KinPFN: Bayesian Approximation of RNA Folding Kinetics using Prior-Data Fitted Networks

Dominik Scheuer[1,*], Frederic Runge[1,*], Jörg K.H. Franke[1], Michael T. Wolfinger[2,3], Christoph Flamm[2], Frank Hutter[1,4]

[1]University of Freiburg, Germany [2]University of Vienna, Austria [3]RNA Forecast e.U., Vienna, Austria [4]ELLIS Institute Tübingen, Germany [*]Equal contributions

dom.scheuer@gmail.com runget@cs.uni-freiburg.de

Summary

RNA folding kinetics describe the **probabilistic dynamics** of the RNA folding process.

RNA folding times allow to analyse the **folding efficiency** with applications in **synthetic biology** and candidate selection for **drug discovery**.

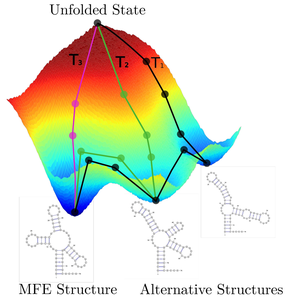
Problem: Current **RNA kinetics simulators** are **costly** and **scale exponentially** with the RNA length.

We present **KinPFN**, a novel approach for RNA folding kinetics based on **prior-data fitted networks (PFNs)** [1, 2].

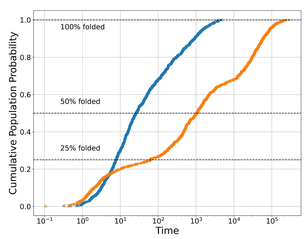
Trained on a **synthetic prior representing RNA folding times**, KinPFN achieves comparable results while **reducing simulator costs by ≥95%.**

Background: RNA Folding Kinetics

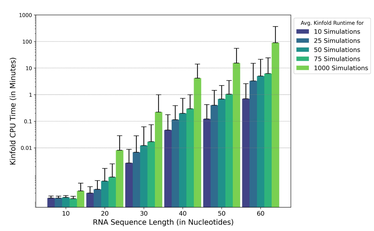
During the folding process, RNA traverses through a series of intermediate structural states, with each transition occurring at variable rates that collectively influence the time required to reach the functional form.



RNA Folding Times, the time required to fold into the structural form, allow to analyse the folding efficiency with applications in synthetic biology and drug discovery.

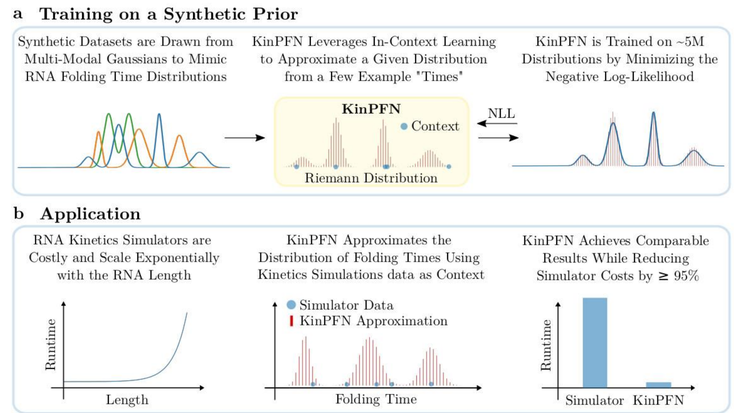


Problem: Existing RNA folding kinetics simulators are costly and scale exponentially with the RNA length, which makes them inapplicable to applications such as kinetic RNA design.



Benchmarking runtime using [3] on a single AMD Milan EPYC 7513 CPU with 2,6 GHz.

Our Approach: KinPFN



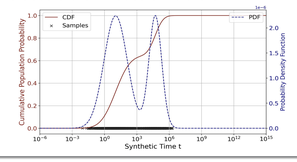
Synthetic RNA Folding Time Prior

Challenges:

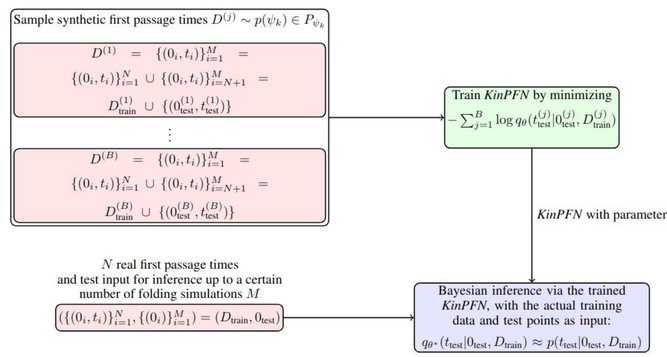
- RNA kinetics data is rare due to exponential costs of kinetic simulators.
- We have no access to the combination of RNA folding times and specific features like sequence or energy.

Approach:

- We train on synthetic datasets drawn from parameterized multi-modal Gaussians representing RNA folding time distributions.
- We leverage in-context learning at test time to accelerate kinetics simulators.



Training on the Synthetic Prior



From Synthetic to Real Data

Setup: Use Simulator Data as Context.

Data: 635 randomly generated RNA sequences with 1000 Simulations from [3].

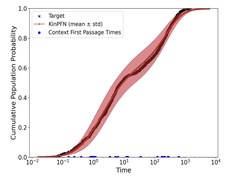
Results: Strong approximation performance of KinPFN across varying context sizes.

Method	First Passage Times N				
	10	**25**	**50**	**75**	**100**
KinPFN	**1.3739**	**1.2435**	**1.2047**	**1.1916**	**1.1858**
GMM_2	2.3122	1.3612	1.2355	1.2036	1.1933
GMM_3	5.2469	1.5830	1.2838	1.2132	1.1910
GMM_4	13.1325	1.9922	1.3676	1.2480	1.2119
GMM_5	37.5845	2.7708	1.4957	1.2953	1.2374
$DP\text{-}GMM_2$	1.6285	1.3529	1.2618	1.2305	1.2150
$DP\text{-}GMM_3$	1.6268	1.3549	1.2653	1.2323	1.2155
$DP\text{-}GMM_4$	1.6294	1.3558	1.2663	1.2337	1.2169
$DP\text{-}GMM_5$	1.6256	1.3572	1.2675	1.2337	1.2175
KDE	1.4370	1.2559	1.2133	1.2003	1.1957

Performance is Independent of RNA Features and Kinetics Simulators

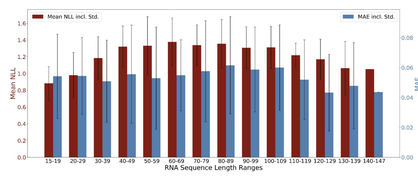
Setup: Use kinetics simulations from [4].

Results: KinPFN approximation is independent of the simulator.



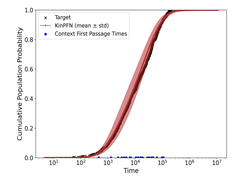
Setup: Analyze performance across RNA sequence lengths.

Results: KinPFN performance is constant across sequence lengths.



Setup: Analyze performance with different start and stop structures.

Results: KinPFN is independent of the start and stop structure.



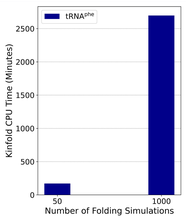
KinPFN is exclusively trained on synthetic folding times without knowledge about the underlying RNA features.

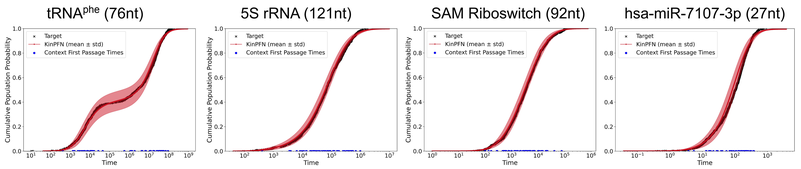
Case Study: Kinetics of Natural RNAs

Setup:
We use 50 context RNA folding times from 1,000 simulations of [3].

Data:
Four natural RNAs: $tRNA^{phe}$, 5S rRNA (both *S. cerevisiae*), SAM Riboswitch (*B. subtilis*), micro RNA (*H. sapiens*).

Results:
95% runtime improvement (~2 days → ~3 hours) with minimal accuracy loss.





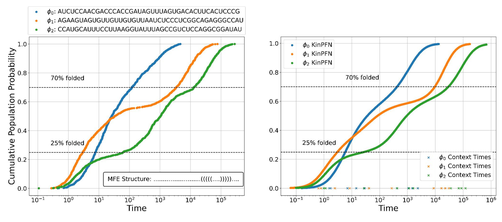
KinPFN requires only 5% of the compute!

Case Study: Folding Efficiency Analysis

Setup: Compare the folding efficiency of three 43nt RNAs (ϕ_0, ϕ_1, ϕ_2) with the same minimum free energy (MFE) structure.

Data: 10 context RNA folding times from 1,000 simulations of [3] for each RNA.

Results: 100× speed-up per RNA at comparable performance.



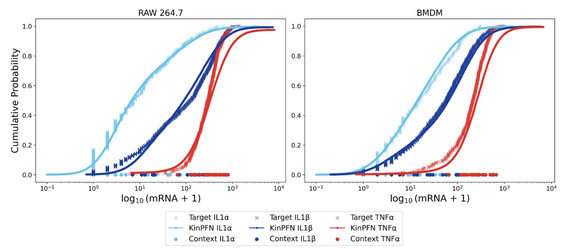
KinPFN requires only 1% of the compute per RNA!

Generalization to Gene Expression Data

Setup: Evaluate KinPFN generalization performance by using data from different biological context.

Data: smFISH counts from [5] for the expression of Interleukin-1 (IL-1α, IL-1β) and tumor necrosis factor alpha (TNF-α) mRNA in two immune cell lines, established RAW 264.7 macrophage cells and bone-marrow-derived macrophages (BMDM)) stimulated with Lipid A.

Results: KinPFN achieves accurate approximations of the gene expression using only 8% of the count data.



KinPFN requires only 8% of the data!

References

[1] Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., & Hutter, F. Transformers Can Do Bayesian Inference. In *International Conference on Learning Representations 2022*.

[2] Adriaensen, S., Rakotoarison, H., Müller, S., & Hutter, F. (2023). Efficient bayesian learning curve extrapolation using prior-data fitted networks. *Advances in Neural Information Processing Systems*, *36*, 19858-19886.

[3] Flamm, C., Fontana, W., Hofacker, I. L., & Schuster, P. (2000). RNA folding at elementary step resolution. *Rna*, *6*(3), 325-338.

[4] Dykeman, E. C. (2015). An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update. *Nucleic acids research*, *43*(12), 5708-5715.

[5] Bagnall, J., Rowe, W., Alachkar, N., Roberts, J., England, H., Clark, C., ... & Paszek, P. (2020). Gene-specific linear trends constrain transcriptional variability of the toll-like receptor signaling. *Cell Systems*, *11*(3), 300-314.

Paper

Code

Funded by the European Union