RMSAGen: Integrating Multiple Sequence Alignment for Function RNA Design

Jiyue Jiang[1,*], Yanyu Chen[1,*], Qingchuan Zhang[1], Jiayi Li[1], Xiangyu Shi[1], Chang Zhou[1], Ziqian Lin[1], Jiuming Wang[1], Dongchen He[1], Liang Hong[1], Qintong Li[2], Pengan Chen[1], Jiayang Chen[1], Xinrui Zhang[1], Jiao Yuan[3,4,†], Tianqing Zhang[5,†], Yu Li[1,†]

[1]The Chinese University of Hong Kong
[2]The University of Hong Kong
[3]Guangzhou National Laboratory
[4]Guangzhou Medical University
[5]Hangzhou Institute of Medicine, Chinese Academy of Sciences
{jiangjy, chenyanyu.cse}@link.cuhk.edu.hk, yuan_jiao@gzlab.ac.cn, ztq.edu@gmail.com, liyu@cse.cuhk.edu.hk

Abstract

Biological sequences, including RNAs and proteins, share similarities with natural languages, enabling the application of advanced language models to various biological tasks. However, due to its flexibility and lack of experimental data, RNA is a particularly challenging biological "language" compared to other biological sequences like proteins. RNA multiple sequence alignments (MSAs), which align evolutionarily related RNA sequences, can greatly enhance RNA biology modeling, as evidenced by their significant roles in structure prediction and function annotation. This raises the question of whether RNA MSAs can also benefit RNA design, which remains unexplored. This paper introduces RMSAGen, a model comprising RMSA-Encoder and RMSA-Decoder, that leverages MSAs to design functional RNA sequences. RMSA-Encoder effectively extracts MSA features, enhancing performance in functional prediction and solvent accessibility prediction tasks and supporting RMSA-Decoder in accurate RNA generation. RMSAGen can design RNA sequences that effectively bind to target RNA-binding proteins, and the design performance improves with an increasing number of sequences. In addition, the ribozymes designed with structural features by RMSAGen show strong computational metrics and exhibit biological activity during gel electrophoresis. These results highlight the effectiveness of RMSAGen, establishing it as a powerful tool and a new direction for RNA design.

Introduction

RNA plays a critical role in the central dogma of molecular biology, and RNA design has long been a research focus (Shen et al. 2024a; Jiang et al. 2025b). Previous studies have primarily concentrated on proteins (Ferruz, Schmidt, and Höcker 2022a; Zhou et al. 2025); however, RNA, which serves as a biological "language" more challenging to interpret than proteins (Shen et al. 2024a; Zhang et al. 2023), presents greater difficulties in design, especially when dealing with multiple sequence alignments containing evolutionary information (Zhang et al. 2023).

Compared to proteins, RNA molecules exhibit more flexible conformations (Shen et al. 2024a), exacerbating the challenges of RNA design. MSAs contain conserved genetic information (Zhang, Zhang, and Pyle 2023; Zhang et al. 2023); effectively leveraging RNA MSAs can enhance the performance and stability of designed RNA sequences. Existing models predominantly rely on MSAs for structure prediction and similar characterization tasks, demonstrating that MSAs can stabilize conformations and enable models to predict more accurate structures (Zhang, Zhang, and Pyle 2023; Shen et al. 2024a).

In this paper, we introduce RMSAGen, which integrates MSAs to design RNA sequences, enhancing the performance and stability of the designed RNAs. RMSAGen consists of two components: RMSA-Encoder and RMSA-Decoder. RMSA-Encoder is utilized to extract features from MSAs. Its outstanding performance in function prediction and solvent accessibility prediction tasks show that the model can effectively learn MSA features, thus better supporting RMSA-Decoder in generating RNAs. We construct MSA data from certain highly conserved RNAs capable of binding to proteins and employ RMSAGen to design the corresponding sequences. We find that the designed sequences achieve effects comparable to those of natural sequences, effectively binding to specific proteins. As the number of sequences in the input MSA increases, the binding effectiveness also improves, and upon reaching a higher sequence count, its performance approaches, and shows the potential to rival, that of natural sequences. In addition, RMSAGen also designs ribozymes (hammerhead) based on MSA conservation and structure feature. These designed ribozymes not only perform well in computational metrics but also exhibit activity in the gel electrophoresis experiment. Through various computational and biological experiments, we demonstrate that it is an effective RNA design tool.

Our contributions are summarized as follows: (1) We have introduced a new task focused on designing RNA based on RNA MSA and have developed a novel model for RNA MSA that undertakes both characterization and RNA gener-

[*]These authors contributed equally to this work.
[†]Corresponding authors.

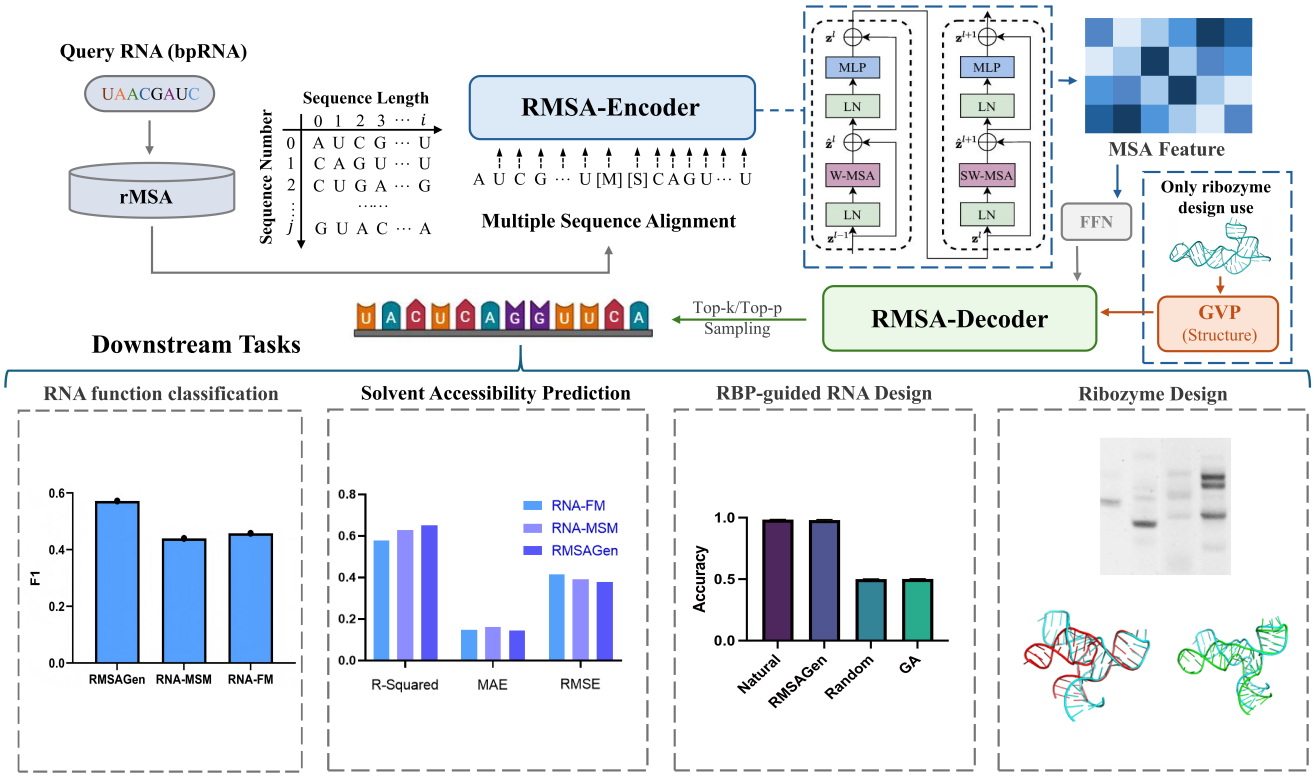

Figure 1: Overview of the RMSAGen architecture. The query RNAs from the bpRNA database are processed using the rMSA to construct RNA MSAs. MSA features are extracted using RMSA-Encoder, and these features are input into RMSA-Decoder through a feed-forward network to guide the generation of RNA sequences. RMSAGen performs excellently in the following three downstream tasks: RNA family classification, RNA solvent accessibility, RBP-guided RNA design, and ribozyme design.

ation tasks. (2) RMSA-Encoder extracts features from RNA MSAs and validates these features through classification and regression, achieving state-of-the-art results compared to existing methods. (3) In RBP-guided RNA design tasks, RMSAGen not only proves effective in designing RNA sequences that bind to proteins but also shows increased effectiveness as the number of sequences grows. (4) In the critical field of ribozyme design, sequences designed with structural features by RMSAGen demonstrate biological activity in biological validation, confirming the effectiveness of RMSAGen.

Methodology: RMSAGen

Transformers (Vaswani et al. 2017) demonstrate outstanding textual understanding and generating capabilities even on non-natural languages such as biological molecular sequences (Ji et al. 2021; Ferruz, Schmidt, and Höcker 2022b; Madani et al. 2023; Jiang et al. 2025a; Wang et al. 2025). We propose a transformer architecture for RNA design based on MSAs. The RMSA-Encoder takes MSA sequences and processes them through twelve self-attention layers to capture the contextual dependencies between nucleotides. The decoder generates the aligned sequences auto-regressively by attending to both the encoder outputs and previously generated tokens. However, when applying the original Transformer to the MSA task by concatenating all the sequences, it works but leads to a sharp increase in memory utilization: with M (up to 128) sequences of length N (up to 512), the space complexity is $O((MN)^2)$. We address this issue by using 2D positional embeddings and computing the attention separately in the row and column directions of the MSA, as demonstrated in (Rao et al. 2021).

2D Attention Mechanism

Following the work of (Rao et al. 2021), we operate the attention mechanism independently over rows and columns of MSAs (i.e., sequence length and sequence number of MSA), reducing the space complexity from $O((MN)^2)$ to $max\{O(NM^2), O(MN^2)\}$, and a tied row attention is applied. Structure is a critical determinant of the properties and functions of biomolecules (Watson and Crick 1953; Pauling 1951). In a MSA, all sequences perform the same function, thus possessing the same or similar structures, which implies that the MSA exhibits the same or similar attention weights across all sequences. Therefore, we also adopt the tied attention mechanism:

$$\text{Attention}(Q_m, K_m, V_m) = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d_k}}\right) V_m \quad (1)$$

where Q_m and K_m are the queries and keys for each row

of MSA, and d_K is the normalization factor of dot-product attention as described in (Vaswani et al. 2017).

Model Architecture and Training

RMSAGen includes an encoder (RMSA-Encoder) and a decoder (RMSA-Decoder). In the task of ribozyme design, it integrates structural features of ribozymes for improved design.

Encoder: RMSA-Encoder To better extract the features of RNA MSAs, we first train a 12-layer representation model (Devlin et al. 2019) on masked language model (MLM) task to encode MSAs. We randomly select 15% of the nucleobases and replace 80% of them with a special mask token [MASK], then calculate the loss function:

$$\mathcal{L}_{enc} = -\sum_{i \in \mathcal{M}} \log p(x_i | f(x_i)), \tag{2}$$

where $\mathcal{M}$ is the indices of masked nucleobases of MSAs, $f(x_i)$ is the last hidden states from the RMSA-Encoder model encoding the input nucleobase x_i.

To validate the effectiveness of the RMSA-Encoder component in extracting MSA features, the RNA family classification task divides the functions of MSAs into 11 categories, namely 5S ribosomal RNA, U2 spliceosomal RNA, tRNA, U5 spliceosomal RNA, U6 spliceosomal RNA, group I catalytic intron, FMN riboswitch aptamer (RFN element), purine riboswitch aptamer, glmS glucosamine-6-phosphate activated ribozyme aptamer, SAM/SAH riboswitch aptamer, bacterial large subunit ribosomal RNA. We feed the output of the last hidden layer of the RMSA-Encoder model H_e into a convolutional neural network (CNN), which consists of two convolutional layers and two linear layers to predict scores over 11 functions of MSAs.

$$S^* = \mathrm{ConvNet}(H_e),$$
$$\mathcal{L}_C = -\sum_{c=1}^{11} [c == \mathrm{argmax}(S^*)] \log \frac{\exp(S_c^*)}{\sum_{i=1}^{11} \exp(S_i^*)}, \tag{3}$$

where $S*$ is the logits predicted by the CNN and $\mathcal{L}_C$ denotes the classification loss.

Encoder-Decoder: RMSAGen In the RBP-guided RNA design task, we train an RNA MSA generative model (24-layer model): RMSA-Decoder. We connect the RMSA-Encoder model and the RMSA-Decoder model through a feed-forward network (FFN), forming RMSAGen model. Although RMSAGen follows an encoder-decoder architecture, during training, RMSA-Encoder only takes MSAs as input, while the target for RMSA-Decoder is the concatenation of the source MSAs and the target sequences. This design allows both the RMSA-Encoder and the RMSA-Decoder to be used independently and to generalize to downstream tasks. Finally, the RMSAGen is optimized by minimizing the negative log-likelihood of predicting the next token/nucleobase:

$$\mathcal{L} = \sum_{i=1}^{N} -\log p(y_i | y_{<i}, x_{1:N}). \tag{4}$$

Fusion Structure to Design Ribozyme To enhance the design of ribozymes, we incorporate structural features to assist. The geometric vector perceptron (GVP) (Jing et al. 2021; Huang et al. 2024) model can encode the geometric structure of an RNA molecule. We select the coordinates of the P, C1, and C3 atoms to form a point cloud $C \in \mathbb{R}^{3 \times N \times 3}$ as the geometric feature input, and use the K-nearest neighbor algorithm to construct a topological graph, which forms the 4-layer GVP graph network to encode RNA structural representations. We fuse the MSA and structural features using an 8-head attention mechanism to constrain the structure of generated RNA sequences:

$$G = \mathrm{GVP}(C), \tag{5}$$
$$\mathcal{H} = \frac{W_q H_d \cdot (W_k C)^T}{\sqrt{d_k}} W_v C + H_d, \tag{6}$$

where H_d is the hidden states of the last layer of the RMSA-Decoder part, W_q, W_k, W_v are trainable parameters of an attention head, and $d_k = 512$ is the embedding size of the GVP feature. $\mathcal{H}$ replaces H_d to output layer and fine-tunes the generation.

Generating by Sampling

During inference, the RNA sequence is generated by sampling method (Fan, Lewis, and Dauphin 2018), with top-k, top-p sampling. Following (Ferruz, Schmidt, and Höcker 2022b; Madani et al. 2023), biological sequences such as RNAs can have vast and highly diverse range of valid structures and functions, the sampling method, compared to the commonly used beam search method in natural language generation, can generate more diverse sequences and avoid over-fitting to few results.

When selecting these parameters, we refer to the selection method used by (Ferruz, Schmidt, and Höcker 2022b; Madani et al. 2023). The top-k ranges from 100 to 1200, selecting every 50 increment. top-p ranges from 0.6 to 1.0, selecting every 0.05 increment. For each type of sampling, the hyper-parameters in the unconditional mode generate 100 sequences. The optimal parameters are those for which base composition ratio is closest in comparison to the most similar sequences from RNAcentral (rna 2019) (the sequences of RNAcentral is natural sequences). We have determined that the optimal top-k is 1000 and top-p is 0.7. To ensure diversity in the generation, the temperature is set to 1.0.

Experiment

Implementation Details

All models are implemented using PyTorch (Paszke et al. 2019) across eight NVIDIA A100 GPUs. The models are trained employing the AdamW optimizer (Loshchilov and Hutter 2017) with a batch size of 1. During training, the learning rate is varied following the strategy outlined by (Vaswani et al. 2017). For inference, the temperature is set to 1.0, top-k to 1000, and top-p to 0.7.

Datasets

Datasets for RMSAGen Pre-Training We utilize data from the bpRNA database (Danaee et al. 2018) as query se-

quences and employ the rMSA method (Zhang, Zhang, and Pyle 2023) to construct corresponding MSAs. We divide the rest every 128 sequences to form new MSAs (to better capture the features of the MSA, we ensure that the first 10% of each new MSA retains sequences similar to the query). The output sequences are the same as the query sequences.

Datasets for Predicting RNA Family We obtain new query sequences from the mmcif2pdb[1], and employ the rMSA method (Zhang, Zhang, and Pyle 2023) to construct corresponding MSAs, using the same method as for the pre-training data construction. We use the cm models in Infernal (Nawrocki and Eddy 2013) for functional/family annotation, and we select the top 11 most populous functions/families to ensure the model can learn the features, namely 5S ribosomal RNA (RF00001), U2 spliceosomal RNA (RF00004), tRNA (RF00005), U5 spliceosomal RNA (RF00020), U6 spliceosomal RNA (RF00026), group I catalytic intron (RF00028), FMN riboswitch aptamer (RFN element) (RF00050), purine riboswitch aptamer (RF00167), glmS glucosamine-6-phosphate activated ribozyme aptamer (RF00234), SAM/SAH riboswitch aptamer (RF01727), bacterial large subunit ribosomal RNA (RF02541). These RNA families are deliberately selected to represent a diverse spectrum of functional RNA classes, thereby providing a rigorous and comprehensive evaluation of classification of RMSAGen performance across different RNA families.

Datasets for Predicting RNA Solvent Accessibility We extract RNA data and atomic coordinates of each nucleotide from the PDB database. Using solvent accessibility tools, we calculate the accessible surface area for each base. Incomplete or fragmented RNA entries are filtered out to maintain data quality. Finally, we construct data correlating RNA sequences with base solvent accessibility.

Datasets for RBP-guided RNA Design For this task, we utilize RNAcompete data (Ray et al. 2017) as the dataset. We select these datasets because they adequately represent the categories of sequences of RBPs, with the RNAcompete data comprising 133 RBPs.

We use the MAFFT (Katoh et al. 2002) tool to convert the sequences of these RBPs into the format of MSAs. For MSAs inputted with sequence counts of 1, 10, 50, and 128, we prioritize and organize them based on the number of conserved bases in the sequences, selecting those with the highest number of conserved bases first for the construction of the input MSAs. The output sequences are consistent with the pre-training, serving as the query sequences. The output sequence is the query sequence of the MSA, which is consistent with the pre-training.

Datasets for Ribozyme Design We obtain tertiary structure data from the RNAsolo (Adamczyk, Antczak, and Szachniuk 2022) database and annotate sequences extracted from these structures using the cm models from Infernal tool (Nawrocki and Eddy 2013). We select structures of RNA families with aligned sequence lengths less than 512, along with their corresponding RNA family MSAs, as input for training data. The output consists of sequences extracted from the structures.

[1]https://mmcif.wwpdb.org/

Evaluation

For the task of predicting RNA family, we use **Accuracy** and **F1** score as evaluation metrics.

For the task of predicting RNA solvent accessibility, we use **R-squared**, **MAE** and **RMSE** score as evaluation metrics. For the RBP-guided RNA design task, we employ PrismNet (Sun et al. 2021) as a tool to evaluate whether RNA and protein can bind (Sun et al. 2021; Jiang et al. 2025c). We use the **Accuracy**, **AUC**, and **AUPRC** metrics from the PrismNet tool as evaluation metrics of this task. For the ribozyme design task, we employ a secondary structure similarity evaluation method (**F1**) and a tertiary structure similarity evaluation method (**RMSD**) to verify if the structures of the generated sequences resemble those of the wild type. **Sequence similarity** is used to evaluate the resemblance to the wild type sequence, with lower values indicating higher diversity.

Baselines

For task of prediction of RNA family and solvent accessibility, we adopt existing language models proficient in RNA classification or RNA MSA as baselines. These include RNA-FM (Shen et al. 2024a; Chen et al. 2022), a foundation model for RNA based on the bert architecture, and RNA-MSM (Zhang et al. 2023), a language model for RNA multiple sequence alignment.

For the RBP-guided RNA design task, given its novelty and the absence of appropriate baselines, we opt to compare our approach with natural sequences, as well as genetic algorithm (GA) and random splicing methods as baselines.

This implies that if the sequences generated by RMSAGen perform similarly to or exceed natural sequences in effectiveness, and can demonstrate a significant distinction from sequences generated by GA and random methods, it indicates that the generated sequences are of high quality.

For ribozyme design, in order to evaluate whether the ribozyme sequences designed by RMSAGen are active, and to validate RMSAGen as a viable tool in biology, we therefore proceed without baselines for this application.

Results and Analysis

Prediction of RNA Family

The performance of the RMSA-Encoder (component within RMSAGen) is systematically evaluated through a classification task spanning eleven distinct classes. RMSAGen demonstrates superior classification capabilities in its performance distribution, achieving an accuracy (ACC) of 0.692 and an F1 score of 0.572. These results notably surpass those of RNA-FM (ACC: 0.600, F1: 0.458) and RNA-MSM (ACC: 0.578, F1: 0.440). The detailed results in Table 1 highlight distinct performance variations across the eleven RNA classes. RMSAGen consistently outperforms both RNA-FM and RNA-MSM in every RNA family, demonstrating significant improvements in classification accuracy and F1 scores. Particularly striking are the

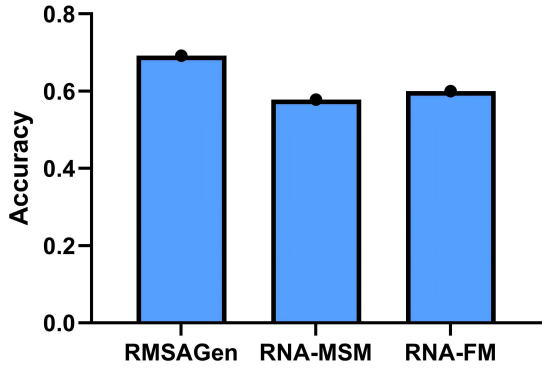

Figure 2: Comparison of Accuracy results for different methods on the classification tasks across eleven classes. The average quantitative results of Accuracy are separately annotated corresponding to each method.

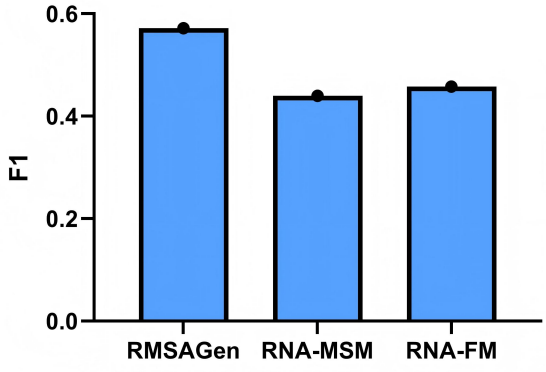

Figure 3: Comparison of F1 results for different methods on the classification tasks across eleven classes. The average quantitative results of F1 are separately annotated corresponding to each method.

results for U2 spliceosomal RNA (RF00004), where RMSAGen achieves an impressive ACC of 0.795 and F1 of 0.654, demonstrating superior embedding information extraction capabilities. The consistent superiority across diverse RNA families underscores the effectiveness of the RMSA-Encoder component in capturing distinctive molecular features, suggesting its potential for effectively distinguishing between different RNA families.

By predicting the function of RNA MSAs, we can determine that the RMSA-Encoder component of RMSAGen extracts MSA features more accurately compared to RNA-FM and RNA-MSM, which provides MSA feature representations for subsequent RNA design and generation.

Prediction of RNA Solvent Accessibility

In the RNA solvent accessibility prediction task (Table 2), RMSAGen consistently achieves the best performance with respect to R-squared, MAE, and RMSE. Compared to RNA-FM and RNA-MSM, RMSAGen attains an R-squared value of 0.6512, indicating that it more accurately captures structural information. Furthermore, its MAE and RMSE scores are 0.1463 and 0.3793, respectively, surpassing the other models and demonstrating higher predictive precision for solvent accessibility. In conjunction with the results of functional classification tasks, these findings suggest that RMSAGen effectively learns and generalizes key evolutionary and structural features from MSAs.

Data	RNA-FM		RNA-MSM		RMSAGen	
	ACC	F1	ACC	F1	ACC	F1
RF00001	0.58	0.418	0.558	0.407	**0.673**	**0.523**
RF00004	0.7	0.523	0.678	0.505	**0.795**	**0.654**
RF00005	0.635	0.467	0.613	0.45	**0.736**	**0.584**
RF00020	0.575	0.429	0.553	0.415	**0.677**	**0.536**
RF00026	0.605	0.453	0.583	0.435	**0.702**	**0.566**
RF00028	0.615	0.498	0.593	0.48	**0.711**	**0.623**
RF00050	0.53	0.398	0.508	0.38	**0.628**	**0.498**
RF00167	0.545	0.402	0.523	0.385	**0.642**	**0.503**
RF00234	0.615	0.45	0.593	0.43	**0.711**	**0.562**
RF01727	0.59	0.433	0.568	0.415	**0.688**	**0.541**
RF02541	0.525	0.37	0.503	0.35	**0.623**	**0.462**

Table 1: The quantitative results of classification tasks on different methods validate the effectiveness of RMSAGen (RMSA-Encoder) in extracting embedding information across eleven RNA classes, with ACC and F1 metrics. The **best** performance is highlighted.

Models	**R-squared ↑**	**MAE ↓**	**RMSE ↓**
RNA-FM	0.5795	0.1491	0.4165
RNA-MSM	0.6269	0.1608	0.3923
RMSAGen	**0.6512**	**0.1463**	**0.3793**

Table 2: The quantitative results of RNA solvent accessibility prediction task on different methods validate the effectiveness of RMSAGen (RMSA-Encoder) in extracting embedding information, with R-squared, MAE and RMSE metrics. The **best** performance is highlighted.

RNA design using RNA-binding protein

RMSAGen demonstrates exceptional capabilities in designing RNA sequences based on RBP, as validated through multiple complementary analyses on the RNAcompete data.

First, on the RNAcompete dataset, the AUPRC scores of sequences generated by RMSAGen show a low correlation with those from the Random and GA methods. This suggests that RMSAGen explores a sequence space distinct from random mutations or genetic algorithms.

Experiments on this dataset also show that as the MSA depth increases from 1, 10, 50, to 128, the AUC value steadily improves, highlighting the beneficial role of leveraging MSA information for enhancing RMSAGen's performance. Collectively, the analyses on the RNAcompete dataset underscore the model's strength. The low AUPRC

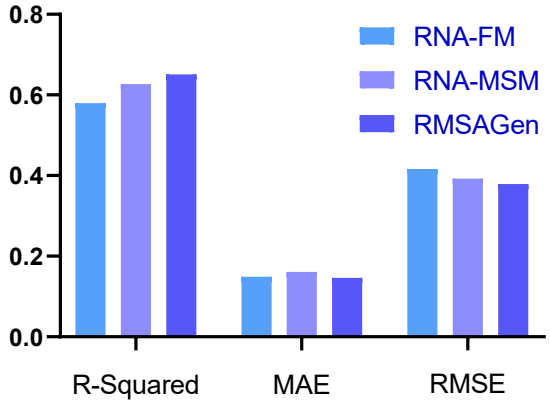

Figure 4: In the solvent accessibility prediction task, the result of the RMSA-Encoder.

correlation with Random and GA methods confirms the uniqueness of the generated sequences, while the steady increase in AUC with MSA depth, from 1 to 128 sequences, decisively shows that leveraging more evolutionary information enhances design quality. These results strongly support that RMSAGen is a effective approach for designing functional RNA sequences.

Overall, the results indicate that the RMSAGen method is a effective approach for designing RNA sequences that can effectively bind to target RBPs. Using MSA data further enhances the performance of RMSAGen, showcasing the importance of leveraging evolutionary information in the RNA design process.

Ribozyme Design

We further validate RMSAGen through computational and biological verification in ribozyme design.

Results of Computational Experiment. RMSAGen generates the ribozymes (hammerhead). Unlike previous methods, we first obtain information on aligned secondary structures of the type of ribozyme of Rfam (RF00163), processed through R-scape to acquire the MSA. We then integrate this information with the wild type structural data of the three ribozymes to generate sequences (for the fusion structure method). We observe that the designed sequences exhibit a high degree of secondary structural similarity, with RNAfold (Lorenz et al. 2011) predicting structures that match the wild type with a similarity score of 1.00. In addition, the tertiary structures predicted by Alphafold3 show an average Root Mean Square Deviation (RMSD) from the wild type of 5.71Å, indicating that the designed ribozymes closely resemble the wild type structures. Notably, the sequence similarity between the RMSAGen-designed sequences and the wild type averages at 52.37%, further demonstrating that RMSAGen effectively designs RNA with excellent computational metrics.

Results of Biological Experiment Hammerhead RNAs (RF00163) generated by RMSAGen not only achieve excellent performance in computational evaluations, but also

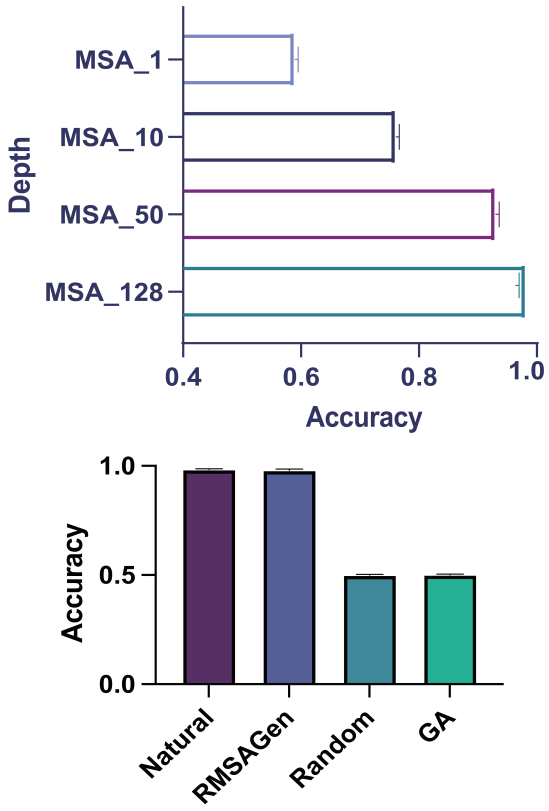

Figure 5: Performance evaluation of different methods on the RNAcompete data.

exhibit clear catalytic activity in biological experiments. To assess their functionality, biological validation was conducted as follows: Single-stranded DNA oligos and T7 transcription-related enzymes were sourced from a commercial supplier. All chemicals used were of reagent grade and applied without further purification.

Transcription reactions were performed in a 20 μL system containing standard T7 transcription buffer, 1mM DTT, 1μL T7 RNA polymerase (50U/μL), 1μM T7 promoter oligo, and 1 μM template oligo comprising the complementary sequence of the hammerhead RNA and the T7 promoter binding site. Additionally, 0.5 μL RNase inhibitor was included to prevent RNA degradation. The reaction mixtures were incubated at 37 °C for 16 hours to allow in vitro transcription. Following transcription, 1 μL DNase-I was added and the mixtures were incubated for 1 hour to ensure complete digestion of DNA templates. Transcription products were analyzed directly by denaturing polyacrylamide gel electrophoresis (urea-PAGE), using a 10% polyacrylamide gel containing 8 M urea in 1x TBE buffer. After electrophoresis, gels were imaged using an Amersham Typhoon Scanner (GE Healthcare), and band intensities were quantified with ImageJ software. The results (Figure 6c), demonstrate that the hammerheads generated by RMSAGen are catalytically active, as evidenced by the presence of cleavage products in the electrophoresis gel images. This confirms that the designed sequences are not only computationally promising,

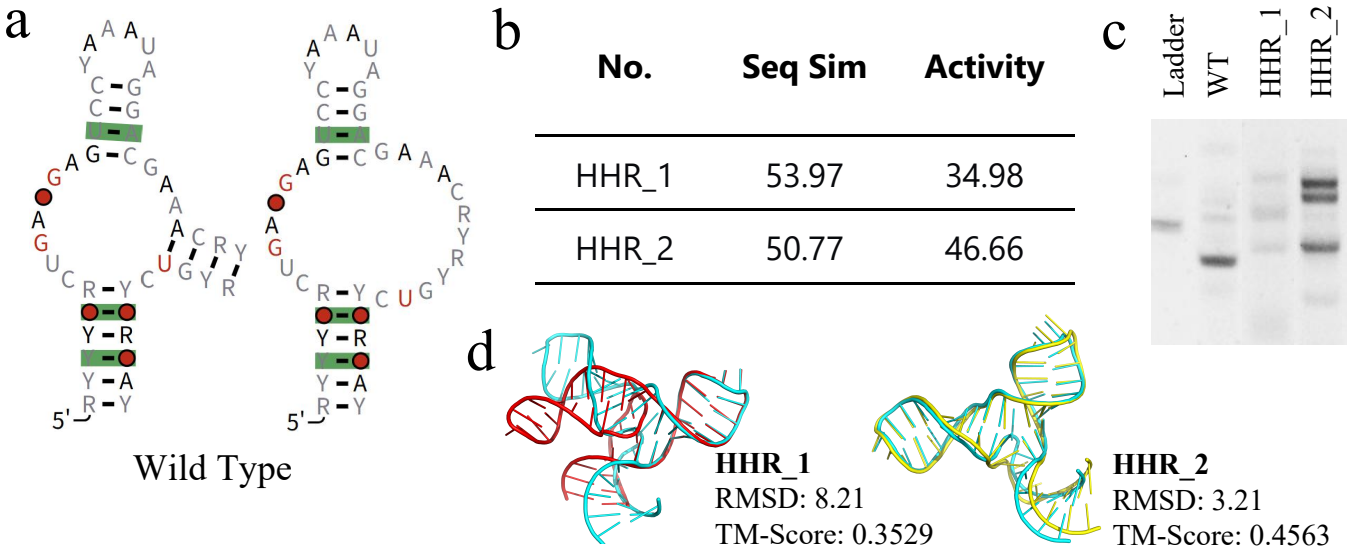

Figure 6: Results of Hammerhead design. **a** is secondary structure of the wild type (The structural image is from Rfam). **b** is sequence information. **c** is biological gel electrophoresis image. **d** is tertiary structure of the designed sequence aligned with the wild type.

but also function effectively in biological settings.

Related Works

Multiple Sequence Alignment Modeling

In the field of bioinformatics, multiple sequence alignment (MSA) emerges as a powerful computational approach for understanding evolutionary relationships and functional characteristics of biological sequences (Cozzetto et al. 2016). Previous research predominantly focuses on protein sequence analysis (Ovchinnikov et al. 2017; Wang et al. 2019), demonstrating the critical importance of conserved genetic information in deciphering sequence features and functional mechanisms. Notably, advanced computational methods such as AlphaFold2 (Jumper et al. 2021), AlphaFold3[2] and DeepMSA2 (Zheng et al. 2024) leverage MSA to revolutionize protein structure prediction and functional annotation, showcasing the technique's potential for extracting meaningful evolutionary signals.

RNA Design

RNA design is of great importance in modern biomedicine and drug discovery (Mortimer, Kidwell, and Doudna 2014; Kulkarni et al. 2021), but RNA sequences present a uniquely challenging area due to their inherent flexibility and structural complexity (Xu et al. 2022; Liu et al. 2022). MSA provides additional information that aids in protein modeling, suggesting that it could also benefit RNA modeling. Currently, approaches like DeepFoldRNA (Pearce, Omenn, and Zhang 2022), trRosettaRNA (Wang et al. 2023), and RhoFold+ (Shen et al. 2024b) utilize MSAs to characterize RNA, but these methods do not explore leveraging MSAs for direct sequence generation. Other methods rely on single-sequence models like DRFold2 (Li et al. 2023), leaving a critical gap in MSA-driven RNA sequence design. This gap highlights the need for an innovative method that can directly utilize MSA information for functional RNA sequence generation, bridging the current limitations in computational RNA design. Therefore, we propose RMSAGen, that leverages MSAs to design functional RNA sequences.

[2]https://github.com/google-deepmind/alphafold3

Conclusion and Outlook

RMSAGen combines MSA with an RNA generation model, offering an approach to the design of functional RNAs. This paper demonstrates, through a multidimensional evaluation of RNA family prediction, protein-binding RNA design, and ribozyme design, that RMSAGen can not only capture key MSA features but also design RNA sequences with biological activity. Ultimately, by bridging evolutionary insights with generative AI, RMSAGen provides a framework for overcoming the scarcity of functional RNA data, paving the way for more RNA engineering.

Looking ahead, we aim to scale RMSAGen to longer sequences for complex RNA machinery and optimize efficiency via advanced attention mechanisms. Furthermore, incorporating chemical modification data could significantly enhance the therapeutic stability and efficacy of the designed RNAs. We also envision extending this framework to co-design RNA-protein complexes, thereby expanding the toolkit for synthetic biology and next-generation RNA therapeutics. Additionally, integrating wet-lab feedback loops directly into the training process will be crucial for refining the model's predictive accuracy and accelerating the design-build-test cycle.

Acknowledgments

We want to thank our anonymous AC, SPC and reviewers for their feedback. This work was supported by the Major Project of Guangzhou National Laboratory (Grant No. GZNL2024A01003, GZNL2023A02007, GZNL2025C02028 to Jiao Yuan), National Natural Science Foundation of China (Grant No. 32400547 to Jiao Yuan), Pearl River Talent Recruitment Program (2023QN10Y296 to Jiao Yuan), Guangzhou Young Top Talent Program, National Key R&D Program of China (2023YFF1204701 to Jiao Yuan), the Chinese University of Hong Kong (CUHK; award numbers 4937025, 4937026, 5501517 and 5501329 to Yu Li), Shenzhen Medical Research Fund (Grant No. A2503002 to Yu Li), the IdeaBooster Fund (IDBF23ENG05 and IDBF24ENG06 to Yu Li), partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Hong Kong SAR), China (project no. CUHK 24204023 to Yu Li), a grant from the Innovation and Technology Commission of the Hong Kong SAR, China (project no. GHP/065/21SZ and ITS/247/23FP to Yu Li), and the Research Matching Grant Scheme at CUHK (award numbers 8601603 and 8601663 to Yu Li) from the Research Grants Council, Hong Kong SAR, China.

References

2019. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Research*, 47(D1): D221–D229.

Adamczyk, B.; Antczak, M.; and Szachniuk, M. 2022. RNAsolo: a repository of cleaned PDB-derived RNA 3D structures. *Bioinformatics*, 38(14): 3668–3670.

Chen, J.; Hu, Z.; Sun, S.; Tan, Q.; Wang, Y.; Yu, Q.; Zong, L.; Hong, L.; Xiao, J.; Shen, T.; et al. 2022. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*.

Cozzetto, D.; Minneci, F.; Currant, H.; and Jones, D. T. 2016. FFPred 3: feature-based function prediction for all Gene Ontology domains. *Scientific reports*, 6(1): 31865.

Danaee, P.; Rouches, M.; Wiley, M.; Deng, D.; Huang, L.; and Hendrix, D. 2018. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic acids research*, 46(11): 5381–5394.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. arXiv:1805.04833.

Ferruz, N.; Schmidt, S.; and Höcker, B. 2022a. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1): 4348.

Ferruz, N.; Schmidt, S.; and Höcker, B. 2022b. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1): 4348.

Huang, H.; Lin, Z.; He, D.; Hong, L.; and Li, Y. 2024. RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models. *Bioinformatics*, 40(Supplement_1): i347–i356.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Jiang, J.; Chen, P.; Wang, J.; He, D.; Wei, Z.; Hong, L.; Zong, L.; Wang, S.; Yu, Q.; Ma, Z.; et al. 2025a. Benchmarking large language models on multiple tasks in bioinformatics nlp with prompting. *arXiv preprint arXiv:2503.04013*.

Jiang, J.; Li, Y.; Cao, S.; Shan, Y.; Liu, Y.; Fei, T.; Yu, Y.; Feng, Y.; Li, Y.; Li, Y.; and Yuan, J. 2025b. Artificial intelligence in bioinformatics: a survey. *Briefings in Bioinformatics*, 26(6): bbaf576.

Jiang, J.; Xu, Y.; Wang, Z.; Ye, Y.; Shao, Y.; Shan, Y.; Wang, J.; Fan, X.; Yuan, J.; and Li, Y. 2025c. RBPtool: A Deep Language Model Framework for Multi-Resolution RBP-RNA Binding Prediction and RNA Molecule Design. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2170–2185. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Jing, B.; Eismann, S.; Suriana, P.; Townshend, R. J. L.; and Dror, R. 2021. Learning from Protein Structure with Geometric Vector Perceptrons. In *International Conference on Learning Representations*.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873): 583–589.

Katoh, K.; Misawa, K.; Kuma, K.-i.; and Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14): 3059–3066.

Kulkarni, J. A.; Witzigmann, D.; Thomson, S. B.; Chen, S.; Leavitt, B. R.; Cullis, P. R.; and van der Meel, R. 2021. The current landscape of nucleic acid therapeutics. *Nature nanotechnology*, 16(6): 630–643.

Li, Y.; Zhang, C.; Feng, C.; Pearce, R.; Lydia Freddolino, P.; and Zhang, Y. 2023. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nature Communications*, 14(1): 5745.

Liu, D.; Thélot, F. A.; Piccirilli, J. A.; Liao, M.; and Yin, P. 2022. Sub-3-Å cryo-EM structure of RNA enabled by engineered homomeric self-assembly. *Nature Methods*, 19(5): 576–585.

Lorenz, R.; Bernhart, S. H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; and Hofacker, I. L. 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology*, 6: 1–14.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8): 1099–1106.

Mortimer, S. A.; Kidwell, M. A.; and Doudna, J. A. 2014. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15(7): 469–479.

Nawrocki, E. P.; and Eddy, S. R. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22): 2933–2935.

Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P.-S.; Pavlopoulos, G. A.; Kim, D. E.; Kamisetty, H.; Kyrpides, N. C.; and Baker, D. 2017. Protein structure determination using metagenome sequence data. *Science*, 355(6322): 294–298.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pauling, L. 1951. Quantum theory and chemistry. *Science*, 113(2926): 92–94.

Pearce, R.; Omenn, G. S.; and Zhang, Y. 2022. De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. *bioRxiv*.

Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; and Rives, A. 2021. MSA Transformer. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8844–8856. PMLR.

Ray, D.; Ha, K. C.; Nie, K.; Zheng, H.; Hughes, T. R.; and Morris, Q. D. 2017. RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins. *Methods*, 118: 3–15.

Shen, T.; Hu, Z.; Sun, S.; Liu, D.; Wong, F.; Wang, J.; Chen, J.; Wang, Y.; Hong, L.; Xiao, J.; et al. 2024a. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 1–12.

Shen, T.; Hu, Z.; Sun, S.; Liu, D.; Wong, F.; Wang, J.; Chen, J.; Wang, Y.; Hong, L.; Xiao, J.; et al. 2024b. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 1–12.

Sun, L.; Xu, K.; Huang, W.; Yang, Y. T.; Li, P.; Tang, L.; Xiong, T.; and Zhang, Q. C. 2021. Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. *Cell research*, 31(5): 495–516.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Feng, C.; Han, R.; Wang, Z.; Ye, L.; Du, Z.; Wei, H.; Zhang, F.; Peng, Z.; and Yang, J. 2023. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nature Communications*, 14(1): 7266.

Wang, Y.; Shi, Q.; Yang, P.; Zhang, C.; Mortuza, S.; Xue, Z.; Ning, K.; and Zhang, Y. 2019. Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome biology*, 20: 1–14.

Wang, Z.; Wang, Z.; Jiang, J.; Chen, P.; Shi, X.; and Li, Y. 2025. Large language models in bioinformatics: A survey. *arXiv preprint arXiv:2503.04490*.

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

Xu, B.; Zhu, Y.; Cao, C.; Chen, H.; Jin, Q.; Li, G.; Ma, J.; Yang, S. L.; Zhao, J.; Zhu, J.; et al. 2022. Recent advances in RNA structurome. *Science China Life Sciences*, 65(7): 1285–1324.

Zhang, C.; Zhang, Y.; and Pyle, A. M. 2023. rMSA: A Sequence Search and Alignment Algorithm to Improve RNA Structure Modeling. *Journal of Molecular Biology*, 435(14): 167904. Computation Resources for Molecular Biology.

Zhang, Y.; Lang, M.; Jiang, J.; Gao, Z.; Xu, F.; Litfin, T.; Chen, K.; Singh, J.; Huang, X.; Song, G.; Tian, Y.; Zhan, J.; Chen, J.; and Zhou, Y. 2023. Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1): e3–e3.

Zheng, W.; Wuyun, Q.; Li, Y.; Zhang, C.; Freddolino, P. L.; and Zhang, Y. 2024. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nature Methods*, 21(2): 279–289.

Zhou, C.; Shan, Y.; Chen, P.; Shi, X.; Wang, Z.; Li, Y.; and Jiang, J. 2025. LM2Protein: A Structure-to-Token Protein Large Language Model. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 7023–7029. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.