RNAGYM: BENCHMARKS FOR RNA FITNESS AND STRUCTURE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the structure and the effects of mutations in RNA are pivotal for numerous biological and medical applications. However, the evaluation of machine learning-based RNA models has been hampered by disparate and limited experimental datasets, along with inconsistent model performances across different RNA types. To address these limitations, we introduce RNAGym, a comprehensive and large-scale benchmark specifically tailored for RNA fitness and structure prediction. This benchmark suite includes over 30 standardized deep mutational scanning assays, covering hundreds of thousands of mutations, and curated RNA structure datasets. We have developed a robust evaluation framework that integrates multiple metrics suitable for both predictive tasks while accounting for the inherent limitations of experimental methods. RNAGym is designed to facilitate a systematic comparison of RNA models, offering an essential resource to enhance the development and understanding of these models within the computational biology community.

1 INTRODUCTION

RNA, once considered a passive intermediate between DNA and protein, is now recognized as a dynamic and crucial agent in cellular process regulation. The complexity of RNA lies in its secondary structure — the base-pairing patterns that form the backbone of its three-dimensional architecture — and in its functional versatility, which ranges from catalytic activities to gene regulation. Predicting RNA secondary structure and assessing the functional impact of sequence variations, ie. RNA fitness, are key challenges in computational biology and machine learning. These interrelated tasks are critical for advancing our understanding of RNA biology and its applications in fields such as drug discovery and synthetic biology.

The prediction of RNA secondary structure remains a significant challenge. While computational methods have made substantial progress, they still face considerable hurdles, especially for larger RNAs ($>$100 nt) with complex features like multi-branched loops and pseudoknots. Experimental methods like nuclear magnetic resonance, Cryo-EM, and X-ray crystallography can determine RNA 3D structures, but they face technical limitations when applied to RNA, resulting in RNA structures comprising less than 1% of entries in the Protein Data Bank (PDB) (Burley et al., 2017). This scarcity of experimental data further complicates the development and validation of computational prediction methods.

An equally critical task is predicting RNA fitness – the functional capacity of RNA sequences when subjected to mutations. Understanding the impact of these mutations on RNA function is crucial for advancing our knowledge of RNA evolution and its role in cellular processes. This task is also vital for the development of RNA-based therapeutics and the expansion of synthetic biology applications, such as designing riboswitches for gene regulation or engineering RNA sensors for metabolite detection. Despite its importance, accurately predicting the functional consequences of RNA mutations, especially from sequence data alone, remains a significant challenge in the field. Both structure and fitness prediction can benefit from evolutionary information. For structure prediction, approaches such as maximum entropy models leverage sequence co-variation to infer evolutionary constraints (Weinreb et al., 2016; Hopf et al., 2017; Frazer et al., 2021). Similarly, fitness prediction methods can utilize evolutionary data to identify functionally crucial sequence

features. However, robust methodologies for integrating this information and accurately predicting both structure and fitness, particularly in zero-shot scenarios, remain elusive.

To address these challenges and support progress in the field, we present RNAGym, a comprehensive benchmarking framework designed to evaluate and compare computational methods for RNA secondary structure and fitness prediction. RNAGym provides a diverse collection of curated RNA mutational scanning assays and chemical mapping data for structure prediction, multiple metrics for model evaluation, and assesses the relative performance of diverse baselines across both tasks.

RNAGym aims to accelerate progress in computational RNA biology by offering a common platform for assessing different approaches. By providing a systematic way to evaluate model performance across various RNA types and prediction tasks, RNAGym can help identify strengths and weaknesses of current methods, guide the development of more accurate algorithms, and ultimately contribute to advancing our understanding of RNA biology and its applications in areas such as personalized medicine, RNA-based drug design, and engineered RNA devices for synthetic biology.

2 Related Work and Background

2.1 Prior RNA benchmarks

Existing RNA benchmarks for variant effect prediction have been limited and fragmented, primarily focusing on testing individual models rather than serving as comprehensive benchmarking platforms. For instance, the RfamGen model was evaluated using five assays, including datasets on ribozymes and tRNAs (Sumi et al., 2024). Similarly, the Evo model was assessed using seven assays, incorporating ncRNA mutational scanning datasets (Nguyen et al., 2024). Both studies relied on overlapping but distinct datasets to evaluate their models, making it difficult to compare performance metrics between studies directly. These small benchmark sets restrict the ability to generalize findings and were primarily used to test the respective models' performance, rather than providing a broad and standardized benchmarking framework spanning the diversity of RNA types.

This limited scope stands in stark contrast to the field of protein research, where platforms like ProteinGym have been established to offer extensive and standardized benchmarking datasets (Notin et al., 2023). RNAGym addresses this gap by introducing a comprehensive benchmarking platform for RNA variant effect prediction that offers more than four times the number of assays compared to previous efforts, across a broader array of RNA classes including mRNAs, tRNAs, aptamers, and ribozymes.

With respect to 3D RNA structure prediction, several competitive benchmarks have been developed, including the Critical Assessment of Structure Prediction (CASP) and RNA-Puzzles. CASP15 (Kryshtafovych et al., 2023), the latest iteration of CASP, introduced a dedicated category for RNA structure prediction, reflecting the growing recognition of RNA's importance and the need for accurate computational models. RNA-Puzzles (Cruz et al., 2012), on the other hand, is a community-driven initiative that presents real-world challenges to participants, who submit their models to be evaluated against experimentally determined RNA structures. Notably, only a few RNA molecules are evaluated at CASP and through RNA-Puzzles, limiting the ability of these high quality datasets to act as benchmarking standards.

Train-test splits are commonly used to evaluate 2D RNA structure prediction models (Penić et al., 2024; Chen et al., 2022). Intra-family splits involve training models on RNA sequences from the same family, with sequences from these families appearing in both the training and test sets (Singha et al., 2019). This approach tests a model's ability to learn and predict structures within known families. In contrast, inter-family splits ensure that sequences from the same RNA family in the test set are excluded from the training set (Penić et al., 2024). This method assesses whether a model can generalize to entirely new RNA families that were not included in the training data. While existing benchmarks offer valuable insights, they often lack the scale and diversity to comprehensively evaluate model performance across various RNA types and structures. Zero-shot benchmarks for secondary structure prediction models are crucial, as they avoid biases inherent in supervised approaches and potential overfitting, thus providing a more robust assessment of true generalization capabilities.

2.2 BACKGROUND: THE DIVERSITY OF RNA MOLECULES

RNA molecules exhibit a remarkable range of structures and functions, highlighting their essential role in both the fundamental processes of biology and their growing utility in medical and biotechnological applications. From the synthesis and regulation of proteins to the catalysis of key biochemical reactions, RNA types such as mRNA, tRNA, ribozymes, and aptamers demonstrate the diversity of RNA molecular diversity and complexity. **mRNAs (Messenger RNAs)** act as the intermediary transcript that carry genetic information from DNA to the ribosome, where they serve as a template for protein synthesis. The sequence of mRNA dictates the amino acid sequence in a protein, thereby directly influencing gene expression and regulatory mechanisms. **tRNAs (Transfer RNAs)** play a crucial role in translation, the process of protein synthesis in the ribosome. Each tRNA molecule transports a specific amino acid to the ribosome; its anticodon loop pairs with the corresponding codon in the mRNA, ensuring that the correct amino acid is incorporated into the growing protein chain. **Ribozymes** are catalytic RNA molecules that perform specific biochemical reactions akin to protein enzymes. These include critical activities such as RNA splicing during gene expression, where ribozymes help in the excision of introns from a pre-mRNA. **Aptamers** are short, single-stranded RNA molecules designed to bind with high specificity and affinity to certain targets, including proteins, small molecules, and various cellular components. Their high specificity and adaptability make aptamers highly valuable for therapeutic uses, as well as in diagnostic and biosensing applications.

3 RNAGYM BENCHMARKS

3.1 OVERVIEW

RNAGym is a comprehensive benchmark suite advancing the development and analysis of machine learning RNA models. It comprises three integrated layers: datasets, models, and analytics (Fig. 1), supporting two core RNA tasks:

- **Fitness prediction:** Zero-shot prediction of RNA functionality across diverse RNA types, leveraging a broad set of deep mutational scanning assays.
- **Structure prediction:** Zero-shot prediction of RNA secondary structure, focusing on identifying nucleotide contacts, which is crucial for understanding RNA function.

These tasks, evaluated in a zero-shot setting, challenge models to generalize across varied RNA contexts without task-specific fine-tuning. Our data layer includes curated datasets that are specifically structured for these two tasks. These datasets are enriched with detailed annotations for a variety of RNA types and are classified by mutation depth, enhancing the granularity of the data available for analysis. Across both tasks, RNAGym integrates a diverse array of 10 predictive models, each tailored to address the nuances of the specific tasks at hand—whether predicting RNA fitness or determining RNA structure. The analytics layer of RNAGym is designed to provide a deep and comprehensive evaluation of model performance. It utilizes five distinct performance metrics to assess the effectiveness of each model in a clear and quantifiable manner. Further, the framework allows for detailed exploration of model performance across different RNA types and mutation depths, with the goal to understand model strengths and limitations in varied biological contexts.

3.2 DATASETS

3.2.1 FITNESS PREDICTION ASSAYS

Screening methodology We conducted a broad PubMed search for RNA mutational studies that yielded over 11,000 results, which we then screened using a LLM with carefully designed prompts adapted from systematic review methods. After narrowing down to 52 studies through the LLM screening, we conducted expert manual review using specific inclusion/exclusion criteria to ensure data quality and relevance. All details regarding search terms, prompts and inclusions/exclusions criteria are provided in Appendix B.

Selected assays RNAGym includes 31 Deep Mutational Scanning assays containing over 350,000 variants across various mRNA, tRNA, aptamers, and ribozymes (Table 1). This represents a *fourfold increase* in size over the largest prior RNA benchmarks. Notably, eleven of these assays had MSAs

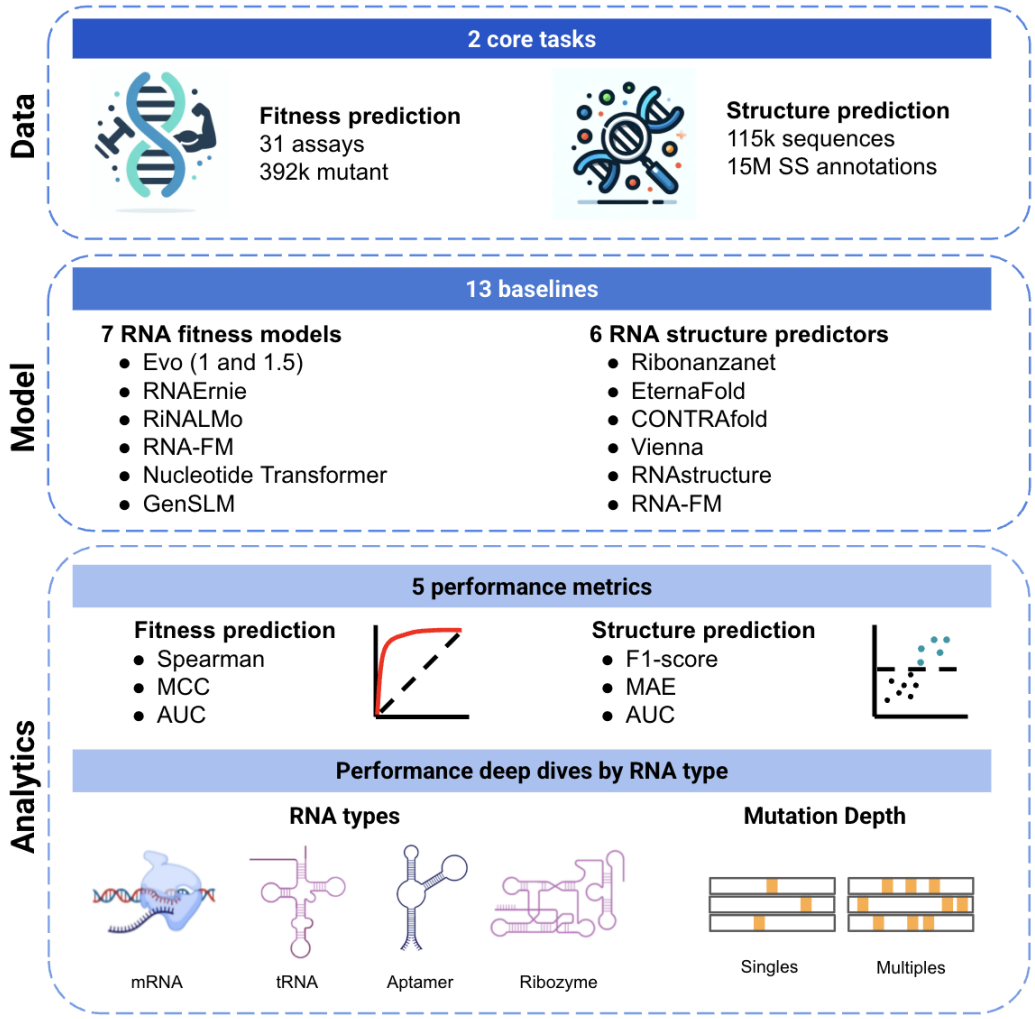

Figure 1: **RNAGym benchmarks.** RNAGym is a comprehensive RNA analysis framework designed specifically for fitness and structure prediction tasks. It evaluates the performance of diverse baselines across these tasks, and offers in-depth assessments by RNA type and mutation depth.

readily available through the RFAM database, while 20 of these assays were synthetic constructs or had no MSA available. Unlike previous efforts, these assays are integrated into a standardized, reusable resource, making RNAGym a more accessible and broadly applicable tool for RNA fitness prediction. Our final processed datasets all have a consistent format with the same 3 fields across: "Mutant" (mutation triplets), "Sequence" (mutated sequence), and "DMS score" (experimental measurement). We also corrected the *directionality* of each measured experimental phenotype, to ensure that higher DMS scores always translate to higher fitness across assays.

3.2.2 Structure prediction data

In preparing the benchmark for our research paper, we utilized the dataset from the Stanford Ribonanza Challenge, which contains DMS (dimethyl sulfate) assay data. DMS data is critical for understanding RNA secondary structures as it selectively methylates unpaired adenine and cytosine bases, thereby providing a chemical footprinting method to infer RNA folding. This type of data is invaluable for validating computational models of RNA secondary structure prediction, as it offers direct evidence of the RNA's physical structure under in vivo-like conditions. The private test set derived from the Ribonanza Challenge incorporates a diverse array of RNA sequences, curated to represent a broad spectrum of RNA types and complexities. To focus on relevant data for our

RNA Type	Description	# Assays	# Singles	# Multiples
Messenger RNA (mRNA)	Splicing ability	2	0.3k	22k
Transfer RNA (tRNA)	Stability and growth	3	0.4k	70k
Aptamer	Target binding ability	7	0.4k	40k
Ribozyme	Splicing ability	19	1.7k	226k
Total		31	2.9k	358k

Table 1: **RNAGym fitness benchmark summary.** RNAGym includes a large collection of DMS assays about diverse RNA types. The table reports the number of assays and number of single and multiple mutants per RNA type.

zero-shot analysis (see Appendix B), the dataset was refined to include approximately $115k$ distinct sequences, covering over $15M$ nucleotide positions for structural predictions. The evaluation dataset provides a DMS score for each nucleotide for all RNA sequences (1 row per nucleotide), reflecting the propensity of that nucleotide to *not* be in contact in the RNA structure.

3.3 Baselines

We benchmarked several RNA models including RiNALMo, Evo 1 and 1.5, RNA-FM, GenSLM, RNAErnie and Nucleotide Transformer for fitness prediction, as well as EternaFold, CONTRAfold, Vienna, RNAstructure, RNA-FM, and Ribonanzanet for structure prediction. All details about baselines are provided in Appendix C.

3.4 Evaluation

For **fitness prediction**, the evaluation was primarily based on the Spearman's rank correlation between the model predictions and experimental measurements, the Area Under the Curve (AUC) and the Matthews Correlation Coefficient (MCC). These metrics are complementary and were chosen to provide a comprehensive evaluation: Spearman correlation assesses the overall ranking of predictions, AUC measures the model's ability to distinguish between functional and non-functional mutations, while MCC offers a balanced measure for potentially imbalanced datasets. To mitigate biases associated with uneven assay distributions across different RNA types, we calculated an average performance for each RNA type separately and then computed the overall performance as the mean of these RNA-type-level averages. This approach ensures that our results are robust and reflective of true model capabilities across varied biological categories.

For the **structure prediction task**, we employed three standard metrics: F1-score, Area Under the Curve (AUC), and Mean Absolute Error (MAE). F1-score provides a balanced measure of precision and recall in identifying nucleotide pairings. AUC assesses the model's ability to distinguish between paired and unpaired nucleotides. MAE offers a direct measure of prediction accuracy by quantifying the average magnitude of errors.

Importantly, all evaluations for both tasks were conducted in a zero-shot setting, where models were tested without any fine-tuning on task-specific labels, emphasizing their generalizability and robustness in unseen scenarios.

4 Results

4.1 Fitness prediction performance

Performance on all assays. The overall fitness prediction benchmark results (see Table 2) show Evo (1.5) and RNAErnie as the leading performers, with nearly identical top metrics - Spearman correlations of 0.222 and 0.221 respectively, and comparable AUC and MCC. These results indicate a modest yet leading capability in predicting RNA fitness outcomes based on experimental data (statistical significance analysis is included in Appendix G.2). The relatively low scores across all models, particularly when compared to the stronger correlations reported for protein language models Notin et al. (2023), suggest substantial room for improvement and warrant deeper investigation.

Several factors may contribute to this performance gap. A primary consideration is the limited availability of large-scale, diverse RNA datasets for model training compared to the abundance of protein sequence data. Additionally, there may be a potential misalignment between the training data used for these models and the taxonomical and functional distribution of our fitness landscapes. Lastly, differences in evolutionary conservation patterns between RNAs (especially non-coding RNAs) and proteins could also play a role, potentially affecting the models' ability to capture fitness-relevant features.

Rank	Model name	Spearman	AUC	MCC
1	Evo 1.5	**0.222**	0.606	**0.163**
2	RNAErnie	0.221	**0.609**	**0.163**
3	Evo 1	0.220	0.606	0.162
4	RNA-FM	0.205	0.598	0.150
5	RiNALMo	0.170	0.583	0.121
6	Nucl. Transformer	0.157	0.582	0.117
7	GenSLM	0.118	0.558	0.082

Table 2: **RNAGym - Fitness prediction overall benchmark.** Average of Spearman's rank correlation, AUC and MCC between model scores and experimental measurements on the full RNAGym fitness prediction benchmark.

Performance by RNA type. When examining performance by RNA type (Table 3), several models show specialized strengths across different RNA categories. RNA-FM achieves the highest correlation for tRNAs (0.463), while RiNALMo leads in mRNA predictions (0.273), followed closely by Nucleotide Transformer (0.245). For aptamers, Evo 1 demonstrates the strongest performance (0.200), while Evo 1.5 performs best with ribozymes (0.170). These performance variations likely reflect the diverse training data of each model. RNA-FM's particular strength with tRNAs aligns with its training on non-coding RNAs from RNAcentral, a database rich in these RNA types. RiNALMo and Nucleotide Transformer's proficiency in mRNA predictions may stem from their exposure to coding sequences during training. The consistently strong performance of Evo models across different RNA types, particularly in aptamers and ribozymes, suggests their training approach may capture broader sequence-function relationships. These observations underscore the importance of targeted model training and selection based on the specific RNA type being studied. They also suggest that performance could potentially be improved by more tailored training data selection or by developing ensemble methods that leverage the strengths of different models for specific RNA types.

Rank	Model name	mRNA	tRNA	Aptamer	Ribozyme	All
1	Evo 1.5	0.16	0.376	0.181	**0.170**	**0.222**
2	RNAErnie	0.207	0.396	0.153	0.127	0.221
3	Evo 1	0.142	0.377	**0.2**	0.16	0.22
4	RNA-FM	0.08	**0.463**	0.131	0.148	0.205
5	RiNALMo	**0.273**	0.258	0.057	0.093	0.17
6	Nucl. Transformer	0.245	0.093	0.178	0.111	0.157
7	GenSLM	0.123	0.116	0.113	0.119	0.118
-	All	0.129	0.265	0.114	0.126	0.137

Table 3: **RNAGym - Fitness prediction by RNA type.** Average of Spearman's rank correlation between model scores and experimental measurements by RNA type and overall.

Performance by mutation type. The fitness prediction results segmented by mutation type (Table 4) show that Evo models consistently outperform others across both single and multiple mutations. Evo 1.5 achieves the highest correlations for both single mutations (0.240) and multiple mutations (0.195), with Evo 1 following closely behind (0.239 and 0.188 respectively). While these results establish Evo models as the current leaders in mutation effect prediction, the relatively low correlation values indicate substantial room for improvement in capturing RNA sequence-function relationships. Notably, all models show somewhat stronger performance on single mutations compared to multiple

mutations, highlighting the increased challenge of predicting fitness effects for more complex genetic variations. This performance gap between single and multiple mutations points to opportunities for improving model architectures and training approaches to better handle combinatorial effects of mutations.

Rank	Model name	Singles	Multiples
1	Evo 1.5	**0.240**	**0.195**
2	RNAErnie	0.178	0.170
3	Evo 1	0.239	0.188
4	RNA-FM	0.193	0.171
5	RiNALMo	0.144	0.119
6	Nucl. Transformer	0.136	0.131
7	GenSLM	0.137	0.120

Table 4: **RNAGym - Fitness prediction by mutation type.** Average of Spearman's rank correlation between model scores and experimental measurements by mutation type.

Future Directions. To advance the field of RNA fitness prediction, several promising avenues of investigation emerge. First, developing models specifically trained on diverse RNA fitness landscapes could potentially improve performance by more closely aligning the training data with the prediction task. Additionally, incorporating RNA secondary structure predictions or experimental structure data into fitness prediction models may enhance their accuracy by capturing the important relationship between RNA structure and function. Comparing zero-shot performance with fine-tuned models could provide valuable insights into the generalizability of learned RNA features, potentially guiding future model development strategies. Lastly, exploring new architectural elements or pre-training objectives that better capture RNA-specific properties might lead to more robust and accurate predictions.

4.2 Structure prediction performance

Rank	Model name	F1-score ↑	AUC ↑	MAE ↓
1	Ribonanzanet	**0.793**	**0.869**	**0.146**
2	CONTRAfold	0.610	0.650	0.372
3	RNAstructure	0.603	0.639	0.326
4	Vienna	0.602	0.638	0.324
5	EternaFold	0.600	0.640	0.360
6	RNA-FM	0.561	0.584	0.278

Table 5: **RNAGym - Structure prediction benchmark.** F1-score, AUC and MAE between model predictions and experimental DMS measurements on the RNAGym structure prediction benchmark.

Performance. The RNAGym structure prediction benchmark (Table 5) reveals interesting performance patterns across supervised and unsupervised approaches. The supervised model Ribonanzanet demonstrates superior performance, achieving an F1-score of 0.793, AUC of 0.869, and MAE of 0.146, significantly outperforming unsupervised methods. Among unsupervised models, we observe a remarkably tight performance distribution. CONTRAfold shows a slight edge with an F1-score of 0.610, followed closely by RNAstructure (0.603), Vienna (0.602), and EternaFold (0.600). This close grouping is maintained across all three evaluation metrics, suggesting that current unsupervised approaches may be approaching a performance ceiling within their current methodological framework.

Future Directions. Advancing RNA secondary structure prediction requires addressing several key challenges. First, the substantial performance gap between Ribonanzanet and unsupervised methods raises concerns about potential data leakage between training and test sets. A rigorous analysis of sequence redundancy in the Ribonanza dataset is necessary to ensure fair evaluation. Current performance metrics, while encouraging, may not fully capture a model's ability to generalize to

truly novel RNA sequences. Future work should enforce stricter evaluation protocols that control for sequence similarity between training and test sets. Additionally, extending these approaches to tertiary structure prediction remains an important but significantly more complex challenge.

5 RESOURCES

Codebase. We open source under an MIT License all resources curated for the RNAGym benchmark via our GitHub repository. In particular, we consolidate of the numerous RNA structure and fitness prediction models discussed in Appendix C and make them available via a common interface, which will facilitate the seamless integration and evaluation of new models as they are developed. This resource aims to provide researchers with robust tools, reducing the technical barrier to entry for conducting advanced RNA analysis and enhancing the reproducibility of results across the scientific community.

Processed datasets. We have made available all datasets used in our fitness and structure prediction benchmarks, including both raw and processed versions, as detailed in Section 3.2. Our GitHub repository provides instructions for downloading these resources. To enhance the utility of our benchmarks, we have included several additional components. For the fitness benchmark, where available, we provide tertiary structure PDB files and multiple sequence alignments for the relevant protein families. In the case of the secondary structure prediction benchmark, we have mapped all sequences in the test set to similar RNA sequences found in the PDB, RFam, and PseudoBase databases, providing easy access to the rich annotations contained in these databases. Furthermore, to support researchers interested in supervised learning approaches, we offer training datasets for both the fitness and secondary structure prediction tasks (Appendix B).

6 CONCLUSION

RNAGym addresses the significant gap in large-scale benchmarks for the robust evaluation of models tailored for RNA structure prediction and fitness assessment. It enables the direct comparison of methods across several dimensions of interest (e.g., RNA type, mutation type). We anticipate that the RNAGym benchmarks and the accompanying data assets we release to the public will serve as invaluable resources for the Machine Learning and Computational Biology communities. We plan to continually update the benchmarks as new data and baseline models become available.

REFERENCES

Johan O. L. Andreasson, Andrew Savinov, Steven M. Block, and William J. Greenleaf. Comprehensive sequence-to-function mapping of cofactor-dependent rna catalysis in the glms ribozyme. *Nature Communications*, 2020. doi: https://doi.org/10.1038/s41467-020-15540-1. URL https://www.nature.com/articles/s41467-020-15540-1#citeas.

David Baker and George Church. Protein design meets biosecurity. *Science*, 383:349 – 349, 2024. URL https://api.semanticscholar.org/CorpusID:267212249.

James D Beck, Jessica M Roberts, Joey M Kitzhaber, Ashlyn Trapp, Edoardo Serra, Francesca Spezzano, and Eric J Hayden. Predicting higher-order mutational effects in an rna enzyme by machine learning of high-throughput experimental data. *Frontiers Mol. Biosci.*, 2022. doi: 10.3389/fmolb.2022.893864. URL https://www.frontiersin.org/articles/10.3389/fmolb.2022.893864/full.

Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): the single global macromolecular structure archive. *Protein Crystallography*, pp. 627–641, 2017.

Christian Cao, Jason Sang, Rohit Arora, Robert Kloosterman, Matthew Cecere, Jaswanth Gorla, Richard Saleh, David Chen, Ian Drennan, Bijan Teja, Michael Fehlings, Paul Ronksley, Alexander A Leung, Dany Weisz, Harriet Ware, Mairead Whelan, David B Emerson, Rahul Krishan Arora, and Niklas Bobrovitz. Prompting is all you need: Llms for systematic review screening. *medRxiv*, 2024. doi: 10.1101/2024.06.01.24308323.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions, 2022.

José Almeida Cruz, Marc-Frédérick Blanchet, Michal J. Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay V. Dokholyan, Samuel Coulbourn Flores, Lili Huang, Christopher A. Lavender, Véronique Lisi, François Major, Katarzyna Mikolajczak, Dinshaw J. Patel, Anna Philips, Tomasz Puton, John SantaLucia, Fredrick Sijenyi, Thomas Hermann, Kristian Rother, Magdalena Rother, Alexander Serganov, Marcin Skorupski, Tomasz Soltysinski, Parin Sripakdeevong, Irina Tuszynska, Kevin M. Weeks, Christina Waldsich, Michael Wildauer, Neocles B. Leontis, and Eric Westhof. Rna-puzzles: a casp-like evaluation of rna three-dimensional structure prediction. *RNA*, 18 4:610–25, 2012. URL https://api.semanticscholar.org/CorpusID:263498187.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679. URL https://www.biorxiv.org/content/early/2023/09/19/2023.01.11.523679.

Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22 14:e90–8, 2006. URL https://api.semanticscholar.org/CorpusID:1646946.

Júlia Domingo, Guillaume Diss, and Ben Lehner. Pairwise and higher-order genetic interactions during the evolution of a trna. *Nature*, 2018. doi: 10.1038/s41586-018-0170-7. URL https://api.semanticscholar.org/CorpusID:240071819.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan N. Gomez, Joseph K Min, Kelly P. Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599: 91 – 95, 2021. URL https://api.semanticscholar.org/CorpusID:240071819.

Andreas R. Gruber, Ronny Lorenz, Stephan H. Bernhart, Richard Neuböck, and Ivo L. Hofacker. The vienna rna websuite. *Nucleic Acids Research*, 36:W70 – W74, 2008. URL https://api.semanticscholar.org/CorpusID:6481000.

Michael P. Guy, David L. Young, Matthew J. Payea, Xiaoju Zhang, Yoshiko Kon, Kimberly M. Dean, Elizabeth J. Grayhack, David H. Mathews, Stanley Fields, and Eric M. Phizicky. Identification of the determinants of trna function and susceptibility to rapid trna decay by high-throughput in vivo analysis. *Genes and Development*, 2014. doi: 10.1101/gad.245936.114. URL https://genesdev.cshlp.org/content/28/15/1721.long.

Shujun He, Rui Huang, Jill Townley, Rachael C. Kretsch, Thomas G. Karagianes, David B.T. Cox, Hamish Blair, Dmitry Penzar, Valeriy Vyaltsev, Elizaveta Aristova, Arsenii Zinkevich, Artemy Bakulin, Hoyeol Sohn, Daniel Krstevski, Takaaki Fukui, Fumiya Tatematsu, Yusuke Uchida, Donghoon Jang, Jun Seong Lee, Roger Shieh, Tom Ma, Eduard Martynov, Maxim V. Shugaev, Habib S.T. Bukhari, Kazuki Fujikawa, Kazuki Onodera, Christof Henkel, Shlomo Ron, Jonathan Romano, John J. Nicol, Grace P. Nye, Yuan Wu, Christian A. Choe, Walter Reade, and Rhiju Das. Ribonanza: deep learning of rna structure through dual crowdsourcing. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:268064302.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2): 128–135, 2017.

Evan Janzen, Yuning Shen, Alberto Vázquez-Salazar, Ziwei Liu, Celia Blanco, Josh Kenchel, and Irene A. Chen. Emergent properties as by-products of prebiotic evolution of aminoacylation ribozymes. *Nature Communications*, 2022. doi: 10.1038/s41467-022-31387-0. URL https://www.nature.com/articles/s41467-022-31387-0.

Philippe Julien, Belén Miñana, Pablo Baeza-Centurion, Juan Valcárcel, and Ben Lehner. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nature communications*, 2016. doi: https://doi.org/10.1038/ncomms11558. URL https://www.nature.com/articles/ncomms11558.

Shengdong Ke, Vincent Anquetil, Jorge Rojas Zamalloa, Alisha Maity, Anthony Yang, Mauricio A. Arias, Sergey Kalachikov, James J. Russo, and Jingyue Juand Lawrence A. Chasin. Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Research*, 2017. doi: 10.1101/gr.219683.116. URL https://genome.cshlp.org/content/28/1/11.long.

Shungo Kobori, Yoko Nomura, Anh Miu, and Yohei Yokobayashi. High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Research*, 43(13):e85–e85, 03 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv265. URL https://doi.org/10.1093/nar/gkv265.

Shungo Kobori, Kei Takahashi, and Yohei Yokobayashi. Deep sequencing analysis of aptazyme variants based on a pistol ribozyme. *ACS Synthetic Biology*, 6(7):1283–1288, 2017. doi: 10.1021/acssynbio.7b00057. URL https://doi.org/10.1021/acssynbio.7b00057. PMID: 28398719.

Andriy Kryshtafovych, Maciej Antczak, Marta Szachniuk, Tomasz Zok, Rachael C. Kretsch, Ramya Rangan, Phillip Pham, Rhiju Das, Xavier Robin, Gabriel Studer, Janani Durairaj, Jerome Eberhardt, Aaron Sweeney, Maya Topf, Torsten Schwede, Krzysztof Fidelis, and John Moult. New prediction categories in casp15. *Proteins*, 91:1550 – 1557, 2023. URL https://api.semanticscholar.org/CorpusID:259138147.

Chuan Li, Wenfeng Qian, Calum J. Maclean, and Jianzhi Zhang. The fitness landscape of a trna gene. *Science*, 352(6287):837–840, 2016. doi: 10.1126/science.aae0568. URL https://www.science.org/doi/abs/10.1126/science.aae0568.

Aditi T. Merchant, Samuel H. King, Eric Nguyen, and Brian L. Hie. Semantic mining of functional de novo genes from a genomic language model. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:274893904.

Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, Stefano Ermon, Stephen A Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D Hsu, and Brian L Hie. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, 2024. doi: 10.1101/2024.02.27.582234.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36, 2023.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks, 2024.

Gianluca Peri, Clémentine Gibard, Nicholas H Shults, Kent Crossin, and Eric J Hayden. Dynamic RNA Fitness Landscapes of a Group I Ribozyme during Changes to the Experimental Environment. *Molecular Biology and Evolution*, 39(3):msab373, 01 2022. ISSN 1537-1719. doi: 10.1093/molbev/msab373. URL https://doi.org/10.1093/molbev/msab373.

Jason N. Pitt and Adrian R. Ferré-D'Amaré. Rapid construction of empirical rna fitness landscapes. *Science*, 330(6002):376–379, 2010. doi: 10.1126/science.1192001. URL https://www.science.org/doi/abs/10.1126/science.1192001.

Jessica S. Reuter and David H. Mathews. Rnastructure: software for rna secondary structure prediction and analysis. *BMC Bioinformatics*, 11:129 – 129, 2010. URL https://api.semanticscholar.org/CorpusID:10356201.

Jessica M Roberts, James D Beck, Tanner B Pollock, Devin P Bendixsen, and Eric J Hayden. Rna sequence to structure analysis from comprehensive pairwise mutagenesis of multiple self-cleaving ribozymes. *eLife*, 12: e80360, jan 2023. ISSN 2050-084X. doi: 10.7554/eLife.80360. URL https://doi.org/10.7554/eLife.80360.

Jaswinder Singha, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications*, 2019. doi: https://doi.org/10.1038/s41467-019-13395-9. URL https://www.nature.com/articles/s41467-019-13395-9.

Valerie W. C. Soo, Jacob B. Swadling, Andre J. Faure, and Tobias Warnecke. Fitness landscape of a dynamic rna structure. *PLOS Genetics*, 17(2):1–21, 02 2021. doi: 10.1371/journal.pgen.1009353. URL https://doi.org/10.1371/journal.pgen.1009353.

Shunsuke Sumi, Michiaki Hamada, and Hirohide Saito. Deep generative design of rna family sequences. *Nature Methods*, 21:435–443, 2024.

Jacob M Tome, Abdullah Ozer, John M Pagano, Dan Gheba, Gary P Schroth, and John T Lis. Comprehensive analysis of rna-protein interactions by high-throughput sequencing–rna affinity profiling. *Nature methods*, 2014. doi: 10.1038/nmeth.2970. URL https://www.nature.com/articles/nmeth.2970.

Brent Townshend, Andrew B Kennedy, Joy S Xiang, and Christina D Smolke. High-throughput cellular rna device engineering. *Nature Methods*, 2015. doi: https://doi.org/10.1038/nmeth.3486. URL https://www.nature.com/articles/nmeth.3486.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

F. H. D. van Batenburg, Alexander P. Gultyaev, Cornelis W. A. Pleij, J. Ng, and J. Oliehoek. Pseudobase: a database with rna pseudoknots. *Nucleic acids research*, 28 1:201–4, 2000. URL https://api.semanticscholar.org/CorpusID:27636094.

Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 6(5):548–557, 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00836-4. URL https://doi.org/10.1038/s42256-024-00836-4.

Hannah K. Wayment-Steele, Wipapat Kladwang, Alexandra I. Strom, Jeehyung Lee, Adrien Treuille, Alexander J Becka, and Rhiju Das. Rna secondary structure packages evaluated and improved by high-throughput experiments. *bioRxiv*, 2020. URL https://api.semanticscholar.org/CorpusID:219311051.

Caleb Weinreb, Adam J Riesselman, John B Ingraham, Torsten Gross, Chris Sander, and Debora S Marks. 3D RNA and functional interactions from evolutionary couplings. *Cell*, 165(4):963–975, May 2016.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, Defne G. Ozgulbas, Natalia Vassilieva, James Gregory Pauloski, Logan Ward, Valerie Hayot-Sasson, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, and Arvind Ramanathan. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 37(6):683–705, 2023. doi: 10.1177/10943420231201154.

APPENDIX

A DATASET COLLECTION

Prioritization of studies for expert review The selection process for prioritized studies for expert review was structured as follows. We initiated with a targeted PubMed search, utilizing specific queries to ensure a comprehensive capture of relevant literature:

1. **Deep Mutational Scan:** (deep[All Fields] OR comprehensive[All Fields]) AND (mutational[All Fields] OR muta*) AND (scan OR scans OR scanning) AND RNA
2. **Saturation Mutagenesis:** (saturat* muta*) AND RNA
3. **MAVE:** ("Multiplex* assay" AND "variant*") AND RNA
4. **MPRA:** ("MPRA" OR "Massively parallel reporter assay*") AND RNA NOT protein
5. **Other:** ("Fitness Landscape" AND muta*) AND RNA

These initial searches proved to be either overly restrictive or too broad, which complicated the manual screening process. Ultimately, this approach resulted in the identification of only 20 primary articles. To enrich our review, an additional 10 articles were identified by scraping references from other pertinent studies, including those cited in previous research such as (Sumi et al., 2024; Nguyen et al., 2024), and RNA-related datasets from studies like ProteinGym (Notin et al., 2023).

Hypothesizing that our initial search strategy may have missed relevant studies, we conducted a comprehensive PubMed search using the following query:

Broad Search: (deep OR comprehensive OR MPRA OR multiplex assay OR massively parallel OR landscape OR saturation) AND (muta* OR variant OR variants) AND (scan OR scans OR screen OR landscape OR assay) AND (RNA OR ribozyme* OR microRNA OR miRNA OR siRNA OR snoRNA OR tRNA OR lncRNA OR (RNA AND aptamer) OR circRNA)

A.1 LITERATURE PRE-SCREENING WITH LLM

The prior search yielded an overwhelming 11,635 results. To efficiently handle this volume, we utilized a large language model (LLM), specifically GPT4-0125-preview, for secondary screening. We adapted a recent prompting approach designed for systematic review screening (Cao et al., 2024). The LLM was instructed with clear study objectives and specific inclusion/exclusion criteria, effectively narrowing down the pool to fewer than 500 articles, thereby making manual curation manageable. To enhance the sensitivity of this process, the LLM's prompt was refined using an initial set of 30 positively identified articles as a control group. This novel use of LLMs for data extraction markedly improved our capacity to pinpoint relevant studies. Consequently, we were able to incorporate an additional 22 studies into our initial screen, resulting in a total of 52 studies ready for manual expert review.

We used the following prompt to pre-screen relevant studies during our extensive literature search:

"The goal of this study is to create a benchmark that contains RNA deep mutational scanning or fitness landscape datasets. We are generating these datasets to benchmark RNA fitness prediction algorithms, and need our datasets we evaluate to have information on RNA mutants/variants and their relative 'fitness'.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled): 1. Studies involve RNA. We are also interested in RNA subclasses such as Ribozyme, lncRNA, tRNA, rRNA, microRNA (miRNA), Aptamer, Riboswitch, mRNA 2. Studies report on fitness prediction. Other terms for fitness prediction can include deep mutational scans, comprehensive multiplex assays, or comprehensive fitness landscapes, among others 3. Studies with greater than 100 experimental measurements 4. Studies that report on mutant fitness through reporter assays, bulk RNA-sequencing, single-cell RNA sequencing assay, fluorescence in-situ hybridization (FISH) assay, flow cytometry assay, imaging mass cytometry assay, evolution of ligands by exponential enrichment assay, single cell imaging, multiplexed fluorescent antibody imaging, binding assays, cell proliferation assay, splicing assays, survival assessment assay selection types, or similar. 5. Studies that report on enzymatic activity, binding affinity, stability, fluorescence, proliferation selection assays, or similar assays. 6. The study must be primary research and generate a novel dataset

Exclusion Criteria (if any met then exclude): 1. Studies only reporting on protein mutational scans, with no relevance or mention of RNA being mutated 2. Studies that do not focus on fitness quantification 3. Review articles (systematic reviews, case reports, case series, etc.) or other non-primary research sources.

Instructions

We now assess whether the paper should be included in the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met. Studies that may not fully align with the primary focus of our inclusion criteria but provide data or insights potentially relevant to our review deserve thoughtful consideration. Given the nature of abstracts as concise summaries of comprehensive research, some degree of interpretation is necessary. Our aim should be to inclusively screen abstracts, ensuring broad coverage of pertinent studies while filtering out those that are clearly irrelevant. We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.

Title and Abstract in investigation:

Title: #Insert title of study#

Abstract: #Insert abstract of paper#"

Expert review process The process for accepting a paper involved several steps to ensure the quality and relevance of the data. First, we checked whether the data was openly available and could be integrated into our benchmark. If data was not accessible, study authors were contacted.

Next, we used the following inclusion and exclusion criteria during our through expert review process:

Inclusion Criteria

- Assay must focus on RNA
- Assay must have at least 100 experimental variants tested, with a sufficiently wide dynamic range
- Assay must be relevant to fitness prediction, and report on mutant fitness
- Assay must only focus on substitutions, not insertions or deletions

Exclusion Criteria

- Assays focusing on DNA or Proteins
- Assays that are not primary research
- Assays with mutants of varying lengths

B DATASETS DETAILS

B.1 FITNESS ASSAYS

References An exhaustive list of the publications from which the assays included our fitness benchmark originated from is provided in Table B.1.

Licenses All fitness assays were licensed under CC-BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`), or the ACS AuthorChoice Usage Agreement (`https://pubs.acs.org/page/policy/authorchoice_termsofuse.html`).

Cross-validation splits For users interested in supervised RNA fitness prediction, we provide two types of cross-validation splits:

- **Random:** a random 80%-20% train-test split;
- **Minimum similarity:** a 80%-20% split in which we minimized the sequence similarity between training and validation RNA sequences.

B.2 STRUCTURE PREDICTION DATASET

We constructed the data for our structure prediction challenge from the DMS data from the Ribonanza Challenge hosted on Kaggle (`https://www.kaggle.com/competitions/stanford-ribonanza-rna-folding/data`). To ensure the integrity of our dataset and prevent data leakage, we removed sequences that were previously utilized by the Ribonanzanet model in the training dataset, as well as the 'public' portion of the evaluation dataset. The final dataset comprises $115,000$

Title	Year	RNA type	# Assays	Reference	License
Saturation mutagenesis reveals manifold determinants of exon definition	2017	mRNA	1	(Ke et al., 2017)	CC BY 4.0
Fitness landscape of a dynamic RNA structure	2021	ribozyme	1	(Soo et al., 2021)	CC BY 4.0
Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme	2020	ribozyme	1	(Andreasson et al., 2020)	CC BY 4.0
High-throughput assay and engineering of self-cleaving ribozymes	2015	ribozyme	3	(Kobori et al., 2015)	CC BY 4.0
Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis	2014	tRNA	1	(Guy et al., 2014)	CC BY 4.0
Deep sequencing analysis of aptazyme variants based on a pistol ribozyme	2018	ribozyme	1	(Kobori et al., 2017)	ACS Author-Choice Usage Agreement
High-throughput cellular RNA device engineering	2015	aptamer	5	(Townshend et al., 2015)	CC BY 4.0
Rapid Construction of Empirical RNA Fitness Landscapes	2010	ribozyme	1	(Pitt & Ferré-D'Amaré, 2010)	CC-BY 4.0
Dynamic RNA Fitness Landscapes of a Group I Ribozyme during Changes to the Experimental Environment	2022	ribozyme	1	(Peri et al., 2022)	CC-BY 4.0
RNA sequence to structure analysis from comprehensive pairwise mutagenesis of multiple self-cleaving ribozymes	2023	ribozyme	5	(Roberts et al., 2023)	CC-BY 4.0
Pairwise and higher-order genetic interactions during the evolution of a tRNA	2018	tRNA	1	(Domingo et al., 2018)	CC-BY 4.0
Emergent properties as by-products of prebiotic evolution of aminoacylation ribozymes	2022	ribozyme	5	(Janzen et al., 2022)	CC-BY 4.0
Predicting higher-order mutational effects in an RNA enzyme by machine learning of high-throughput experimental data	2022	ribozyme	1	(Beck et al., 2022)	CC-BY 4.0
The fitness landscape of a tRNA gene	2016	tRNA	1	(Li et al., 2016)	CC-BY 4.0
Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling	2014	aptamer	2	(Tome et al., 2014)	CC-BY 4.0
The complete local genotype-phenotype landscape for the alternative splicing of a human exon	2016	mRNA	1	(Julien et al., 2016)	CC-BY 4.0

Table A1: **RNAGym fitness prediction data.** We developed our fitness prediction benchmark by curating and processing 31 assays from 16 publications.

distinct sequences, encompassing over $15,000,000$ positions where structural predictions have been applied. This extensive dataset underpins the robustness and comprehensive nature of our RNA structure prediction challenge. The original data from the challenge is made available under a CC-BY 4.0 license.

C Baselines

C.1 RNA fitness prediction models

Our fitness prediction benchmarks currently include the following 7 baselines:

- **RiNALMo** (Penić et al., 2024) is a 650 million parameter RNA language model trained on 36 million non-coding RNA sequences, achieving high performance in RNA structural and functional prediction

tasks. Variants were scored using the masked marginal scoring strategy from the ESM sequence modeling framework.

- **Evo** (Nguyen et al., 2024) is a 7 billion parameter model trained on 2.7M prokaryotic and phage genomes to generate DNA sequences using a context length of 8k (further extended to131k tokens). It is based on StripedHyena, a deep signal processing architecture designed to improve efficiency and quality over the prevailing Transformer architecture. We use the 8k version of the model, refered as Evo 1. The recently released Evo 1.5 builds upon Evo 1 by increasing the pretraining data from 300 billion to 450 billion tokens Merchant et al. (2024).
- **RNA-FM** (Chen et al., 2022) is a RNA foundation model based on the BERT language model architecture. It was trained on 23 million unlabeled non-coding RNA sequences from over 800,000 species, collected from RNA-central. Variants were scored using the wild-type marginal scoring strategy from the ESM sequence modeling framework, which RNA-FM builds upon.
- **GenSLM** (Zvyagin et al., 2023) includes an autoregressive RNA language model with codon-level tokenization, along with stable diffusion, to model genome-scale interactions and predict SARS-CoV-2 evolution. It was trained on 110 million prokaryotic coding sequences from BV-BRC. Variant sequence likelihoods were scored using the 2.5 billion parameter language model.
- **RNAErnie** (Wang et al., 2024) is a 12-layer transformer-based RNA language model pre-trained on 23 million ncRNA sequences using masked language modeling. With 105 million trainable parameters, it achieves high performance in RNA sequence classification, RNA–RNA interaction, and RNA secondary structure prediction.
- **Nucleotide Transformer** (Dalla-Torre et al., 2023) is a 2.5B parameter model trained on 850 species. The 2.5b-multi-species version was used.

C.2 RNA STRUCTURE PREDICTION MODELS

Our structure prediction benchmarks currently include the following 5 baselines (in addition to RNA-FM, introduced earlier):

- **EternaFold** (Wayment-Steele et al., 2020) is built on the principles derived from the Eterna massive open online game, where players design RNA sequences that fold into target shapes. This model incorporates crowd-sourced insights from thousands of players to refine its algorithms, significantly enhancing its ability to predict RNA structures under varied environmental conditions and complexities.
- **CONTRAfold** (Do et al., 2006) is a machine learning-based RNA secondary structure prediction model that utilizes conditional log-linear models for structure inference.
- **Vienna** (Gruber et al., 2008) is one of the most widely used RNA secondary structure prediction tools. It employs dynamic programming algorithms based on thermodynamic models to accurately predict RNA secondary structures, including handling pseudoknotted structures as extensions.
- **RNAstructure** (Reuter & Mathews, 2010) is a model designed for the prediction and analysis of RNA secondary structures. It is known for its dual ability to use either thermodynamic or machine learning-based methods to predict RNA folding patterns.
- **RNA-FM** (Chen et al., 2022) as described above was also used for RNA structure prediction by utilizing its downstream second-order structure prediction module.

We also contextualize our zero-shot results by contrasting performance with the supervised model **Ribonanzanet** (He et al., 2024). This model is a sequence-only deep neural network trained on the data from the Stanford Ribonanza Challenge. It was further fine tuned on pseudo labels derived from the predictions of the top 3 models from the Kaggle challenge.

D LIMITATIONS

Experimentally assaying RNA fitness, while resource-intensive, provides critical insights that help advance our understanding of RNA function. However, such experimental assays may not always accurately mimic the cellular environment, which can lead to variations between observed in vitro results and actual in vivo functionality. Chemical mapping experiments using dimethyl sulfate (DMS) offer valuable data on RNA secondary structures by identifying accessible adenine and cytosine bases that interact with DMS. This technique, although powerful for revealing the in-vivo-like structure of RNA in a relatively high-throughput manner, has its limitations. DMS mapping can be affected by incomplete coverage, as it primarily marks only two of the four nucleotide types. Additionally, the resolution of DMS mapping might not always distinguish closely spaced structural features, potentially obscuring important details about RNA folding and interaction sites. The accuracy of predictions from DMS data also heavily depends on the computational tools used to interpret the chemical

reactivity patterns, necessitating ongoing improvements in both experimental and analytical methodologies to enhance the precision and utility of RNA structural studies. There is also sampling bias for the RNA families that were assayed using high-throughput fitness assays, with not all families equally represented. The majority of fitness assays were focused on ribozymes while other RNA families such as mRNA and tRNA had far fewer available datasets to evaluate.

E SOCIETAL IMPACT

The advancement of RNA models holds transformative potential across a spectrum of applications. By accurately predicting RNA structure and fitness, researchers can unlock new therapies by targeting previously intractable genetic conditions, enhance crop resilience through agricultural biotechnology, and even engineer microbial systems for cleaner energy production. The creation of benchmarks like RNAGym is crucial in this endeavor, as they drive the field forward by setting standards for model performance and fostering innovation through competition and collaboration. However, as it is the case for any approach that facilitates the development of novel biological sequences for good, the potential misuse of these technologies to create harmful biological agents cannot be ignored (Urbina et al., 2022). It is imperative to proceed with a careful framework that promotes secure use, ethical guidelines, and synthesis monitoring (Baker & Church, 2024) to mitigate risks associated with dual-use capabilities. Ultimately, benchmarks like RNAGym not only validate the effectiveness of emerging RNA models but also, by highlighting the methods leading to step-change performance improvements, encourage their integration into real-world applications, ensuring that these innovations contribute positively to society.

F COMPUTE RESOURCES

In our benchmarking work for RNA fitness prediction and structure analysis, we primarily rely on GPUs as hardware accelerators to handle the computationally intensive tasks involved. We estimate our compute budget for both fitness and structure prediction benchmarks to approximately 30 V100 GPU days.

G DETAILED EXPERIMENTAL RESULTS

G.1 FITNESS PREDICTION PERFORMANCE BY ASSAY

We report the assay-level Spearman performance, across all assays in the RNAGym fitness prediction benchmark in Fig G.1.

G.2 STATISTICAL SIGNIFICANCE OF FITNESS PREDICTION PERFORMANCE

We report the statistical significance for the relative Spearman performance by RNA type in Table A2. We follow the same methodology as in ProteinGym (Notin et al., 2023) and assess statistical significance by computing the non-parametric bootstrap standard error of the difference between the Spearman performance of a given model and that of the best overall model.

Rank	Model name	mRNA		tRNA		Aptamer		Ribozyme		All	
		Diff	SE	Diff	SE	Diff	SE	Diff	SE	Diff	SE
1	Evo 1.5	-0.114	0.027	-0.087	0.05	-0.02	0.013	0	0	0	0
2	RNAErnie	-0.066	0.022	-0.067	0.096	-0.048	0.043	-0.043	0.021	-0.001	0.017
3	Evo 1	-0.132	0.029	-0.085	0.052	0	0	-0.011	0.011	-0.002	0.008
4	RNA-FM	-0.193	0.002	0	0	-0.07	0.07	-0.022	0.026	-0.016	0.023
5	RiNALMo	0	0	-0.204	0.151	-0.144	0.052	-0.077	0.024	-0.051	0.02
6	Nucl. Transformer	-0.029	0.093	-0.369	0.091	-0.022	0.044	-0.059	0.019	-0.065	0.019
7	GenSLM	-0.151	0.068	-0.346	0.121	-0.087	0.046	-0.051	0.025	-0.104	0.023

Table A2: **Fitness prediction by RNA type - Difference in Spearman to best score by category** Difference in average of Spearman's rank correlation between model scores and experimental measurements to the best model by category, by RNA type and overall. The standard error reported corresponds to the non-parametric bootstrap standard error of the difference between the Spearman performance of a given model and that of the best overall model for a given category, computed over 10k bootstrap samples from the set of assays in the RNAGym fitness benchmark.

Figure 2: **RNAGym fitness prediction benchmark - Detailed performance.** Spearman's rank correlation between model predictions and experimental values for each assay in the RNAGym fitness prediction benchmark.

G.3 STRUCTURE PREDICTION - PSEUDOKNOTS

We mapped RNA sequences in our evaluation set to pseudoknot annotations from PseudoBase (van Batenburg et al., 2000) (358 test sequences mapped), and report the corresponding global F1 score and crossed pair F1 score (Table A3). Out of our various structure prediction baselines, only RNAstructure, Ribonanzanet and RNA-FM are able to score pseudoknots.

Rank	Model name	Global F1 score	Crossed Pair F1 score
1	Ribonanzanet	0.758	0.583
2	RNA-FM	0.654	0.423
3	RNAstructure	0.639	0.211
4	CONTRAfold	0.596	N/A
5	EternaFold	0.588	N/A
6	Vienna	0.579	N/A

Table A3: **Structure prediction - Pseudoknots.** Global and Crossed Pair F1 score on subset of eval sequences with annotations in PseudoBase.