Chambers, Nikita Chopra, Christopher Higgs, Koushik Kasavajhala, Hubert Li, Prajwal Nandekar, et al. Benchmarking refined and unrefined alphafold2 structures for hit discovery. *Journal of Chemical Information and Modeling*, 63(6):1656–1667, 2023.

[266] David A Pearlman and Paul S Charifson. Improved scoring of ligand- protein interactions using owfeg free energy grids. *Journal of medicinal chemistry*, 44(4):502–511, 2001.

[267] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024.

[268] Yunrui Qiu, Richard John, Lukas Herron, and Pratyush Tiwary. Latent thermodynamic flows: unified representation learning and generative modeling of temperature-dependent behaviors from limited data. *Chemical Science*, 2026.

[269] Richard John, Yunrui Qiu, Lukas Herron, and Pratyush Tiwary. Combined representation and generation with diffusive state predictive information bottleneck. *arXiv preprint arXiv:2510.09784*, 2025.

[270] Prasenjit S. Ray, Jing Jia, Ping Yao, Munia Majumder, Maria Hatzoglou, and Paul L. Fox. A stress-responsive RNA switch regulates VEGFA expression. *Nature*, 457:915–919, 2009.

[271] Jessica A. Brown, David Bulkley, Jun Wang, Marissa L. Valenstein, Theresa A. Yario, Thomas A. Steitz, and Joan A. Steitz. Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nature Structural & Molecular Biology*, 21:633–640, 2014.

[272] Anja Donlic, Brandon S. Morgan, Jenny L. Xu, Annie Liu, Connor Jr. Roble, and Amanda E. Hargrove. Discovery of small molecule ligands for MALAT1 by tuning an RNA-binding scaffold. *Angewandte Chemie International Edition*, 57:13242–13247, 2018.

[273] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

[274] Mark James Abraham, Teemu Murtola, Roland Schulz, et al. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[275] 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.

[276] Christina Bergonzo, Niel M Henriksen, Daniel R Roe, and Thomas E Cheatham III. Multi-dimensional replica exchange molecular dynamics yields a converged ensemble of an rna tetranucleotide. *Journal of chemical theory and computation*, 10:492–499, 2014.

[277] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[278] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

[279] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[280] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[281] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[282] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[283] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comp. Bio.*, 13(7):e1005659, 2017.

[284] Jose LF Abascal and Carlos Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *J. Chem. Phys.*, 123(23), 2005.

[285] Lev Davidovich Landau and Evgenii Mikhailovich Lifshitz. *Statistical Physics: Volume 5*, volume 5. Elsevier, 2013.

[286] Sergio Bacallado, John D Chodera, and Vijay Pande. Bayesian comparison of markov models of molecular dynamics with detailed balance constraint. *J. Chem. Phys.*, 131(4), 2009.

[287] Kengo Sato and Michiaki Hamada. Recent trends in rna informatics: a review of machine learning and deep learning for rna secondary structure prediction and rna drug discovery. *Briefings in Bioinformatics*, 24(4):bbad186, 2023.

[288] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical rna structure. *Nature methods*, 7(4):291–294, 2010.

[289] Maria d'Errico, Elena Facco, Alessandro Laio, and Alex Rodriguez. Automatic topography of high-dimensional data sets by non-parametric density peak clustering. *Information Sciences*, 560:476–492, 2021.

[290] Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Dadapy: Distance-based analysis of data-manifolds in python. *Patterns*, 3(10), 2022.

[291] Peter Eastman, Raimondas Galvelis, Raúl P Peláez, Charlles RA Abreu, Stephen E Farr, Emilio Gallicchio, Anton Gorenko, Michael M Henry, Frank Hu, Jing Huang, et al. Openmm 8: molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*, 128(1):109–116, 2023.

[292] Da Teng, Vanessa J. Meraz, Akashnathan Aranganathan, Xinyu Gu, and Pratyush Tiwary. af2rave: protein ensemble generation with physics-based sampling. *Digital Discovery*, 4:2052–2061, 2025.

[293] Richard John, Lukas Herron, and Pratyush Tiwary. A comparison of probabilistic generative frameworks for molecular simulations. *The Journal of Chemical Physics*, 162(11), 2025.

[294] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of rna base pairs. *Rna*, 7(4):499–512, 2001.

[295] Tomasz Waleń, Grzegorz Chojnowski, Przemysław Gierski, and Janusz M Bujnicki. Clarna: a classifier of contacts in rna 3d structures based on a comparative analysis of various classification schemes. *Nucleic acids research*, 42(19):e151–e151, 2014.

[296] Marcin Magnus, Maciej Antczak, Tomasz Zok, Jakub Wiedemann, Piotr Lukasiak, Yang Cao, Janusz M Bujnicki, Eric Westhof, Marta Szachniuk, and Zhichao Miao. Rna-puzzles toolkit: a computational resource of rna 3d structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic acids research*, 48(2):576–588, 2020.