Bridging Sequence-Structure Alignment in RNA Foundation Models

Heng Yang[1], **Renzhi Chen**[2], **Ke Li**[1*]

[1] Department of Computer Science, University of Exeter, EX4 4QF, Exeter, UK
[2] Qiyuan Lab, Beijing, China
{hy345, k.li}@exeter.ac.uk, {chengrenzhi1989}@gmail.com

Abstract

The alignment between RNA sequences and structures in foundation models (FMs) has yet to be thoroughly investigated. Existing FMs have struggled to establish sequence-structure alignment, hindering the free flow of genomic information between RNA sequences and structures. In this study, we introduce `OmniGenome`, an RNA FM trained to align RNA sequences with respect to secondary structures based on structure-contextualised modelling. The alignment enables free and bidirectional mappings between sequences and structures by utilising the flexible RNA modelling paradigm that supports versatile input and output modalities, i.e., sequence and/or structure as input/output. We implement RNA design and zero-shot secondary structure prediction as case studies to evaluate the Seq2Str and Str2Seq mapping capacity of `OmniGenome`. Results on the EternaV2 benchmark show that `OmniGenome` solved 74% of puzzles, whereas existing FMs only solved up to 3% of the puzzles due to the oversight of sequence-structure alignment. We leverage four comprehensive *in-silico* genome modelling benchmarks to evaluate performance across a diverse set of genome downstream tasks, where the results show that `OmniGenome` achieves state-of-the-art performance on RNA and DNA benchmarks, even without any training on DNA genomes.

Code — https://github.com/yangheng95/OmniGenBench
Datasets — https://huggingface.co/spaces/yangheng/OmniGenomeLeaderboard/tree/main/benchmarks
Extended version — https://arxiv.org/abs/2407.11242

1 Introduction

RNA is a critical type of molecule that encodes a vast array of biological regulatory elements that orchestrate crucial aspects of plant growth, development, and adaptation to environmental stresses. To decipher the genomic code in RNA and manipulate RNA engineering and design, current research mainly uses bioinformatics in solving RNA genome-oriented challenges. Recent advancements in large-scale pre-trained foundation models (FMs) have demonstrated their unprecedented potential to back up existing genome analysis, as FMs are capable of learning and predicting the complex 'genomic language' (Nguyen et al. 2023) hidden in genome encoding processes. Existing FMs have been widely employed as basic sequence feature extractors to improve the performance of diverse genome analysis tasks, such as secondary structure prediction (Tan et al. 2017; Danaee et al. 2018; Mathews 2019; Kalvari et al. 2021), degradation rate prediction (Yaish and Orenstein 2022; Wayment-Steele et al. 2022), and mRNA vaccine design (Corbett et al. 2020; Runge et al. 2023). In RNA, it is intriguing that the functionality and stability are intertwined with its complex structures in molecular biology (Ganser et al. 2019). However, the role of the structure as a second 'genomic language' to interact with sequences and solve various RNA downstream tasks has been largely ignored.

Sequence-Structure Alignment in GFMs We define alignment between sequences and secondary structures[1] as the bidirectional information flows. Current FMs have been struggling to establish an alignment between RNA nucleotide sequences and their folded structures, thus impeding bidirectional genomic information flows. There has been a deep scientific challenge to align RNA sequences with structures because it is not deterministic to predict sequences from structures and vice versa. In other words, an identical sequence may be folded into different sub-optimal structures because the folding patterns of RNA sequences depend on various *in-vivo* factors (Tinoco Jr and Bustamante 1999). Further, a structure can be folded from different sequences composed of variational combinations of nucleotide bases. The oversight of such alignment in existing FMs causes outstanding issues in understanding and leveraging RNA structures, such as mRNA design. For example, recent state-of-the-art RNA FMs, RNA-FM (Chen et al. 2022) and RNA-MSM (Zhang et al. 2024), only solved 3 out of 100 puzzles in *in-silico* RNA design (Lee et al. 2014). This is because they fail to decipher corresponding sequences based on structures to guide RNA design.

To address the above two problems, we propose sequence-structure alignment in RNA FMs, which leverages the large-scale annotations of sequences and structures to build reliable structure to sequence (Str2Seq) and sequence to structure (Seq2Str) mappings, leading to an *aligned* FM dubbed `OmniGenome`. The sequence-structure alignment

*Corresponding Author

[1] In this paper, all structures are referred to as the secondary structures.

enables genomic information to freely flow between sequences and structures by introducing a flexible RNA modelling paradigm that supports versatile inputs and outputs modalities, i.e., sequence and/or structure as input/output. The sequence-structure alignment enables genomics information to freely flow between sequences and structures by introducing a flexible RNA modelling paradigm that supports versatile inputs and outputs modalities, i.e., sequence and/or structure as input/output. Furthermore, the sequence-structure alignment is designed to be architecture-agnostic and genome-agnostic. That is to say, it can be easily transferred to large-scale models with new architecture and different genome types like DNA.

Str2Seq Mapping RNA structure serves as a vital input in most of the RNA genome analysis tasks. To induce the ability of Str2Seq mapping in genomic FMs, we formulate a structure-contextualised RNA sequence reconstruction task, which stems from the representation of RNA secondary structures in texts composed of dots and brackets. We first concatenate sequence-structure pairs as inputs and then mask a small portion of nucleotide bases in the sequence. Then, we pre-train `OmniGenome` to reconstruct the masked nucleotide bases given the structure contexts. This simple but effective formulation of Str2Seq mapping realises structure input awareness in genomics FM pre-training and provides substantial compatibility for structure-contextualised tasks, which has been verified in our RNA design benchmark.

Seq2Str Mapping On the other hand, Seq2Str mapping, such as end-to-end secondary structure prediction (SSP) (Sato, Akiyama, and Sakakibara 2021; Fu et al. 2022), is another critical aspect of achieving the alignment. We generalise end-to-end structure pre-training (Yan, Hamilton, and Blanchette 2022) to `OmniGenome` pre-training. This large-scale structure pre-training on diversified genomes supervises `OmniGenome` to perform Seq2Str mapping. The problem of structure pre-training lies in RNA structure annotation scarcity, which leads to biased structure predictions (Chen et al. 2020) and barriers the structure prediction robustness on small datasets. To conduct Seq2Str mapping, tremendous secondary structure annotations are required to avoid data bias. A feasible solution to RNA structure pre-training is leveraging the plausible structures calculated based on the minimum free energy. In this paper, we leverage the popular ViennaRNA (Lorenz et al. 2011) to serve our purpose, 'computing' the structures for millions of RNA sequences and perform structure pre-training in `OmniGenome`.

Evaluations and Results To validate the effectiveness of `OmniGenome`, we designed four large-scale genome benchmarks with diverse genomics tasks. The first one is the RNA genomics benchmark (RGB) compiled in the study, which contains diverse challenging genomics understanding tasks that benefit from the sequence-structure alignment, such as degradation rate prediction. The second benchmark is the plant genomics benchmark (PGB) (Mendoza-Revilla et al. 2023) which contains millions of DNA sequences to evaluate the DNA sequence understanding tasks. In particular, we want to use this benchmark to evaluate the generalisability of `OmniGenome` among diversified species and genomes. The overall performance of `OmniGenome` (up to 186M parameters) on both two benchmarks consistently outperforms existing genomics FMs with up to 35% improvement, even compared with Agro-NT (Mendoza-Revilla et al. 2023) which contains 1 billion parameters. The last two benchmarks, available in the Appendix[2], are the genomics benchmark (GB) (Grešová et al. 2023) and genomics understanding evaluation (GUE) (Zhou et al. 2023), which serve as two additional DNA benchmarks to evaluate generalisability on non-plant genome modelling.

In addition, we also conduct zero-shot Seq2Str and Str2Seq prediction experiments to verify the performance of sequence-structure alignment. As revealed in the experiments in Sections 3.4 and 3.4, `OmniGenome` achieves up to a 74.85% macro-F1 score in zero-shot Seq2Str prediction, i.e., secondary structure prediction, outperforming fine-tuned FMs and bioinformatics methods like ViennaRNA. In terms of Str2Seq prediction performance, we evaluate the performance of `OmniGenome` in the *in-silico* RNA design task. We solved 74% of complex puzzles of the EternaV2 benchmark (Lee et al. 2014), while state-of-the-art FMs such as RNA-MSM and RNA-FM only solved up to 3%. Besides, `OmniGenome` only takes less than one hour to solve most of the puzzles, while most RNA design methods need to take up to 24 hours to solve even a single puzzle.

Open-source Toolkit and Tutorials Open science is always the golden standard to promote this rising area of FM for genome modelling, which unfortunately lacks relevant high-quality resources such as code integrity, data availability, and pre-training pipeline. To address this gap, following the FAIR principles (Wilkinson et al. 2016), we developed an open-source package[3] that includes step-by-step tutorials for FM pre-training and downstream tasks fine-tuning, to name a few. It provides ready-to-use genomics benchmarks and uses the API with only a few lines of code to streamline benchmarking purposes. We believe this will be a valuable resource to make this emerging AI for the RNA community to thrive.

2 Methodology

This section delineates the implementation details of `OmniGenome` including its entire pre-training workflow and downstream benchmarks.

2.1 RNA Tokenization for Alignment

We aim to implement a fine-grained alignment between RNA sequences and structures, where each base in the sequences reflects a structural label in { '`(`', '`)`', '`.`'}. Therefore, we propose an adapted implementation of the single nucleotide tokenization (SNT) method (Nguyen et al. 2023; Chen et al. 2023) in `OmniGenome`, where the whole vocabulary, {'`A`', '`T`', '`C`', '`G`', '`U`', '`N`', '`(`', '`)`', '`.`'}, contains the

[2]Please find the Appendix in https://arxiv.org/abs/2407.11242

[3]https://github.com/yangheng95/OmniGenBench

nucleotide-level structural labels. We illustrate our tokenization based on an example shown in the extended version.

Our adapted SNT features bidirectional mappings between single nucleotide (SN) bases and structural labels required by sequence-structure alignment. Another reason for the adaption of SNT is that, in the realm of RNA genome modelling, the FM performance highly depends on the tokenization resolution (Nguyen et al. 2023; Chen et al. 2023). For example, the k-mers (Yang et al. 2023; Dalla-Torre et al. 2023) and BPE (Devlin et al. 2019; Zhou et al. 2023) tokenization methods combine multiple bases into tokens and embeddings, which compromise modelling resolution and thus fail to the solution of fine-grained genomic tasks like structure prediction as well as base-level degrade rate prediction. Like other encoder-only models, e..g, BERT (Devlin et al. 2019), we incorporated special tokens, e.g., '<mask>', to implement masked language modelling.

2.2 Pre-training Objectives

As discussed in Section 1, a key desideratum for SN-level genome modelling is to build the alignment between RNA sequences with corresponding secondary structures. Bearing this in mind, we formulate two pre-training objectives, i.e., $\mathcal{L}_{\texttt{Str2Seq}}$ and $\mathcal{L}_{\texttt{Seq2Str}}$, for Str2Seq and Seq2Str predictions, respectively. Besides, we aggregate these two objectives with the masked RNA language modelling objective MRLM to pre-train OmniGenome as follows:

$$\mathcal{L}_{\texttt{pre-train}} = \mathcal{L}_{\texttt{Str2Seq}} + \mathcal{L}_{\texttt{Seq2Str}} + \mathcal{L}_{\texttt{MRLM}} + \lambda||\theta||_2, \quad (1)$$

where λ is the ℓ_2 regularisation weight and θ represents the parameters of OmniGenome. The following paragraphs explain the design principles of each objective function used in equation (1).

- $\mathcal{L}_{\texttt{Str2Seq}}$ is designed to enable OmniGenome to predict bases given structure-contextualised sequences with partially masked bases. This objective aims at Str2Seq tasks and teaches OmniGenome to interpret structure information and infer the masked sequences. To achieve this objective, we mask 15% of the bases and structure tokens, encouraging the model to infer masked bases (i.e., {'A', 'T', 'C', 'G', 'U', 'N'}) and structure tokens (i.e., {'(', ')', '.'}). Specifically, $\mathcal{L}_{\texttt{Str2Seq}}$ is defined as the classic cross-entropy loss widely used in the masked language modelling:

$$\mathcal{L}_{\texttt{Str2Seq}} = -\frac{1}{|m|}\sum_{i=1}^{m} \log p(x_i \mid x_{\setminus i}), \quad (2)$$

  where m is the number of masked nucleotide and structure tokens, and $p(x_i|x_{\setminus i})$ indicates the probability of predicting the masked nucleotide x_i based on its context.
- In terms of structure-out modelling, we implement $\mathcal{L}_{\texttt{Seq2Str}}$ to enable OmniGenome for Seq2Str predictions. Instead of directly feeding the secondary structure into OmniGenome as inputs, this objective employs the RNA secondary structures as labels for supervised training. This objective is implemented as a token-level classification, where the $\mathcal{L}_{\texttt{Seq2Str}}$ loss is defined in the following cross-entropy loss:

$$\mathcal{L}_{\texttt{Seq2Str}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} s_{ic} \log(\hat{s}_{ic}), \quad (3)$$

  where s_{ic} denotes the label c of secondary structure at the i-th position, and $\hat{s}_{ic}$ is the probability predicted by a linear classifier deployed on OmniGenome. N is the length of an RNA sequence and $C = 3$ denotes the number of the possible labels of structure, i.e., {'(', ')', '.'}.
- The last objective, $\mathcal{L}_{\texttt{MRLM}}$, is adapted to the conventional masked language modelling loss in NLP. It aims to improve the model's understanding of genomic language in RNA sequences by predicting the masked or replaced 5% of nucleotide bases. The definition of $\mathcal{L}_{\texttt{MRLM}}$ is similar to that of $\mathcal{L}_{\texttt{Str2Seq}}$ which only considers the prediction of masked bases rather than randomly replaced bases. The loss function of MRLM is well-known so we omit its formula here.

We cannot trust structure predictions (in $\mathcal{L}_{\texttt{Seq2Str}}$) while the structures are leaked in inputs (in $\mathcal{L}_{\texttt{Str2Seq}}$), i.e., the sequence inputs and outputs of these two objectives are exclusive. In practice, we only consider objectives either $\mathcal{L}_{\texttt{Seq2Str}} + \mathcal{L}_{\texttt{MRLM}}$ or $\mathcal{L}_{\texttt{Str2Seq}} + \mathcal{L}_{\texttt{MRLM}}$ for each input sequence. In the pre-training, 70% of RNA sequences are used for the first two objectives, while the remaining 30% are used for the latter two objectives. This proportion setting is concluded from our empirical experience to balance the capability of Str2Seq and Seq2Str predictions.

2.3 Model Architecture

OmniGenome adopts the Transformer encoder architecture with bidirectional multi-head attention. We do not adopt recent architectures like Mamba (Gu and Dao 2023; Schiff et al. 2024) and Hyena (Nguyen et al. 2023) because our experiments in Table 4 and Table 5 show that these architectures are not competent at RNA genome understanding. This low performance is probably because RNA sequences are much shorter than DNA sequences in the wild.

We designed two variants, dubbed $\texttt{OmniGenome}^{52\texttt{M}}$ and $\texttt{OmniGenome}^{186\texttt{M}}$ with 52 and 186 million parameters respectively. Some key model specifications are summarised in Table 1.

To improve the reproducibility of OmniGenome, we list the pre-training settings and hyperparameters as follows.

- The learning rate is set to 5×10^{-5} and the weight decay is set to 0.01.
- We use AdamW as the optimiser with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.
- We use a linear decay strategy with a warm-up period of $1,000$ steps in the learning rate scheduler.
- The batch size is set to $2,048$.
- No dropout is applied during pre-training, and we use the rotary position embeddings (Su et al. 2024) to further enhance the model's scalability to long RNA sequences.

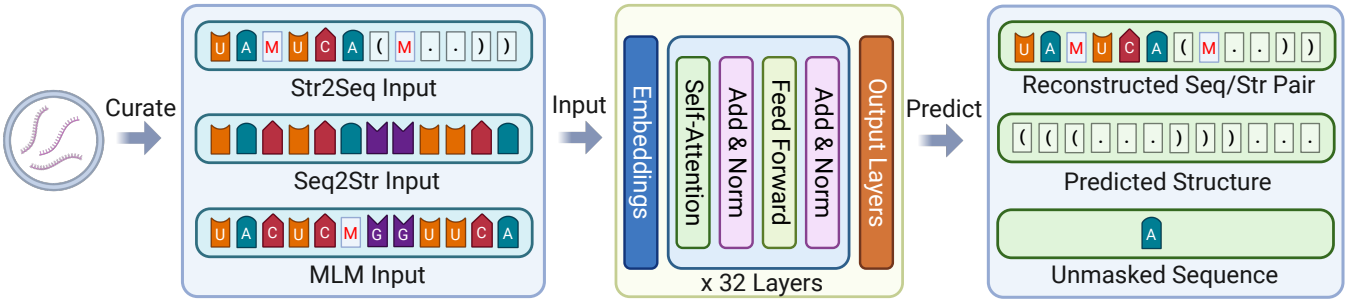

Figure 1: The workflow of OmniGenome pre-training. We craft the inputs for three pre-training objectives described in Section 2.2. The outputs are reconstructed sequences based on the context of structure, predicted secondary structure, and unmasked sequences, respectively. The predictions of shadowed tokens are not calculated in the objective functions.

OmniGenome	52M	186M
# of Layers	16	32
Embedding dimension	480	720
Intermediate dimension	2,400	2,560
# of heads	24	30
# of parameters	52M	186M
Modelling length	4,096*	

Table 1: Summary of some key model specifications of two OmniGenome variants. "*" means that we used a modelling length of 1024 in the pre-training, while the supports up to 4096 in downstream tasks.

- We built a distributed training environment with 8 NVIDIA RTX 4090 GPUs, while its configuration is introduced in the Appendix. The pre-training was finished in approximately 1 and 3 weeks for $\texttt{OmniGenome}^{52M}$ and $\texttt{OmniGenome}^{186M}$, respectively.

2.4 Pre-training Database: OneKP

Recent studies (Chen et al. 2023; Zhou et al. 2023) have shown that data diversity can enhance FM performance without significantly increasing model capacity. For the OmniGenome pre-training, we collected transcriptome data from the OneKP initiative[4] (Carpenter, Leebens-Mack, and et al. 2019), which compiles large-scale RNA raw sequence database from 1,124 plant species. The raw sequences are not available for pre-training before processing and filtering.

We adopt the following raw sequence data curation protocol to fit pre-training.

- To enhance training efficiency and reduce bias, we removed all duplicate sequences.
- To tackle incomplete transcriptome data and other noises, we discard sequences shorter than 50 bases.
- To facilitate the sequence-structure alignment training, we adopt ViennaRNA[5] to obtain the secondary structures for the sequences.
- We use "cd-hit-est" (Li and Godzik 2006) and blast (Altschul et al. 1990) tools to filter the sequences in downstream tasks with similar structures. Please refer to the experiment section for more details.

2.5 Benchmark Suites

OmniGenome is designed as a general-purpose RNA FM that can be fine-tuned for a diverse set of downstream genomics predictive tasks. In this paper, we constructed a large-scale benchmark suite for RNA FMs. According to the category of genomes, we split the benchmark into two parts, i.e, RNA Genomic Benchmark (RGB) and Plant Genomic Benchmark (PGB). Please refer to the appendix for the benchmark details.

3 Experiments

To evaluate the performance of OmniGenome across genome modelling, we implement experiments on diverse downstream tasks. We first evaluate the sequence-structure alignment capability of OmniGenome. Subsequently, we evaluate the overall performance of OmniGenome on two comprehensive genomic modelling benchmarks, i.e., RGB and the PGB, respectively. Finally, we include the GB and GUE in the appendix to evaluate the performance on non-plant genomes.

3.1 RNA Sequence Filtering

The pertaining involves RNA sequences and structures prediction, we take the data and annotation leakage problem seriously.

- To avoid structure annotation leakage of downstream benchmarks, the secondary structure predictors for all FMs were randomly initialised for fair comparisons, which means the pre-trained structure predictor of OmniGenome was not used in benchmarks, except for zero-shot SSP experiments. Please find the source codes for details.
- To reduce sequence leakage caused by evolutionary conservative sequences across multiple species, we use the ch-hit-est tool to calculate the sequence similarity between sequences from the OneKP database and downstream tasks. We adopt the similarity threshold of 80%

[4]https://sites.google.com/a/ualberta.ca/onekp/
[5]https://github.com/ViennaRNA/ViennaRNA

for ch-hit-est to eliminate sequences whose homogeneous sequences appeared in the OneKP database. Subsequently, we exploit the blastn tool to query potentially leaked sequences in downstream benchmark datasets and further alleviate the data leakage problem. The e-value has been set to 1 for rigorous sequence filtering.

3.2 Pre-training and Evaluation Environment

The pre-training of OmniGenome was conducted on a dedicated Linux computation node, equipped with 8 NVIDIA RTX 4090 GPUs. For distributed model training, we employed version 4.37.1 of the Transformers library alongside version 0.26.1 of the Accelerate library. Our implementation framework of choice for OmniGenome was PyTorch, specifically version 2.0.0. The ViennaRNA version is 2.6.4 in our experiments. While some existing code was adapted for the modules within OmniGenome, the majority of the codebase, such as genomic sequences preprocessing, model pre-training, objective functions, and experiments, was meticulously crafted from scratch.

3.3 Comparison Baselines

Apart from OmniGenome, we implement a plus variant, i.e., OmniGenome+. In the context of OmniGenome+, we assume the structure annotation from ViennaRNA is always available for enhancing the model based on structure-contextualised modelling. In SSP tasks, we can also use the ViennaRNA's structure annotations as contexts to improve downstream SSP performance. Please refer to the Appendix for brief introductions of these FMs.

We can compare OmniGenome with the following RNA and DNA FMs shown in the Appendix as baselines to help evaluate the performance of OmniGenome. We are aware that some FMs are also developed for RNA, such as UniRNA (Wang et al. 2023), 5UTR-LM (Chu et al. 2024), etc. However, we cannot compare OmniGenome with them because their source codes are very hard to work with in our efforts or not publicly available.

3.4 Sequence-Structure Alignment Evaluation

In this section, we verify the sequence-structure alignment capability based on two experiments, i.e., Str2Seq prediction and zero-shot Seq2Str prediction via SSP and RNA design tasks, respectively. Overall, the results in Table 2 and Table 3 provide reliable evaluations of the FMs' capabilities in sequence-structure alignment. This underscores OmniGenome's efficacy in enabling genomic information to freely flow between structures and sequences.

RNA Design (Str2Seq) Evaluation we demonstrate the Str2Seq prediction capability of OmniGenome based on RNA design. We employed the Eterna (Lee et al. 2014) V2 benchmark, which consists of 100 specified secondary structures. This task aims to design RNA sequences based on reference structures. We develop a genetic algorithm (GA) which exploits masked nucleotide modelling (a.k.a., masked language modelling) to find plausible RNA sequences that solve RNA design puzzles. The implementation details can be found in the Appendix. In the GA, the population size is set at 1000, with 100 iterations, and the mutation rate for each base is 0.5. The evaluation metric is accuracy following existing works which indicates the number of puzzles solved by FMs. The experimental results are available in Table 2.

Model	**Token.**	**EternaV2 (Acc)**
RNAInverse	—	30
3UTRBERT	k-mers	0
DNABERT2	BPE	0
SpliceBERT	SNT	3
RNA-MSM	SNT	2
RNA-FM	SNT	3
OmniGenome^{52M}+	SNT	71
OmniGenome^{186M}+	SNT	**74**

Table 2: Performance on the EternaV2 RNA design benchmark. The best accuracy is in **bold** face. "Token." indicates the tokenization method.

Model	**Zero-shot SSP (F1)**		
	Archive2	**bpRNA**	**Stralign**
ViennaRNA	73.99	65.04	74.09
OmniGenome^{52M}	69.93	65.85	74.71
OmniGenome^{186M}	74.38	66.19	74.91
OmniGenome^{52M}+	73.58	65.95	75.16
OmniGenome^{186M}+	**74.72**	**66.37**	**75.80**

Table 3: Performance in zero-shot SSP. The results are based on zero-shot inferences without any fine-tuning or domain adaptation. "Stralign" denotes the RNAStralign dataset.

We include a popular baseline of RNAInverse and select recent DNA and RNA FMs which support masked language modelling. We exclude HyenaDNA in this experiment because it does not support masked nucleotide prediction. It is observed from Table 2 that RNAInverse solved 30 of the RNA design puzzles, indicating a promising capability in RNA design. The FMs, such as 3UTRBERT and DNABERT2 fail in RNA design because they cannot handle SN-level modelling. Meanwhile, RNA-MSM, RNA-FM and SpliceBERT demonstrated trivial proficiency in RNA design, solving 2 to 3 puzzles. This observation suggests these FMs cannot precisely predict the bases without any Str2Seq prediction ability. With the help of Str2Seq, i.e., structure-contextualised sequence reconstruction, OmniGenome^{52M}+ and OmniGenome^{186M}+ significantly outperformed other FMs with 71 and 74 puzzles solved, respectively, underscoring the significance of Str2Seq in sequence-structure alignment. Besides, we expect an increase in performance with sufficient computational budgets and the findings provide crucial evidence of the significance of Str2Seq for RNA sequence design.

Zero-shot SSP (Seq2Str) Evaluation This subsection evaluates both Seq2Str and Str2Seq prediction in sequence-

Model	PolyA	LncRNA	Chrom Acc	Prom Str	Term Str	Splice	Gene Exp	Enhancer
	F1	F1	F1	RMSE	RMSE	F1	RMSE	F1
DNABERT2	41.35	72.55	61.49	0.99	0.24	45.34	14.78	36.40
HyenaDNA	83.11	58.21	52.20	0.88	0.26	90.28	14.79	66.17
Caduceus	70.89	68.40	64.53	0.91	0.26	78.51	14.72	60.83
NT-V2	71.26	73.08	65.71	0.81	0.27	95.05	14.79	73.89
Agro-NT	78.89	67.24	63.27	0.94	0.78	88.45	15.56	62.83
SpliceBERT	65.23	71.88	63.62	0.75	0.22	96.45	14.70	69.71
3UTRBERT	76.48	70.75	63.71	1.04	0.36	94.44	14.87	71.67
RNA-BERT	78.54	61.99	48.94	1.81	0.38	94.45	14.89	57.61
RNA-MSM	84.25	67.49	53.52	1.28	0.28	95.49	14.87	61.45
RNA-FM	84.94	68.75	54.92	0.95	0.27	95.95	14.83	57.14
OmniGenome^{52M}	85.47	75.71	64.23	0.67	0.21	97.40	14.76	68.31
OmniGenome^{186M}	86.87	77.53	66.88	0.65	0.19	98.15	14.76	72.45
OmniGenome^{52M}+	87.05	76.23	65.41	0.65	0.20	97.70	14.76	70.71
OmniGenome^{186M}+	**87.55**	**77.96**	**67.69**	**0.59**	**0.18**	**98.41**	**14.71**	**79.77**

Table 4: Performance of OmniGenome and baseline FMs on PGB. "PolyA" stands for Polyadenylation, "Chrom Acc" for Chromatin Accessibility, "Prom Str" for Promoter Strength, "Term Str" for Terminator Strength, "Splice" for Splice Site, "Gene Exp" for Gene Expression, and "Enh Reg" for Enhancer Region. Results for OmniGenome^{186M}+ are excluded due to the time-intensive nature of the experiments.

structure alignment. The evaluation of Seq2Str is based on zero-shot SSP. We use OmniGenome and OmniGenome+ without fine-tuning to predict the secondary structures of sequences from the testing datasets and measure the macro-F1 score, where better structure prediction performance indicates a stronger capability for Seq2Str prediction. The experimental results are available in Table 3.

The results in Table 3 indicate that OmniGenome FMs mirrored the zero-shot secondary structure prediction (i.e., Seq2Str) performance of ViennaRNA. Moreover, OmniGenome^{52M}+ and OmniGenome^{186M}+ outperform OmniGenome FMs based on structure contexts from ViennaRNA. Given the ablation of structure contexts, OmniGenome^{186M} also achieves performance comparable with ViennaRNA on the Archive2, bpRNA and RNAStralign datasets. Besides, we found that OmniGenome+ generally obtains better performance on a wide genome downstream tasks owing to the structure awareness, and random or noise structure contexts have no obvious effects on the structure prediction. We cannot compare with other FMs in the zero-shot SSP experiments, because existing FMs were not pertained for secondary structure prediction.

3.5 Results on RGB

The results in Table 5 demonstrate the performance of OmniGenome and its generalizability across various fine-grained RNA downstream tasks. It is observed that OmniGenome models achieve better results than both RNA and DNA FM baselines, including Agro-NT and DNABERT2, which contain hundreds of millions of parameters. This is because the existing FMs usually adopt k-mers or BPE tokenization that cannot handle SN resolution tasks, e.g., single nucleotide mutation detection and repair, and structure prediction. Because of the Seq2Str pretraining, OmniGenome and OmniGenome+ models exhibit strong results in secondary structure prediction, underscoring OmniGenome's capabilities in SN-level RNA sequence understanding and manipulation.

3.6 Results on PGB

The PGB is a plant-oriented genomic benchmark. Although the benchmark datasets in PGB are DNA-based tasks, we can still evaluate the performance of OmniGenome and its generalizability on multi-modal (i.e., DNA and RNA) genomic tasks. The results in Table 4 reveal substantial variability in the performance of different FMs, where OmniGenome^{52M} outperformed other baseline models across most tasks, particularly in tasks like Polyadenylation, Splice Site, and Enhancer Region classification, where they achieved the highest F1 scores. This suggests that OmniGenome's architecture is particularly adept at handling complex genomic sequences. In comparison, existing FMs, e.g., NT-V2 and Agro-NT, showed lower performance with more parameters than OmniGenome. Besides, the performance of OmniGenome^{52M}+ suggests that the structure context can further enhance the performance of genomic modelling. Overall, OmniGenome models achieve state-of-the-art performance on both benchmarks, especially for OmniGenome+ variants. The results underscore the importance of sequence-structure alignment in achieving complex genomic modelling tasks.

4 Related Works

Biological sequence modelling, including DNA, RNA, and protein, has attracted attention in recent years. Protein modelling, e.g., AlphaFold (Jumper et al. 2021; Evans et al.

Model	mRNA	SNMD	SNMR	Archive2	Stralign	bpRNA
	RMSE	AUC	F1	F1	F1	F1
ViennaRNA	N.A.	N.A.	N.A.	73.99	74.09	65.03
MXFold2	N.A.	N.A.	N.A.	90.09	97.01	64.99
Ufold	N.A.	N.A.	N.A.	89.78	95.76	78.38
DNABERT2	0.8158	49.94	15.86	55.73	64.09	33.77
HyenaDNA	0.8056	53.32	39.80	71.18	91.24	57.43
Caduceus	0.8026	57.01	39.59	74.37	92.28	59.76
NT-V2	0.7826	50.49	26.01	68.36	83.18	56.95
Agro-NT	0.7830	49.99	26.38	62.81	72.54	46.87
SpliceBERT	0.7340	58.11	46.44	79.89	93.81	71.59
3UTRBERT	0.7772	50.02	24.01	68.62	88.55	57.90
RNABERT	0.8087	51.32	29.14	24.66	83.68	47.96
RNA-MSM	0.7321	57.86	45.22	68.72	91.15	64.44
RNA-FM	0.7297	59.02	42.21	82.55	95.07	78.16
OmniGenome^{52M}	0.7191	62.44	49.91	88.48	97.46	80.51
OmniGenome^{186M}	0.7164	63.81	50.80	90.32	97.82	83.09
OmniGenome^{52M}+	0.7174	63.11	51.21	88.58	97.33	81.29
OmniGenome^{186M}+	**0.7121**	**64.13**	**52.44**	**91.89**	**98.21**	**83.18**

Table 5: The performance of OmniGenome and baseline models on the RGB, with results averaged based on five random seeds. "N.A." means not available for predictive tasks.

2021; Abramson et al. 2024) and ESM (Lin et al. 2022), has been studied for many years compared to DNA and RNA modelling. However, the RNA foundation model development has been struggling because the data scale and quality of RNA sequences are limited. Nevertheless, the RNA secondary structures are expensive to verify via in vivo experiments, leading to a grad challenge in the past to model the alignment between RNA sequences and structures.

Current RNA FMs focused on sequence-to-structure mapping, e.g., end-to-end secondary structure prediction. However, to the best of our knowledge, the sequence-structure alignment in RNA genome modelling has yet been investigated in the literature. There have been some preliminary works, such as scBERT (Yang et al. 2022), RNABERT (Akiyama and Sakakibara 2022), RNA-FM (Chen et al. 2022), RNA-MSM (Zhang et al. 2023), and RNAErnie (Wang et al. 2024), to name a few. However, these methods have only trained the FMs on a limited-scale database, as RNA sequences are generally expensive to obtain. Some FMs focus on specific types of RNA sequences, such as coding sequences (CDS) (Hallee, Rafailidis, and Gleghorn 2023), 5' untranslated regions (5'UTR) (Chu et al. 2024), 3' untranslated regions (3'UTR) (Yang et al. 2023), or precursor mRNA sequences (Chen et al. 2023), thus limiting the models' ability to capture the diversity of RNA sequences. Uni-RNA (Wang et al. 2023) has been reported to achieve good performance, however, it is not open-sourced and cannot be compared in the experiments.

In short, the existing RNA FMs neglect the significance of sequence-structure alignment in RNA genome modelling, while the 5UTR-LM (Chu et al. 2024) adopts the secondary structure prediction as a pre-training objective to achieve Seq2Str prediction in pre-training. However, these FMs are not available for Str2Seq mapping and suffer from limited model and data scales that fail to uncover the comprehensive efficacy of sequence-structure alignment on a wide set of genomic tasks. ERNIE-RNA (Yin et al. 2024) feeds the RNA structure along with the sequence into the model and improves the downstream tasks. However, it also ignores the significance of Str2Seq prediction capability. In a nutshell, existing FMs fail to achieve sequence-structure alignment without exception.

5 Conclusion

We introduced OmniGenome to tackle the challenge of sequence-structure alignment in genome modelling, which bridges the gap between sequence and structural information and improves the reliability of genome analysis. Experimental results on four comprehensive in-silico RNA and DNA benchmarks demonstrate that OmniGenome outperforms existing FMs across diversified downstream tasks, e.g., up to 98% F1 score for SSP and 74% accuracy of RNA design. The superior performance highlights the potential of sequence-structure alignment in the field of genomics.

Acknowledgements

This work was supported in part by the UKRI Future Leaders Fellowship under Grant MR/S017062/1 and MR/X011135/1; in part by NSFC under Grant 62376056 and 62076056; in part by the Royal Society under Grant IES/R2/212077; in part by the EPSRC under Grant 2404317; in part by the Kan Tong Po Fellowship (KTP\R1\231017); and in part by the Amazon Research Award and Alan Turing Fellowship.

References

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1–3.

Akiyama, M.; and Sakakibara, Y. 2022. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics*, 4(1): lqac012.

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3): 403–410.

Carpenter, E. J.; Leebens-Mack, J. H.; and et al., M. S. B. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780): 679–685.

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. *bioRxiv*, 2022–08.

Chen, K.; Zhou, Y.; Ding, M.; Wang, Y.; Ren, Z.; and Yang, Y. 2023. Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction. *bioRxiv*, 2023–01.

Chen, X.; Li, Y.; Umarov, R.; Gao, X.; and Song, L. 2020. RNA Secondary Structure Prediction By Learning Unrolled Algorithms. In *International Conference on Learning Representations*.

Chu, Y.; Yu, D.; Li, Y.; Huang, K.; Shen, Y.; Cong, L.; Zhang, J.; and Wang, M. 2024. A 5' UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence*, 1–12.

Corbett, K. S.; Edwards, D. K.; Leist, S. R.; Abiona, O. M.; Boyoglu-Barnum, S.; Gillespie, R. A.; Himansu, S.; Schäfer, A.; Ziwawo, C. T.; DiPiazza, A. T.; et al. 2020. SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. *Nature*, 586(7830): 567–571.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023–01.

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.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; Ronneberger, O.; Bodenstein, S.; Zielinski, M.; Bridgland, A.; Potapenko, A.; Cowie, A.; Tunyasuvunakool, K.; Jain, R.; Clancy, E.; Kohli, P.; Jumper, J.; and Hassabis, D. 2021. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*.

Fu, L.; Cao, Y.; Wu, J.; Peng, Q.; Nie, Q.; and Xie, X. 2022. UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic acids research*, 50(3): e14–e14.

Ganser, L. R.; Kelly, M. L.; Herschlag, D.; and Al-Hashimi, H. M. 2019. The roles of structural dynamics in the cellular functions of RNAs. *Nature reviews Molecular cell biology*, 20(8): 474–489.

Grešová, K.; Martinek, V.; Čechák, D.; Šimeček, P.; and Alexiou, P. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1): 25.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hallee, L.; Rafailidis, N.; and Gleghorn, J. P. 2023. cdsBERT-Extending Protein Language Models with Codon Awareness. *bioRxiv*.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kalvari, I.; Nawrocki, E. P.; Ontiveros-Palacios, N.; Argasinska, J.; Lamkiewicz, K.; Marz, M.; Griffiths-Jones, S.; Toffano-Nioche, C.; Gautheret, D.; Weinberg, Z.; et al. 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1): D192–D200.

Lee, J.; Kladwang, W.; Lee, M.; Cantu, D.; Azizyan, M.; Kim, H.; Limpaecher, A.; Gaikwad, S.; Yoon, S.; Treuille, A.; et al. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, 111(6): 2122–2127.

Li, W.; and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13): 1658–1659.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022: 500902.

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.

Mathews, D. H. 2019. How to benchmark RNA secondary structure prediction accuracy. *Methods*, 162: 60–67.

Mendoza-Revilla, J.; Trop, E.; Gonzalez, L.; Roller, M.; Dalla-Torre, H.; de Almeida, B. P.; Richard, G.; Caton, J.;

Lopez Carranza, N.; Skwark, M.; et al. 2023. A Foundational Large Language Model for Edible Plant Genomes. *bioRxiv*, 2023–10.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A. W.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C. M.; Massaroli, S.; Bengio, Y.; Ermon, S.; Baccus, S. A.; and Ré, C. 2023. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *CoRR*, abs/2306.15794.

Runge, F.; Franke, J. K.; Fertmann, D.; Backofen, R.; and Hutter, F. 2023. Partial RNA Design. *bioRxiv*.

Sato, K.; Akiyama, M.; and Sakakibara, Y. 2021. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature communications*, 12(1): 941.

Schiff, Y.; Kao, C.-H.; Gokaslan, A.; Dao, T.; Gu, A.; and Kuleshov, V. 2024. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*.

Su, J.; Ahmed, M. H. M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568: 127063.

Tan, Z.; Fu, Y.; Sharma, G.; and Mathews, D. H. 2017. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic acids research*, 45(20): 11570–11581.

Tinoco Jr, I.; and Bustamante, C. 1999. How RNA folds. *Journal of molecular biology*, 293(2): 271–281.

Wang, N.; Bian, J.; Li, Y.; Li, X.; Mumtaz, S.; Kong, L.; and Xiong, H. 2024. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 1–10.

Wang, X.; Gu, R.; Chen, Z.; Li, Y.; Ji, X.; Ke, G.; and Wen, H. 2023. UNI-RNA: universal pre-trained models revolutionize RNA research. *bioRxiv*, 2023–07.

Wayment-Steele, H. K.; Kladwang, W.; Watkins, A. M.; Kim, D. S.; Tunguz, B.; Reade, W.; Demkin, M.; Romano, J.; Wellington-Oguri, R.; Nicol, J. J.; et al. 2022. Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nature Machine Intelligence*, 4(12): 1174–1184.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Yaish, O.; and Orenstein, Y. 2022. Computational modeling of mRNA degradation dynamics using deep neural networks. *Bioinformatics*, 38(4): 1087–1101.

Yan, Z.; Hamilton, W.; and Blanchette, M. 2022. Integrated pretraining with evolutionary information to improve RNA secondary structure prediction. *bioRxiv*, 2022–01.

Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; and Yao, J. 2022. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mac. Intell.*, 4(10): 852–866.

Yang, Y.; Li, G.; Pang, K.; Cao, W.; Li, X.; and Zhang, Z. 2023. Deciphering 3'UTR mediated gene regulation using interpretable deep representation learning. *bioRxiv*, 2023–09.

Yin, W.; Zhang, Z.; He, L.; Jiang, R.; Zhang, S.; Liu, G.; Zhang, X.; Qin, T.; and Xie, Z. 2024. ERNIE-RNA: An RNA Language Model with Structure-enhanced Representations. *bioRxiv*, 2024–03.

Zhang, Y.; Ge, F.; Li, F.; Yang, X.; Song, J.; and Yu, D.-J. 2023. Prediction of multiple types of RNA modifications via biological language model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Zhang, Y.; Lang, M.; Jiang, J.; Gao, Z.; Xu, F.; Litfin, T.; Chen, K.; Singh, J.; Huang, X.; Song, G.; et al. 2024. Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1): e3–e3.

Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R. V.; and Liu, H. 2023. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *CoRR*, abs/2306.15006.