TorRNA - Improved Prediction of Backbone Torsion Angles of RNA by Leveraging Large Language Models

Sriram Devata and U. Deva Priyakumar*

Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500 032, India

E-mail: deva@iiit.ac.in;sriram.devata@research.iiit.ac.in

Abstract

RNA molecules play a significant role in many biological pathways and have diverse functional roles, which is a result of their structural flexibility to fold into diverse conformations. This structural flexibility makes it challenging to obtain the structures of RNAs experimentally. Deep learning can be used to predict the secondary structures of RNA and other properties such as the backbone torsion angles, to be used as restraints for the computational optimization of the tertiary structures of RNA. TorRNA is a transformer encoder-decoder model, that takes an input RNA sequence and predicts the (pseudo)torsion angles of each nucleotide with a pre-trained RNA-FM model as the encoder. TorRNA is able to achieve a performance boost of $2\% - 16\%$ over the previous (pseudo)torsion angle prediction method for RNAs. We also demonstrate that TorRNA can used as a tool for model quality assessment of candidate RNA structures.

Introduction

RNA molecules play a significant role in modulating many biological pathways, ranging from acting as catalytic ribozymes[1] to controlling gene expression via transcriptional regulation.[2] Recent advances in generation[3] and delivery[4] of RNA make it more feasible for RNA molecules to be used as therapeutic agents[5] to address the underlying pathology of diseases rather than treating the symptomology as done by small molecule-based therapeutics.[6] RNA that are involved in disease pathways can also serve as druggable targets for small molecules to bind to the RNA and modulate their function,[7] increasing the number of ways we can interfere with pathological mechanisms. This functional diversity of RNA molecules is closely tied to their structure, with their ability to fold into various conformations impacting how they interact with other molecules.[8–10] Determining the structures of RNA is important for understanding their mechanisms and to be able to exploit them as therapeutic agents and targets.

RNA molecules fold hierarchically with their secondary structure elements being folded first, which then interact and result in the tertiary structure.[11] RNA molecules fold into their secondary structures and specific sub-structures based on hydrogen bonding between the nucleotides and their stacking, to form helices and unique RNA loops like hairpin loops and pseudoknots. These secondary structure elements interact and form the tertiary structure, and result in the great structural plasticity exhibited by RNA molecules. Determining the tertiary structures of these RNA molecules through experimental means such as nuclear magnetic resonance and X-ray crystallography is challenging due to the resolution limits of these methods and the intrinsic structural plasticity of RNA molecules.[12,13]

To alleviate the struggles of determining the structure of RNA molecules experimentally, a number of computational approaches based on thermodynamic models and Watson-Crick-Franklin (WCF) interactions have been developed to determine the secondary structure of RNA molecules over the years.[14–19] Recently, new methods have made use of Machine learning (ML) algorithms to solve problems in computational chemistry such as predicting

and synthesizing new drug molecules[20–23], performing molecular dynamics simulations[24–26], protein stability and binding site prediction[27,28], and predicting physical molecular properties.[29–31] ML has been employed to predict the secondary structure of RNA as early as the 1990s.[32–34]

Recent advances in deep learning have resulted in improved prediction of macromolecular structures like proteins[35,36] and RNA.[37–39] The breakthroughs in protein structure prediction by deep learning are due to the improved prediction of contact maps and backbone structures, which are used as restraints for modelling the structures. However, there are only a few studies that predict such restraints for RNA molecules.[37,40] With existing methods optimizing the tertiary structure of RNA molecules when given the secondary structures, deep learning can be used to solve the downgraded problem of predicting the secondary structures and other structural properties[41] that can be used as restraints for the optimization. Presented in this manuscript, TorRNA focuses on accurate prediction of the backbone structure of RNA molecules by predicting the torsion and pseudotorsion angles that can characterize the backbone of an RNA molecule.

In proteins, the backbone configuration can be described by only two backbone conformational parameters ϕ and ψ. For nucleic acid structures like RNA and DNA however, the phosphodiester backbone is best characterized by 6 torsion angles ($\alpha, \beta, \gamma, \delta, \epsilon,$ and ζ), and a torsion angle χ that quantifies the orientation of the base with respect to the sugar. For a nucleotide indexed i and the next nucleotide along the $5' - 3'$ direction indexed as $i + 1$, these 7 torsion angles as shown in Figure 1 can be described as the dihedral angle between the atoms $O3'_{i-1} - P_i - O5'_i - C5'_i(\alpha)$, $P_i - O5'_i - C5'_i - C4'_i(\beta)$, $O5'_i - C5'_i - C4'_i - C3'_i(\gamma)$, $C5'_i - C4'_i - C3'_i - O3'_i(\delta)$, $C4'_i - C3'_i - O3'_i - P_{i+1}(\epsilon)$, $C3'_i - O3'_i - P_{i+1} - O5'_{i+1}(\zeta)$, and $O4'_i - C1'_i - (N9_i/N1_i) - (C2_i/C4_i)(\chi)$. To simplify the representation of the RNA backbone configuration, two pseudotorsion angles eta (η) and theta (θ) can be used to describe the RNA backbone configuration[8,42] similar to how ϕ and ψ are used to describe backbone configuration of proteins. These pseudotorsion angles as shown in Figure 1 can be described as

the dihedral angle between the atoms $C4'_{i-1}-P_i-C4'_i-P_{i+1}(\eta)$ and $P_i-C4'_i-P_{i+1}-C4'_{i+1}(\theta)$ where $i-1$, i, and $i+1$ are the indices of three nucleotides in the $5'-3'$ direction. These 9 torsion and pseudotorsion angles are depicted in Figure 1 and are henceforth referred to as (pseudo)torsion in the rest of the manuscript.

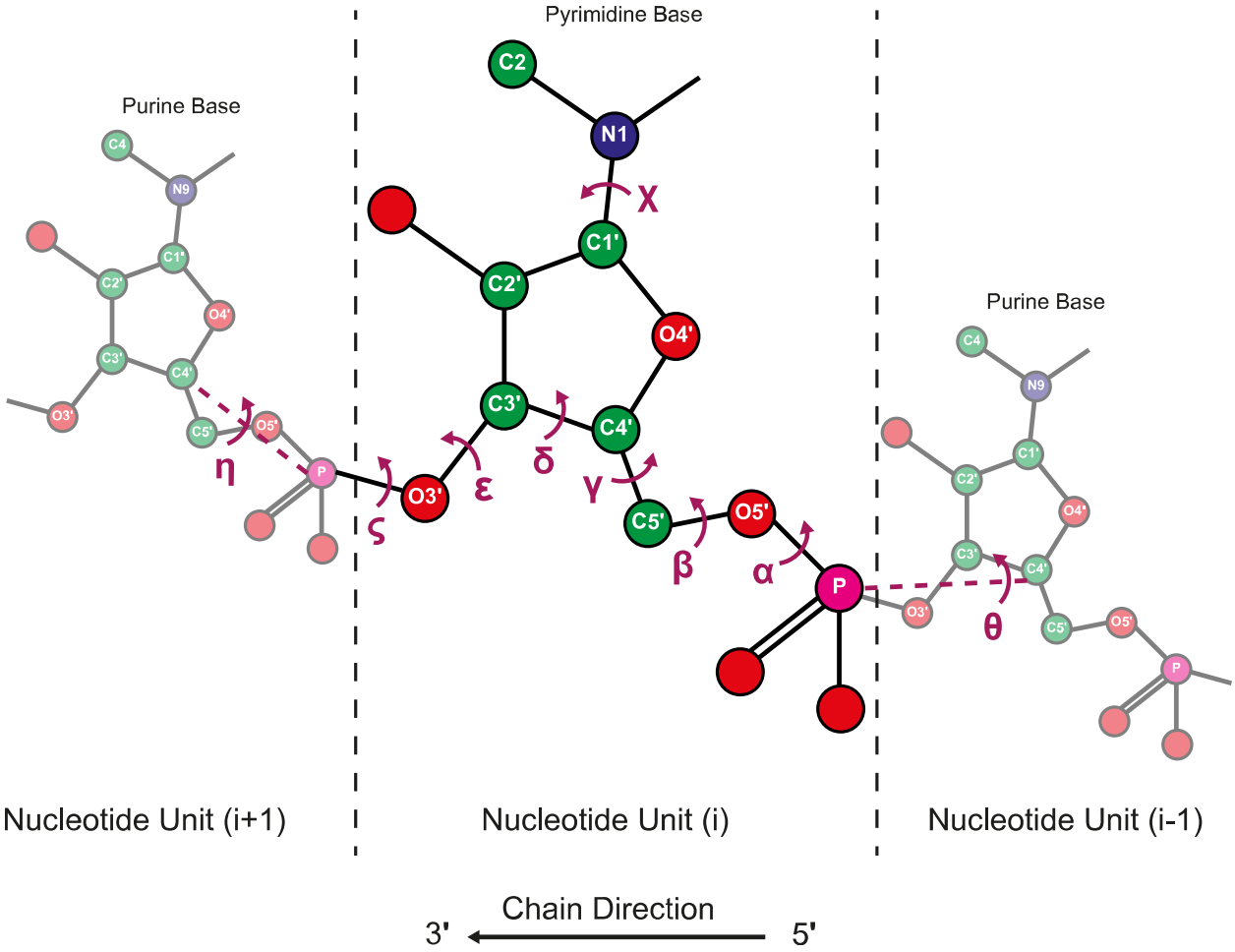


Figure 1: RNA backbone torsion $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \chi)$ and pseudotorsion (η, θ) angles.

SPOT-RNA-1D[40] employed a residual dilated convolutional neural network architecture[43,44] to predict seven torsion and two pseudotorsion angles, and was able to beat a random baseline predictor by achieving a mean absolute error (MAE) between 14° and 44° for the nine (pseudo)torsion angles. The design choice of using a dilated convolutional neural network architecture is justified by the architecture's ability to learn long-range interactions between nucleotides. However, SPOT-RNA-1D and other methods that predict secondary

structures of RNA employ variations of CNNs since the properties they predict are represented by two-dimensional matrices - such as contact maps.

When compared to proteins, PDB[45] has fewer 3D RNA structures. This lack of RNA sequence-structure datapoints is one of the greatest challenge in developing ML-based sequence to structure methods for RNA. The RNA foundation model (RNA-FM)[41] is a foundation model trained in a self-supervised manner to learn any patterns in the RNA sequences and generates sequence encodings that potentially capture the underlying evolutionary, structural, and functional information of the corresponding RNA molecules from their sequences. RNA-FM implicitly learns the co-evolutionary information of RNA sequences from 23 million unlabeled non-coding RNA sequences and performed well in downstream tasks like RNA secondary structure prediction and 3D contact map prediction. RNA-FM has been used by E2Efold-3D[46] to develop the first end-to-end deep learning approach to predict 3D RNA structures directly from the sequence, highlighting the importance of the information contained in the RNA-FM encodings.

In this work, we present TorRNA - a method that focuses on predicting the (pseudo)torsion angles of each residue by using a transformer[47] architecture to predict the (pseudo)torsion angles. TorRNA uses the encodings of all nucleotides of an input RNA sequence as generated by a pre-trained RNA-FM model and predicts the (pseudo)torsion angles using a transformer decoder architecture. The choice of using a transformer is consistent with the choice of using a dilated convolutional neural network since a transformer also contains residual connections[48] to help learn the long-range interactions between nucleotides.

Methods

Dataset

SPOT-RNA-1D's[40] training dataset contains 286 RNA chains, with the validation and test dataset containing 30 and 147 RNA chains respectively. However, this dataset was con-

structed by downloading all RNA structures from PDB[45] with a suitable X-ray resolution on October 3, 2020. To train and test TorRNA, we sought to create a new dataset that contains the RNA structures uploaded to PDB[45] in the recent years.

The dataset of RNAs used for training and testing TorRNA was curated with data from RCSB Protein Data Bank (PDB)[45] and BGSU RNA Representative Sets.[49] More specifically, we assembled the PDB identifiers of RNA structures that were available with a resolution of $< 4\mathring{A}$ from PDB on July 4, 2023 and from Release 3.288 of BGSU RNA Representative Sets. The structures of these RNAs were downloaded from PDB[45] using their PDB identifiers. The downloaded PDB structures are processed using the Biopython[50] package to obtain the structures of individual RNA chains.

We follow the same methodology as SPOT-RNA[37] to make the train, validation, and test splits of the dataset. To remove the redundancies in the dataset, the sequences of all the RNA chains with < 500 nucleotides were clustered using CD-HIT-EST[51] with a sequence identity threshold of 80%. The RNA sequences that do not belong to any clusters are assigned to a noncluster set (NCS), and the clustered RNA sequences are assigned to a cluster set (CS). To ensure an even stronger nonredundancy between NCS and CS, we run the BLAST-N[52] tool on the RNA sequences with an e-value cutoff of 10. Sequences in CS that have hits with sequences in NCS are removed to ensure that sequence homologies between CS and NCS are minimal. The resulting CS is used as the training data, and NCS is randomly divided into validation and test dataset with a 20-80 split.

While dividing NCS into the validation and test datasets, we maintained the RNA sequences from the RNA-Puzzles benchmarking test set[53–57] exclusively in the test dataset for TorRNA to run further experiments on these RNAs as described in the . The final training, validation, and test datasets have 767, 42, and 172 RNAs respectively. When comparing the performance of TorRNA with SPOT-RNA-1D[40] in the Results section, we use the same dataset splits used by SPOT-RNA-1D[40] in one of the results. For the list of curated PDB IDs, we use the DSSR[58,59] software tool to calculate the native torsion angles and to identify

the structural regions from the 3D structures. The final dataset is available on our code repository.

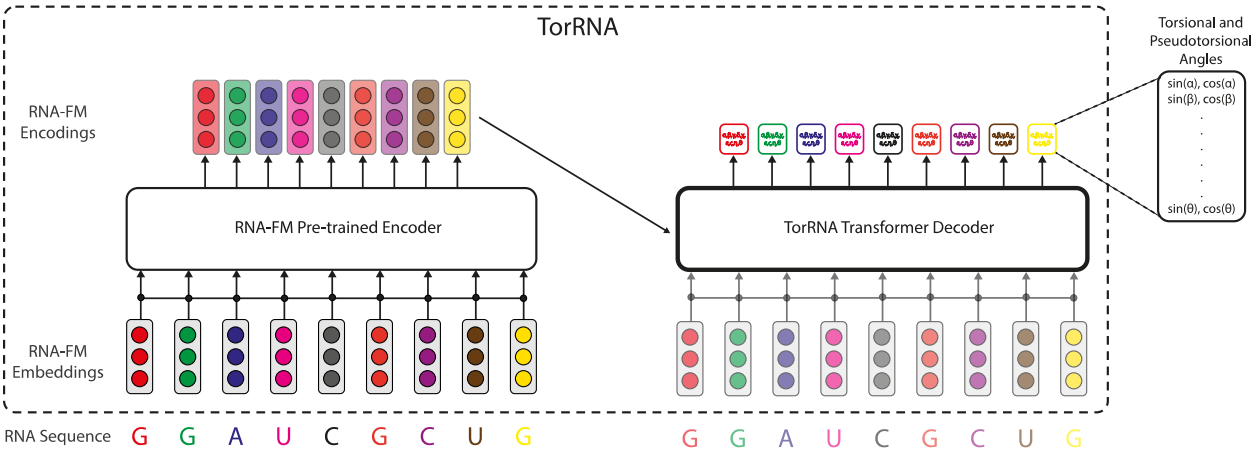
Architecture of TorRNA

TorRNA's overall architecture is a transformer encoder-decoder as shown in Figure 2a, that takes an input RNA sequence and predicts the (pseudo)torsion angles of each nucleotide. TorRNA utilizes a pre-trained RNA-FM[41] model's embedding layer and subsequent transformer encoder blocks[47,60] to obtain encodings for each residue of an RNA sequence. RNA-FM's[41] model architecture as shown in Figure 2b is a stack of 12 transformer encoder blocks, similar to the BERT[60] language model architecture. Each encoder block has a hidden size of 640 and 20 self-attention heads, with layer normalization and residual connections being applied before and after every block. For an RNA sequence as the input, RNA-FM first tokenizes the sequence into the individual nucleotide tokens ('A', 'U', 'G', and 'C' among others). An initial embedding layer maps each of these sequential nucleotide tokens to 640-dimensional vectors. These initial embeddings are passed through the stack of 12 encoder blocks to give final encodings of the same size for each nucleotide. These final encodings of each nucleotide contain information aggregated from the entire RNA sequence.

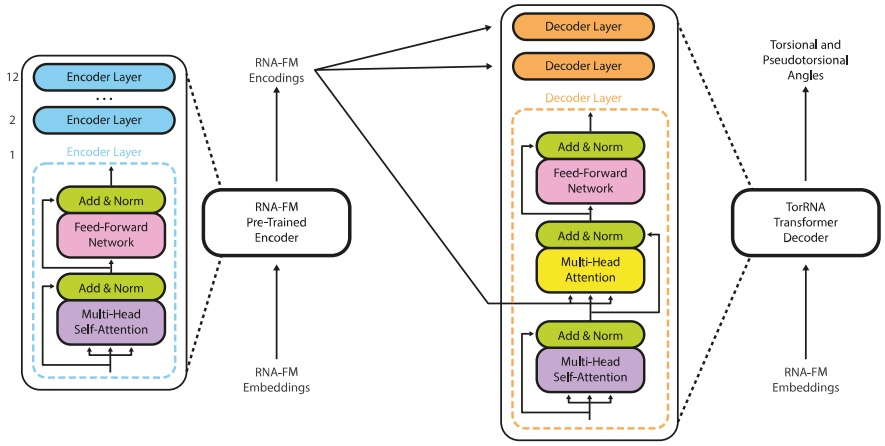
The final encodings of each nucleotide computed by the pre-trained RNA-FM model are then passed to a stack of 3 transformer decoder blocks[47] along with the embeddings of the nucleotides computed by the pre-trained embedding layer of RNA-FM. These decoder blocks as shown in Figure 2b use the embeddings of each nucleotide and perform cross-attention over the RNA-FM encodings to finally predict the (pseudo)torsion angles for each nucleotide. Since these angles are in the range $[-180°, 180°]$, TorRNA predicts the *sine* and *cosine* values of the (pseudo)torsion angles instead of predicting the angles directly to handle the periodicity of the angles as done in previous works that predict torsion angles for RNA and proteins.[40,61] The predicted *sine* and *cosine* values can be used to calculate the angle using the inverse tangent function.

$$angle = \tan^{-1}\left(\frac{\sin(angle)}{\cos(angle)}\right)$$

The transformer decoder layers of TorRNA are trained to minimize the Mean Squared Error (MSE) of the *sine* and *cosine* of the (pseudo)torsion angles using the Adam optimizer[62] with the hyperparameters as chosen in Table 1. The training and testing of TorRNA was done on a system with a Intel Xeon E5-2640 v4 processor and a GeForce RTX 2080 Ti GPU.



(a) Overall architecture of TorRNA is a transformer encoder-decoder that takes an input RNA sequence and predicts the (pseudo)torsion angles of each nucleotide.



(b) Details of the RNA-FM encoder blocks and the TorRNA decoder blocks that shows how the RNA-FM embeddings are used by the decoder blocks to predict the (pseudo)torsion angles.

Figure 2: Overall architecture of TorRNA.

To choose the best hyperparameters for TorRNA's architecture and training procedure,

we conduct grid search of the hyperparameters presented in Table 1. We chose the best values for the hyperparameters when the error of predicting the (pseudo)torsion angles was the lowest for the RNAs in the validation dataset.

The code and datasets for TorRNA are available at https://github.com/devalab/torrna.

Table 1: Search space and the best value for the various hyperparameters for TorRNA

Hyperparameter	Search Space	Best Value
Learning Rate	[0.0001, 0.0002]	0.0002
Hidden Dimension	[256, 512]	256
Number of Attention Heads	[4, 8]	4
Number of Transformer Decoder Layers	[2, 3, 4, 5, 6, 7]	3
Dropout	[0.1, 0.2]	0.2
Tolerance	[3, 5]	5

Results

To evaluate the performance of TorRNA, we use the Mean Absolute Error (MAE), which is the average absolute error between the predicted and ground truth (pseudo)torsion angles. To handle the periodicity of the angles in the MAE calculation, we consider $min(d, 360^\circ - d)$, where d is the absolute difference between two angles. We compare the results of TorRNA with SPOT-RNA-1D[40] and a random predictor. The random predictor works by constructing a histogram of the native angles from the RNAs in the training dataset with a bin-width of 2°, and returns the mean of 100 random predictions using the normalized frequency of each bin as the discrete probability distribution for the center of each bin.

TorRNA outperforms SPOT-RNA-1D and the random baseline predictor

Table 2 and Figure 3 compare the performance of TorRNA, SPOT-RNA-1D, and the random baseline predictor in predicting the (pseudo)torsion angles. To provide a direct comparison

with SPOT-RNA-1D , we show the performance of TorRNA when trained and tested on dataset splits curated in this work in Table 2, and on the dataset splits used by SPOT-RNA-1D in Table 3.

TorRNA shows improved performance in predicting all torsion angles $(\alpha, \beta, \gamma, \delta, \chi, \epsilon, \zeta)$ and both pseudotorsion angles (η, θ) when compared to both SPOT-RNA-1D and the random baseline predictor. The common trend exhibited by the ML-based prediction methods is that the prediction of the angle delta (δ) has the least average error and the angle alpha (α) has the highest average error. TorRNA and SPOT-RNA-1D have MAEs of 14.26° and 17.1° when predicting the angle delta (δ), and MAEs of 42.1° and 46.1° when predicting the angle alpha (α) . TorRNA predicts the angle delta (δ) with the least error, followed by the angles epsilon (ϵ), chi (χ), beta (β), zeta (ζ), gamma (γ), and alpha (α). When compared to SPOT-RNA-1D, TorRNA achieves an improvement ranging from 2.7% for angle beta (β) to 16.5% for angle delta (δ).

Since the available source code for SPOT-RNA-1D does not allow the model to be re-trained with new dataset splits, to obtain a direct comparision, we retrain and test TorRNA on the same RNA molecules on which SPOT-RNA-1D was trained and tested. The performance of the retrained TorRNA and SPOT-RNA-1D are presented in Table 3, which show that TorRNA has better predictions of 8/9 of the (pseudo)torsion angles when compared to SPOT-RNA-1D. In the Supplementary Information, we compare TorRNA against other predictors submitted to RNA-Puzzles.[53–57] TorRNA consistently performs the best in predicting the torsion angles for most puzzles, and gives comparable predictions to the top RNA puzzle predictor in the remaining puzzles.

Correlation between TorRNA's prediction errors and (pseudo)torsion angle distributions

The boxplot of the prediction errors of the (pseudo)torsion angles shown in Figure 3 shows the distribution of the errors whose averages are presented as the MAEs in Table 2. TorRNA's

Table 2: MAE of TorRNA compared with SPOT-RNA-1D and the random baseline method for all (pseudo)torsion angles on TorRNA dataset splits

(pseudo)torsion angle	Prediction Method		
	TorRNA	SPOT-RNA-1D	Random Baseline
alpha (α)	**42.052**	46.079	73.044
beta (β)	**20.626**	21.209	123.877
gamma (γ)	**36.443**	37.958	59.064
delta (δ)	**14.257**	17.081	19.538
chi (χ)	**20.11**	21.999	46.129
epsilon (ϵ)	**19.306**	20.311	36.209
zeta (ζ)	**29.182**	30.545	50.646
eta (η)	**25.124**	29.114	79.595
theta (θ)	**28.82**	30.725	67.517

Table 3: MAE of TorRNA compared with SPOT-RNA-1D and the random baseline method for all (pseudo)torsion angles on SPOT-RNA-1D dataset splits

(pseudo)torsion angle	Prediction Method		
	TorRNA	SPOT-RNA-1D	Random Baseline
alpha (α)	**38.87**	40.371	72.968
beta (β)	**19.677**	19.82	123.241
gamma (γ)	**31.289**	32.149	55.059
delta (δ)	**12.668**	14.71	17.396
chi (χ)	**16.407**	18.159	48.259
epsilon (ϵ)	19.956	**19.798**	33.564
zeta (ζ)	**27.033**	28.034	49.241
eta (η)	**22.677**	26.537	76.677
theta (θ)	**25.788**	27.887	65.929

prediction errors follow the same trend as SPOT-RNA-1D where the difficulty of predicting the (pesudo)torsion angles depends on the distribution of the (pseudo)torsion angle. As seen in Figure 4, the ground truth values of the angle delta (δ) have a narrow distribution, which explains the low prediction error and the narrow range of the errors in predicting this angle in Figure 3. The wide distribution of the ground truth values of the angle alpha (α) explain the prediction errors having a wide range in Figure 3 and a high MAE as reported in Table 2.

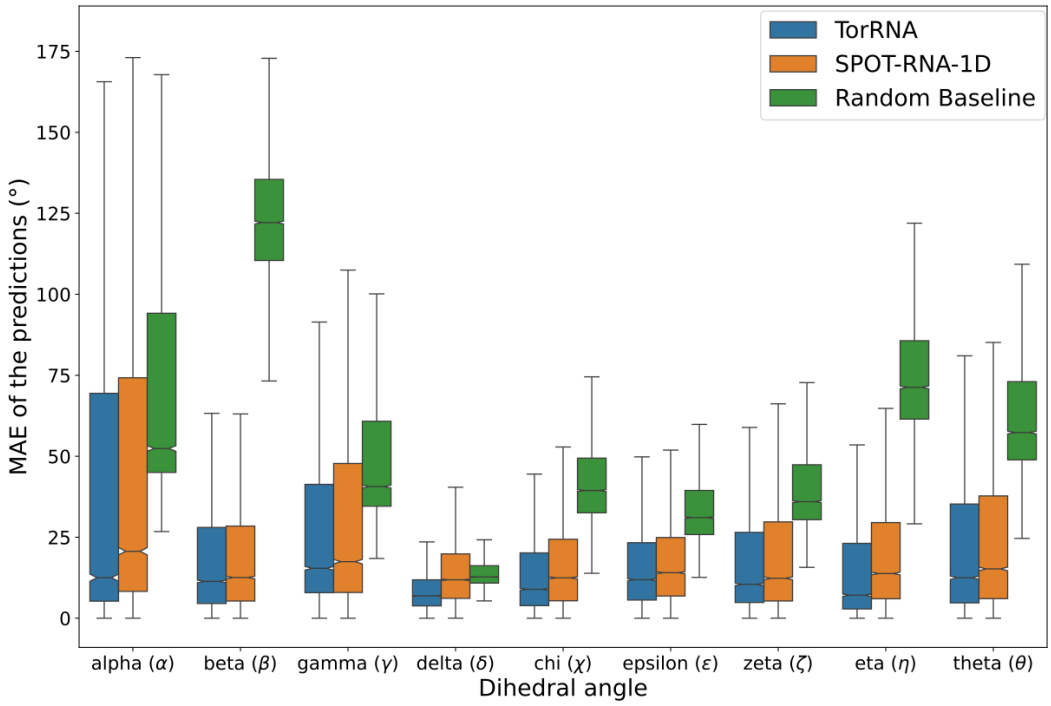


Figure 3: Boxplot of the prediction errors of the (pseudo)torsion angles to compare the distribution of the errors of TorRNA, SPOT-RNA-1D, and the random baseline predictor.

TorRNA's predictive ability for various structural regions of RNA molecules

We investigate TorRNA's (pseudo)torsion angle predictions of nucleotides with various secondary and tertiary interactions with other nucleotides within an RNA molecule. The DSSR[58,59] software tool marks each nucleotide with the type of interaction in which it is involved. Table 4 shows the MAEs obtained by averaging TorRNA's prediction errors for the nucleotides in various structural regions. Figure 5 shows the various structural regions that we consider in Table 4.

The (pseudo)torsion angles of nucleotides that are unpaired ($\sim 28\%$) or are part of hairpin loops ($\sim 12\%$) are the hardest to predict. The difficulty in predicting the (pseudo)torsion

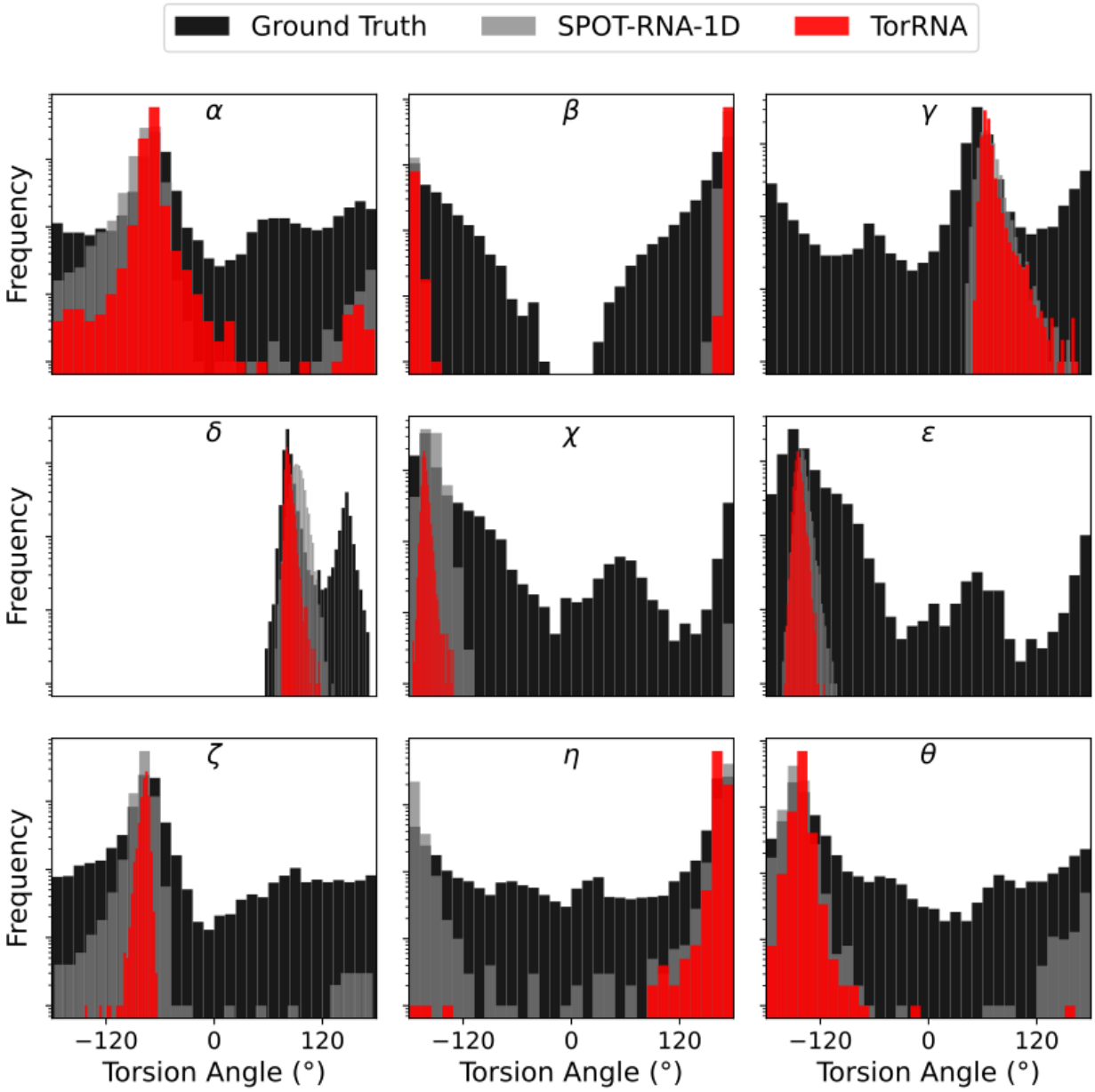
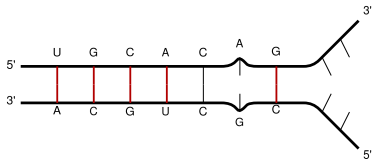
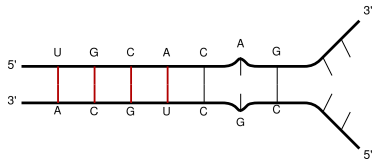


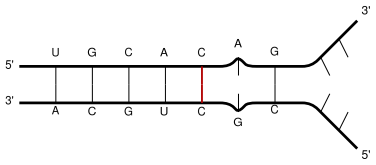
Figure 4: Histograms of ground truth (pseudo)torsion angles and those predicted by TorRNA and SPOT-RNA-1D. The Y-axis uses a logarithmic scale to show the frequency of each frequency bin in the histogram.



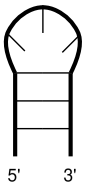
(a) Canonical Base Paired Residues.



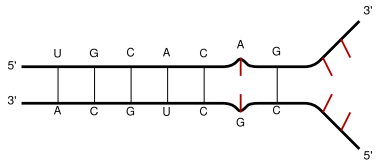
(b) Canonical Nested Base Paired Residues.



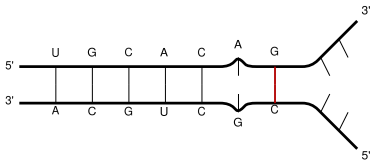
(c) Non-Canonical Base Paired Residues.



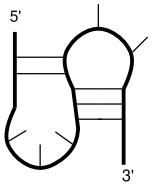
(d) Hairpin Loop Residues.



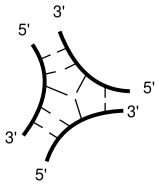
(e) Unpaired Residues.



(f) Lone Base Paired Residues.



(g) Pseudoknot Residues.



(h) Multiplet Residues.

Figure 5: The various structural regions of RNA molecules that we consider. The specific residues are highlighted in red when the region is ambiguous from the figure.

angles of these regions could be due to the unpaired nucleotides being very flexible, and TorRNA having no geometric information to infer the nucleotides in hairpin loops which have a distribution of angles away from the remaining nucleotides. Nucleotides that are a part of canonical nested pairs make up $\sim$ 47% of all nucleotides and are the easiest to predict. As seen in Table 4, TorRNA predicts all (pseudo)torsion angles better than SPOT-RNA-1D across most structural region, and gives comparable results to SPOT-RNA-1D in the remaining cases.

TorRNA's robustness to the length of RNA sequences

The lengths of the longest RNA sequence in the training, validation, and test sets varying greatly could potentially affect the performance of TorRNA on long RNA molecules. To analyse this, Figure 6 shows the MAEs of all (pseudo)torsion angles for RNA molecules of varying sequence lengths. All the (pseudo)torsion angles largely have the same MAEs

Table 4: MAE of angles predicted by TorRNA in various regions of an RNA molecule with the MAE of the predictions by SPOT-RNA-1D in the parenthesis. SPOT-RNA-1D MAEs are in bold when they are lower than the corresponding MAE of TorRNA.

Region (% of all nucleotides present in this region)	alpha (α)	beta (β)	gamma (γ)	delta (δ)	chi (χ)	epsilon (ϵ)	zeta (ζ)	eta (η)	theta (θ)
All Canonical Pairs (50.51%)	29.51 (32.72)	15.06 (15.66)	27.54 (28.26)	7.82 (11.66)	9.37 (12.15)	13.80 (15.44)	15.18 (16.28)	11.10 (15.48)	14.43 (16.05)
Canonical Nested Pairs (46.49%)	33.32 (36.56)	16.89 (17.33)	30.45 (31.25)	9.01 (12.41)	12.43 (14.54)	15.13 (16.73)	18.51 (19.47)	13.95 (17.94)	17.30 (18.79)
Non-Canonical Pairs (25.61%)	47.24 (50.10)	24.14 (24.64)	41.32 (41.66)	15.58 (17.91)	24.48 (25.48)	22.42 (22.66)	37.94 (38.76)	25.94 (28.14)	33.89 (34.86)
Hairpin Loops (11.72%)	62.57 (62.66)	26.55 (27.14)	42.77 (42.92)	22.75 (23.57)	28.52 (**28.47**)	26.57 (**26.55**)	46.18 (46.69)	48.98 (51.52)	45.43 (46.73)
Unpaired (27.77%)	61.56 (63.52)	29.32 (29.97)	48.21 (48.91)	22.45 (23.69)	32.14 (32.58)	26.75 (26.79)	46.76 (47.87)	48.18 (50.83)	48.83 (50.05)
Lone Pairs (5.99%)	44.43 (46.32)	22.18 (23.19)	37.18 (37.55)	16.28 (18.67)	22.48 (23.55)	22.96 (**22.73**)	42.93 (43.84)	25.10 (27.46)	36.16 (36.57)
Pseudoknots (2.71%)	38.30 (40.46)	18.51 (18.61)	30.23 (31.47)	11.88 (13.76)	13.38 (14.58)	18.73 (18.80)	27.96 (28.78)	23.60 (25.31)	27.61 (28.30)
Multiplets (9.24%)	50.89 (53.55)	24.89 (25.33)	42.53 (42.99)	17.36 (18.81)	25.07 (25.29)	22.88 (**22.78**)	42.60 (42.76)	27.60 (28.56)	37.04 (37.53)

for RNAs of all lengths. It can also be noted that there is no clear loss in performance in predicting the (pseudo)torsion angles of RNAs of greater lengths, with some angles even having their lowest prediction errors for the longest RNAs in the test dataset.

Using TorRNA as a model evaluator

While developing Ribonucleic Acids Statistical Potential (RASP)[63] - an all-atom knowledge-based potential for the assessment of 3D RNA structures - the authors use 500 decoy models for each of the 85 native RNA structures in a dataset that they name *randstr* decoy set to test the knowledge-based potential they developed. These decoys were built with the MODELLER computer program[64] using an increasingly smaller subset of Gaussian potentials as restraints on the dihedral angles and atomic distances.

Out of these 85 RNAs, 2 RNAs are present in the testing dataset of TorRNA and are non-redundant with the training dataset. We use these 2 RNAs to explore the connection between the prediction errors of TorRNA and the structural accuracy of the models measured by the root-mean-square deviation (RMSD) and global distance test (GDT) score[65] to their native structures. Figure 7 plots the MAEs between the (pseudo)torsion angles predicted by TorRNA and the angles of the decoy models against the structural accuracy of the models for the PDB IDs 1MZP (Chain B) and 387D (Chain A). The MAE of the predictions increase

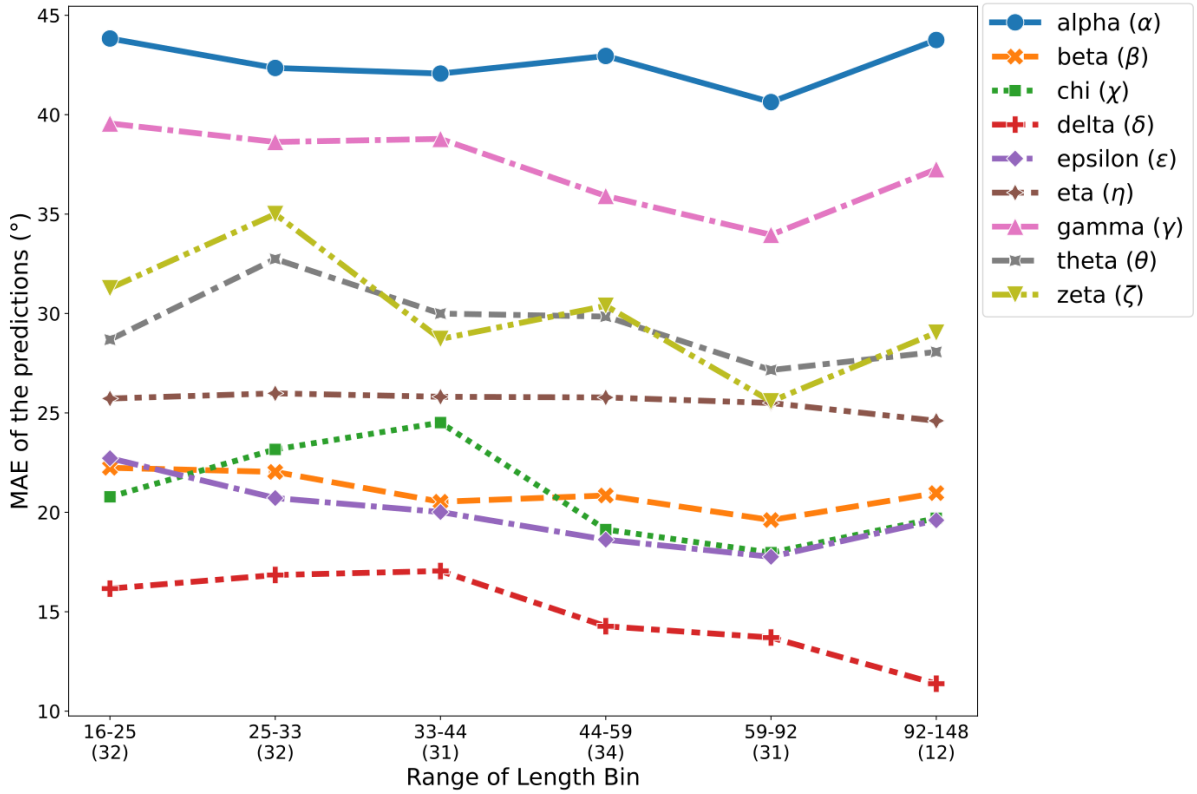


Figure 6: MAEs of the (pseudo)torsion angles for various RNA sequence lengths. The X-axis labels describe the length bins along with the number of RNAs that are in each length bin.

as the structural accuracy of the models decrease, i.e. as the RMSD increases and the GDT decreases. This shows that the MAE of the predictions can serve as an effective proxy when the RMSD and GDT scores are not available, which is the case when generating the structure of a novel RNA.

In Figure 8a, we plot the distribution of the MAEs between the (pseudo)torsion angles predicted by TorRNA and the angles of the decoy models for the decoy models that have the minimum and maximum MAE for each RNA in the *randstr* decoy set. Figure 8b plots the distributions of the RMSDs of decoy models that have the minimum and maximum MAEs against the angles predicted by TorRNA. Both these figures show that the MAEs and RMSDs of the decoy models with minimum and maximum MAEs show disjoint distributions, implying that the MAE calculated against the (pseudo)torsion angles by TorRNA is a good

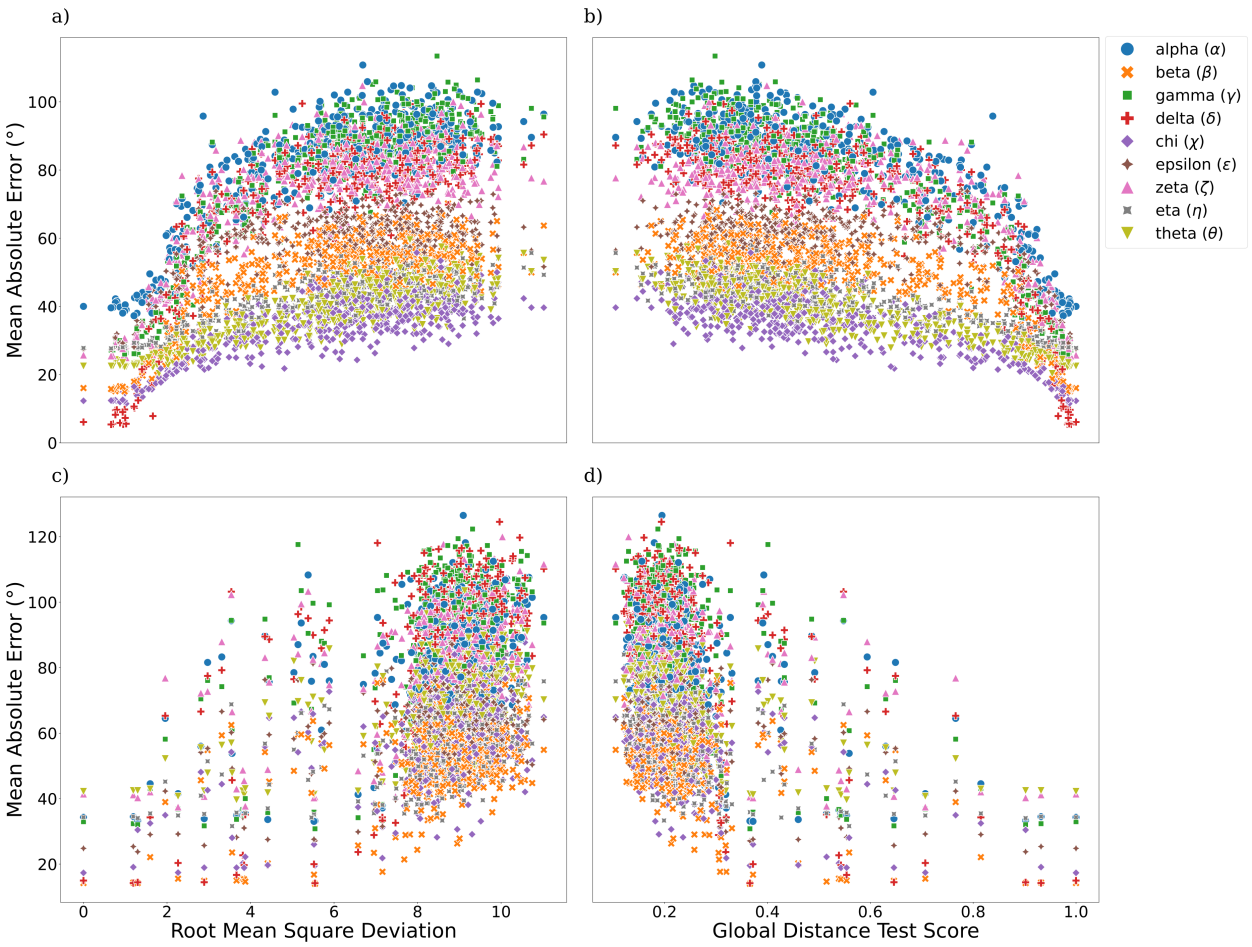


Figure 7: MAE vs RMSD and MAE vs GDT scatterplots for PDB ID 1MZP (Chain B) (a, b) and 387D (Chain A) (c, d)

metric to assess the quality of the decoy models. Figure 9 shows the best (green) and worst (red) decoy models against the native structure (black) of 3 RNAs.

These results show that the difference of the (pseudo)torsion angles predicted by TorRNA from the angles of a candidate model structure could be used as a model quality assessment of the candidate 3D structure of the RNA molecule. TorRNA's MAEs can be used to distinguish and correctly rank candidate models of RNA structures, even when the candidate models have minimal structural deviation. TorRNA can work as a powerful RNA model quality assessment tool to rank candidate models generated by ML-based methods or through other methods.

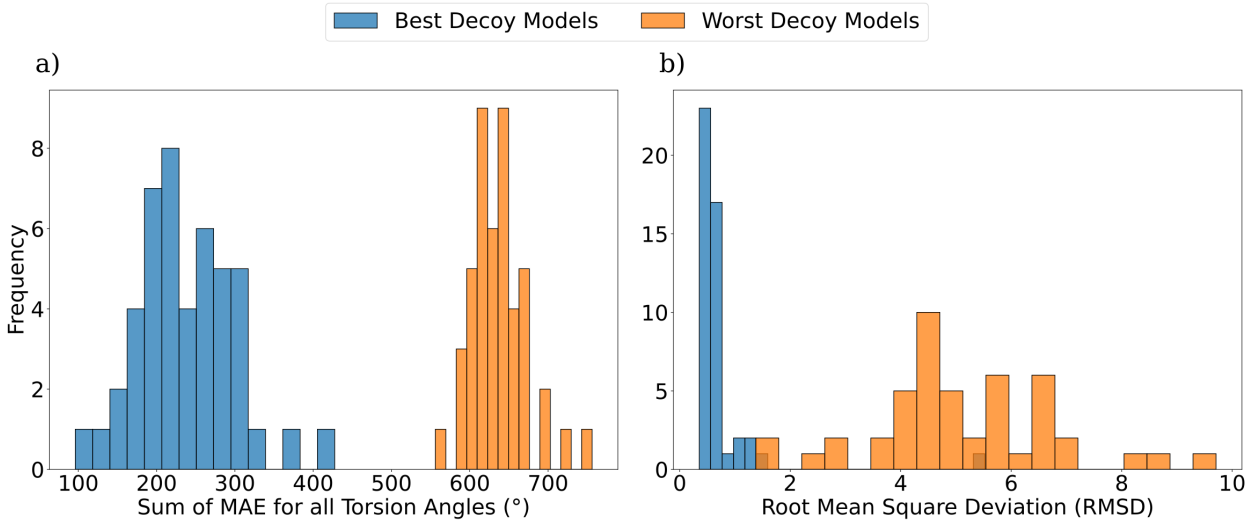


Figure 8: The MAE of a model's angles against TorRNA's predictions separates the best and worst decoy models both in terms of the MAE, and also in terms of the RMSD of the decoy structures with the native structure.

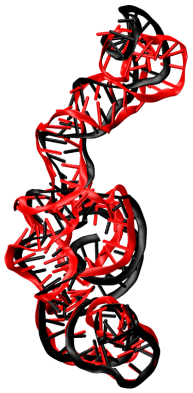
(a) Best model for PDB ID 1Z43 (Chain A) according to TorRNA. RMSD - 0.405 $\mathring{A}$; MAE - 284°.

(b) Worst model for PDB ID 1Z43 (Chain A) according to TorRNA. RMSD - 4.582 $\mathring{A}$; MAE - 608°.

(c) Best model for PDB ID 3CPW (Chain 9) according to TorRNA. RMSD - 0.687 $\mathring{A}$; MAE - 207°.

(d) Worst model for PDB ID 3CPW (Chain 9) according to TorRNA. RMSD - 4.626 $\mathring{A}$; MAE - 583°.

(e) Best model for PDB ID 3F1H (Chain B) according to TorRNA. RMSD - 0.605 $\mathring{A}$; MAE - 169°.



(f) Worst model for PDB ID 3F1H (Chain B) according to TorRNA. RMSD - 4.667 $\mathring{A}$; MAE - 622°.

Figure 9: The native structure (black) of various RNAs and the decoy model with the lowest (green) and highest MAE (red) against the angles predicted by TorRNA to show TorRNA's potential to be used as a model quality assessment tool. The caption of each subfigure also contains the RMSD of the decoy model to the native structure, and the sum of MAE between TorRNA's predictions and the decoy model's (pseudo)torsion angles.

Discussion

TorRNA is a transformer encoder-decoder model, that takes an input RNA sequence and predicts the (pseudo)torsion angles of each nucleotide with a pre-trained RNA-FM model as the transformer encoder. Since the secondary structure being predicted are the (pseudo)torsion angles, TorRNA is able to employ a transformer decoder that takes the encodings from a pre-trained transformer encoder. This sets TorRNA apart from other works that use a CNN-based architecture to predict the secondary structure of proteins and nucleic acids from encodings derived by foundation models. TorRNA also curates new dataset splits of the RNAs that have high-resolution 3D structures available, to take into the account new data that might have been gathered since the previous (pseudo)torsion angle prediction method was released.

TorRNA is able to achieve a performance boost of $2\% - 16\%$ over the previous (pseudo)torsion angle prediction method SPOT-RNA-1D and consequently shows an improved performance over a random baseline predictor as well. TorRNA is also robust in terms of predicting the (pseudo)torsion angles for RNAs of various sizes, and for nucleotides in various structural regions of the RNA molecules. With this improved prediction of the (pseudo)torsion angles, these predictions can be used as restraints on the dihedrals for the optimization of unrefined RNA structures. We also demonstrate the potential of TorRNA to be used as a tool for model quality assessment of candidate RNA structures for a given RNA sequence.

We believe that TorRNA is a valuable contribution that would help spur further research in improving sequence to structure methods for RNA molecules and take a step towards unleashing the therapeutic value of RNA molecules to develop better drugs.

Data Availability Statement

The data and code used in this work can be accessed through the GitHub repository.

Supplementary Data statement

Supplementary Data are available for this manuscript.

Funding

This work was supported by IHub-Data, DST-SERB (CRG/2021/008036), and IIIT Hyderabad's Kohli Center on Intelligent Systems. The funders however did not have any role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest Disclosure

We have no conflicts of interest to declare.

Author contributions

U.D.P. conceptualized the problem statement. S.D. developed and designed the architecture of the model, implemented the code for training and testing the model, curated the dataset, prepared the figures, and wrote the initial draft of the manuscript. S.D. and U.D.P. edited and reviewed the final manuscript. U.D.P. supervised the project.

References

(1) Walter, N. G.; Engelke, D. R. Ribozymes: catalytic RNAs that cut things, make things, and do odd and useful jobs. *Biologist (London)* **2002**, *49*, 199–203.

(2) Long, Y.; Wang, X.; Youmans, D. T.; Cech, T. R. How do lncRNAs regulate transcription? *Science advances* **2017**, *3*, eaao2110.

(3) Shu, Y.; Pi, F.; Sharma, A.; Rajabi, M.; Haque, F.; Shu, D.; Leggas, M.; Evers, B. M.; Guo, P. Stable RNA nanoparticles as potential new generation drugs for cancer therapy. *Advanced Drug Delivery Reviews* **2014**, *66*, 74–89, Cancer nanotechnology.

(4) Muskan, M.; Abeysinghe, P.; Cecchin, R.; Branscome, H.; Morris, K. V.; Kashanchi, F. Therapeutic potential of RNA-enriched Extracellular Vesicles 1: The next generation in RNA delivery via biogenic nanoparticles. *Molecular Therapy* **2024**,

(5) Mollica, L.; Cupaioli, F. A.; Rossetti, G.; Chiappori, F. An overview of structural approaches to study therapeutic RNAs. *Frontiers in Molecular Biosciences* **2022**, *9*.

(6) Zogg, H.; Singh, R.; Ro, S. Current Advances in RNA Therapeutics for Human Diseases. *International Journal of Molecular Sciences* **2022**, *23*.

(7) Childs-Disney, J. L.; Yang, X.; Gibaut, Q. M. R.; Tong, Y.; Batey, R. T.; Disney, M. D. Targeting RNA structures with small molecules. *Nature Reviews Drug Discovery* **2022**, *21*, 736–762.

(8) Grille, L.; Gallego, D.; Darré, L.; da Rosa, G.; Battistini, F.; Orozco, M.; Dans, P. D. The Pseudo-Torsional Space of RNA. *bioRxiv* **2022**,

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

(10) Wu, L.; Belasco, J. G. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Molecular cell* **2008**, *29*, 1–7.

(11) Tinoco, I.; Bustamante, C. How RNA folds. *Journal of Molecular Biology* **1999**, *293*, 271–281.

(12) Zhang, J.; Fei, Y.; Sun, L.; Zhang, Q. C. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nature Methods* **2022**, *19*, 1193–1207.

(13) Ma, H.; Jia, X.; Zhang, K.; Su, Z. Cryo-EM advances in RNA structure determination. *Signal Transduction and Targeted Therapy* **2022**, *7*, 58.

(14) Nussinov, R.; Pieczenik, G.; Griggs, J. R.; Kleitman, D. J. Algorithms for loop matchings. *SIAM Journal on Applied mathematics* **1978**, *35*, 68–82.

(15) Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research* **1981**, *9*, 133–148.

(16) Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **2003**, *31*, 3406–3415.

(17) Markham, N. R.; Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Bioinformatics: structure, function and applications* **2008**, 3–31.

(18) Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic acids research* **2003**, *31*, 3429–3431.

(19) Reuter, J. S.; Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics* **2010**, *11*, 1–9.

(20) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **2019**, *18*, 463–477.

(21) Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M. Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials* **2019**, *18*, 435–441.

(22) Mehta, S.; Laghuvarapu, S.; Pathak, Y.; Sethi, A.; Alvala, M.; Priyakumar, U. D. MEMES: Machine learning framework for Enhanced MolEcular Screening. *Chem. Sci.* **2021**, *12*, 11710–11721.

(23) Devata, S.; Sridharan, B.; Mehta, S.; Pathak, Y.; Laghuvarapu, S.; Varma, G.; Priyakumar, U. D. DeepSPInN – deep reinforcement learning for molecular structure prediction from infrared and 13C NMR spectra. *Digital Discovery* **2024**, *3*, 818–829.

(24) Manzhos, S.; Carrington, T. Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chemical Reviews* **2021**, *121*, 10187–10217.

(25) Pattnaik, P.; Raghunathan, S.; Kalluri, T.; Bhimalapuram, P.; Jawahar, C. V.; Priyakumar, U. D. Machine Learning for Accurate Force Calculations in Molecular Dynamics Simulations. *The Journal of Physical Chemistry A* **2020**, *124*, 6954–6967.

(26) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry* **2020**, *71*, 361–390, PMID: 32092281.

(27) Samaga, Y. B. L.; Raghunathan, S.; Priyakumar, U. D. SCONES: Self-Consistent Neural Network for Protein Stability Prediction Upon Mutation. *The Journal of Physical Chemistry B* **2021**, *125*, 10657–10671.

(28) Aggarwal, R.; Gupta, A.; Chelur, V.; Jawahar, C. V.; Priyakumar, U. D. DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2022**, *62*, 5069–5079.

(29) Pathak, Y.; Laghuvarapu, S.; Mehta, S.; Priyakumar, U. D. Chemically Interpretable Graph Interaction Network for Prediction of Pharmacokinetic Properties of Drug-Like

Molecules. *Proceedings of the AAAI Conference on Artificial Intelligence* **2020**, *34*, 873–880.

(30) Laghuvarapu, S.; Pathak, Y.; Priyakumar, U. D. BAND NN: A Deep Learning Framework for Energy Prediction and Geometry Optimization of Organic Small Molecules. *Journal of Computational Chemistry* **2020**, *41*, 790–799.

(31) Goel, M.; Raghunathan, S.; Laghuvarapu, S.; Priyakumar, U. D. MoleGuLAR: Molecule Generation Using Reinforcement Learning with Alternating Rewards. *Journal of Chemical Information and Modeling* **2021**, *61*, 5815–5826.

(32) Takefuji, Y.; Chen, L. Parallel algorithms for finding a near-maximum independent set of. *IEEE Trans. Neural Networks* **1990**, *1*, 263.

(33) Steeg, E. *chapter Neural Networks, Adaptive Optimization, and RNA Secondary Structure Prediction*; American Association for Artificial Intelligence, 1993; pp 121–60.

(34) Xia, T.; SantaLucia Jr, J.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson- Crick base pairs. *Biochemistry* **1998**, *37*, 14719–14735.

(35) Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710.

(36) Klukowski, P.; Riek, R.; Güntert, P. Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nature Communications* **2022**, *13*, 6151.

(37) Singh, J.; Hanson, J.; Paliwal, K.; Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications* **2019**, *10*, 5407.

(38) Chen, X.; Li, Y.; Umarov, R.; Gao, X.; Song, L. RNA secondary structure prediction by learning unrolled algorithms. *arXiv preprint arXiv:2002.05810* **2020**,

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

(40) Singh, J.; Paliwal, K.; Singh, J.; Zhou, Y. RNA Backbone Torsion and Pseudotorsion Angle Prediction Using Dilated Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2021**, *61*, 2610–2622.

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

(42) Wadley, L. M.; Keating, K. S.; Duarte, C. M.; Pyle, A. M. Evaluating and Learning from RNA Pseudotorsional Space: Quantitative Validation of a Reduced Representation for RNA Structure. *Journal of Molecular Biology* **2007**, *372*, 942–957.

(43) He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. Computer Vision – ECCV 2016. Cham, 2016; pp 630–645.

(44) Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. 2016.

(45) Rose, P. W. et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research* **2016**, *45*, D271–D281.

(46) Shen, T.; Hu, Z.; Peng, Z.; Chen, J.; Xiong, P.; Hong, L.; Zheng, L.; Wang, Y.; King, I.; Wang, S.; others E2Efold-3D: end-to-end deep learning method for accurate de novo RNA 3D structure prediction. *arXiv preprint arXiv:2207.01586* **2022**,

(47) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.;

Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.

(48) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; pp 770–778.

(49) Leontis, N. B.; Zirbel, C. L. *RNA 3D Structure Analysis and Prediction*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 281–298.

(50) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(51) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152.

(52) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **1990**, *215*, 403–410.

(53) Magnus, M.; Antczak, M.; Zok, T.; Wiedemann, J.; Lukasiak, P.; Cao, Y.; Bujnicki, J. M.; Westhof, E.; Szachniuk, M.; Miao, Z. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Research* **2019**, *48*, 576–588.

(54) Cruz, J. A. et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* **2012**, *18*, 610–625.

(55) Miao, Z. et al. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **2015**, *21*, 1066–1084.

(56) Miao, Z. et al. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **2017**, *23*, 655–672.

(57) Miao, Z. et al. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA* **2020**, *26*, 982–995.

(58) Lu, X.; Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research* **2003**, *31*, 5108–5121.

(59) Lu, X.-J.; Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols* **2008**, *3*, 1213–1227.

(60) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,

(61) Xu, Y.-C.; ShangGuan, T.-J.; Ding, X.-M.; Cheung, N. J. Accurate prediction of protein torsion angles using evolutionary signatures and recurrent neural network. *Scientific reports* **2021**, *11*, 21033.

(62) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(63) Capriotti, E.; Norambuena, T.; Marti-Renom, M. A.; Melo, F. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics* **2011**, *27*, 1086–1093.

(64) Sali, A.; Blundell, T. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **1993**, *234*, 779.

(65) Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research* **2003**, *31*, 3370–3374.