# RNA-TorsionBERT: leveraging language models for RNA 3D torsion angles prediction

**Clément Bernard[1,2], Guillaume Postic[1], Sahar Ghannay[2], and Fariza Tahi[1]**

[1]Université Paris Saclay, Univ Evry, IBISC, 91020, Evry-Courcouronnes, France

[2]LISN - CNRS/Université Paris-Saclay, France, 91400, Orsay, France

**Predicting the 3D structure of RNA is an ongoing challenge that has yet to be completely addressed despite continuous advancements. Current approaches often incorporate secondary structure or multiple sequence alignment as features to guide the predictions. Enhancements to predictive models can be achieved by incorporating additional features such as torsional and pseudo-torsional angles. This paper presents a novel approach for directly predicting RNA torsional angles from raw sequence data. Our method draws inspiration from the successful application of language models in various domains and adapts them to RNA. We have developed a language-based model, RNA-TorsionBERT, incorporating better sequential interactions for predicting RNA torsional and pseudo-torsional angles from the sequence only. Through extensive benchmarking, we demonstrate that our method improves the averaged mean absolute error (MAE) by 7.4°, compared to the state-of-the-art. In addition, by using our predictive model, we have inferred a torsion angle-dependent scoring function, called RNA-Torsion-A, that replaces the true reference angles by our model prediction. We show that it accurately evaluates the quality of near-native predicted structures, in terms of RNA backbone torsion and pseudo-torsion angle values. Our work demonstrates promising results, suggesting the potential utility of language models in advancing RNA 3D structure prediction.**

**The source code is freely available on the EvryRNA platform: https://evryrna.ibisc.univ-evry.fr/evryrna/RNA-TorsionBERT.**

**RNA 3D structure | Torsional angles | Language models | Scoring function**

**Correspondence:** ***fariza.tahi@univ-evry.fr***

## 1. Introduction

RNA is a macromolecule that plays various biological functions in organisms. Similarly to proteins, the biological function of an RNA may be directly linked to its 3D structure. Experimental methods such as NMR, X-ray crystallography, or cryo-EM (1) can determine the 3D structure of RNAs, but they remain tedious in cost and time. Computational methods have been developed for predicting the 3D structure from the sequence, with three different approaches: *ab initio*, template-based and deep learning-based (2). Currently, no existing method matches the performance of AlphaFold for proteins (3), as shown with the last results on the CASP-RNA challenge (4). Reaching AlphaFold's level of accuracy is a long shot, notably due to the lack of data (5). AlphaFold 2 also used torsional angles to reconstruct the 3D structure (3), which is not adaptable to RNA. Indeed, a nucleotide comprises a sugar, a phosphate, and a nitrogenous base and can characterize the backbone structure by seven angles (Figure 1 A) or two pseudo-torsion angles (Figure 1 B) compared to two angles for proteins.

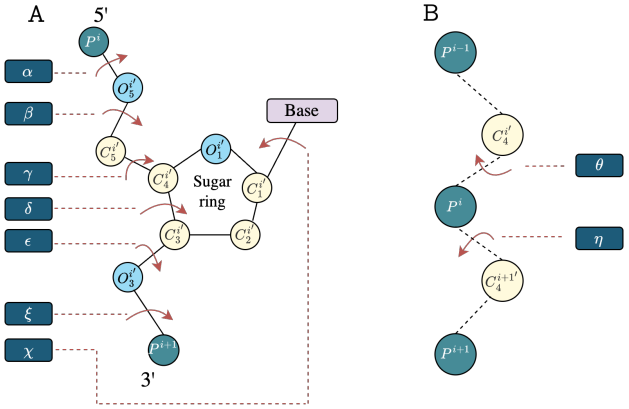


**Figure 1.** (A) RNA backbone torsional angles. The angles are defined around $O3'_{i-1}/P_i/O5'_i/C5'_i$ ($\alpha$), $P_i/O5'_i/C5'_i/C4'_i$ ($\beta$), $O5'_i/C5'_i/C4'_i/C3'_i$ ($\gamma$), $C5'_i/C4'_i/C3'_i/O3'_i$ ($\delta$), $C4'_i/C3'_i/O3'_i/P_{i+1}$ ($\epsilon$), $C3'_i/O3'_i/P_{i+1}/O5'_{i+1}$ ($\zeta$) and the rotation of the base relative to the sugar ($\chi$) $O4'_i/C1'_i/N1_i/C2_i$ for pyrimidines and $O4'_i/C1'_i/N9_i/C4_i$ for purines. (B) RNA pseudo-torsional angles. $\eta$ is defined around $C4'_{i-1}/P_i/C4'_i/P_{i+1}$ and $\theta$ around $P_i/C4'_i/P_{i+1}/C4'_{i+1}$.

The features of existing methods for the prediction of RNA 3D structure are the sequence, the secondary structure, or the multiple sequence alignment (MSA). In recent deep learning methods for the prediction of RNA 3D structure (6–10), they integrate the prediction of distance maps or geometry modules (with torsional angles) into their architecture. However, the contribution of their torsion angle predictions to the quality of the whole predicted structure remains unknown. They do not always predict all the torsional angles, and these angle predictions are hidden in their full model. Their results are not yet to exceed *ab initio* or template-based methods in the CASP-RNA challenge (4). Even the last release of AlphaFold, named AlphaFold 3 (11), has limited results on the CASP-RNA challenge. To widen the scope of molecules, AlphaFold 3 removed the prediction of torsional angles from its previous version AlphaFold 2 (3). It remains unclear if the use of torsional angles would have improved the prediction for RNA structures, as it has not been compared in AlphaFold 3.

With the recent success of language models (12) for sequential data, interest has been raised in adapting them to bi-

ological sequences. DNABERT (13) and its recent version DNABERT 2 (14) have been developed to construct a robust language model for DNA. Another BERT-based method, RNABERT (15), has been developed for RNA. For the proteins, EMSFold (16) has built a successful language model (without the use of MSA) for protein 3D structure prediction, which achieves competitive results. For RNA, methods have been developed to leverage language-based approaches for RNA structural predictions, like RNA-FM (17), Uni-RNA (18) or RNA-MSM (19). Nonetheless, they require multiple sequence alignment (MSA) as inputs, which restricts the use for RNAs. It prevents the prediction of orphan RNAs (like non-coding RNAs), as many RNA families are still unknown. The development of language models from sequence only for RNA remains very limited. No existing language models have been adapted to infer structural features from the sequence alone. To the best of our knowledge, the only method that uses the sequence to predict RNA torsional angles is SPOT-RNA-1D (20). It does not use a language model but rather a residual convolutional approach.

The torsional angles can be adopted to assess the structural quality of an RNA. Knowledge-based scoring functions like 3dRNAScore (21) or rsRNASP (22) incorporate torsional angles in their potential computation. Scoring functions are useful to help predictive models in the selection of the best candidates, increasing the accuracy of RNA structural predictions.

In this article, we propose two main contributions:

- A language-based method, RNA-TorsionBERT, to predict RNA torsional and pseudo-torsional angles from its raw sequence.
- A scoring function, RNA-Torsion-A, that uses the torsional angles and RNA-TorsionBERT to assess structural quality.

This paper is organized as follows: we describe our contributions in two separate points. Then, we detail our experiments for the torsional angles prediction and the structural quality assessment tasks. We discuss the results and the limitations of our approaches before concluding the scope of our contributions. The results and the code of our RNA-TorsionBERT and RNA-Torsion-A are easily reproducible and freely available on the EvryRNA platform: https://evryrna.ibisc.univ-evry.fr/evryrna/RNA-TorsionBERT.

## 2. Methods

This section introduces the prediction problem of torsional angles before diving into language models and their interests. Then, we introduce the scoring function derived from our model and its evaluation criteria.

### A. Torsional angles prediction

***A.1. RNA-TorsionBERT approach.*** Current methods that use sequence only represent sequences as one-hot-encoding vectors. This representation may be too sparse to consider sequential interactions well. This encoding is usually associated with a convolutional neural network, which is commonly limited by long-range interactions. A solution could be using attention mechanisms (23). Attention-based architecture nonetheless requires a huge amount of data to train well, which is not the case for RNA 3D structure data. To counter this problem, we can use models pre-trained on a large amount of unlabeled data. This could bring a better input representation of the raw sequence, which could then be fine-tuned to specific tasks. These pre-trained models could input either RNA or DNA sequences. Indeed, the alphabet of DNA is nearly the same as RNA in terms of vocabulary, with only one nucleotide change: the T for DNA and the U for RNA. Their similarity in sequences could be leveraged into language-based models to create embeddings for RNA sequences. This could mitigate the lack of data from RNA and thus help a model start from an embedding that already has a meaning.

Recent advances in language models started with BERT (12), where the model was pre-trained on masking and next-sentence prediction tasks before being fine-tuned on diverse specific language tasks. DNA or RNA can be seen as a sequence of nucleotides, where their interactions have a biological meaning. Therefore, methods have been adapted from BERT to develop a language-based architecture for either RNA or DNA. The aim is to reproduce the success of language comprehension for another language. As the size of the vocabulary is different, modifications should be made to fit the current language. An example for DNA is DNABERT (13), where the training process was updated compared to the original BERT by removing the next sentence prediction and taking $K$-mers as inputs (contiguous sequence of k nucleotides). It was trained on human-genome data. Its second version, named DNABERT 2 (14), uses Byte Pair Encoding (BPE) instead of $K$-mers. An example of adaptation of BERT for RNA is called RNABERT (15). It is a six-transformer layer pre-trained on two tasks: masking and structural alignment learning (SAL). RNABERT was trained on 76,237 human-derived small ncRNAs.

The aim of our method is, given a pre-trained language model (DNABERT or RNABERT), to adapt its neuronal weights to predict RNA torsional and pseudo-torsional angles from the sequence. We have added layers to adapt the methods to our multi-token label regression task. Each token in the input would have 18 labels: two values (sine and cosine) for each of the seven torsional angles and two pseudo-torsional angles. The use of pre-trained embedding would help the model not to start from scratch and update the learned attention layers for RNA structural features.

***A.2. RNA-TorsionBERT architecture.*** The architecture of our method, when based on DNABERT, is described in Figure 2 (illustrated with 3-mers). An input sequence of size L is tokenized and then fed to the network with token and positional embeddings. The tokenization process usually adds specific tokens (like the *CLS* and *PAD* tokens). As DNABERT could input a maximum of 512 tokens, we set the maximum sequence length to 512 nucleotides. The last hidden state is set

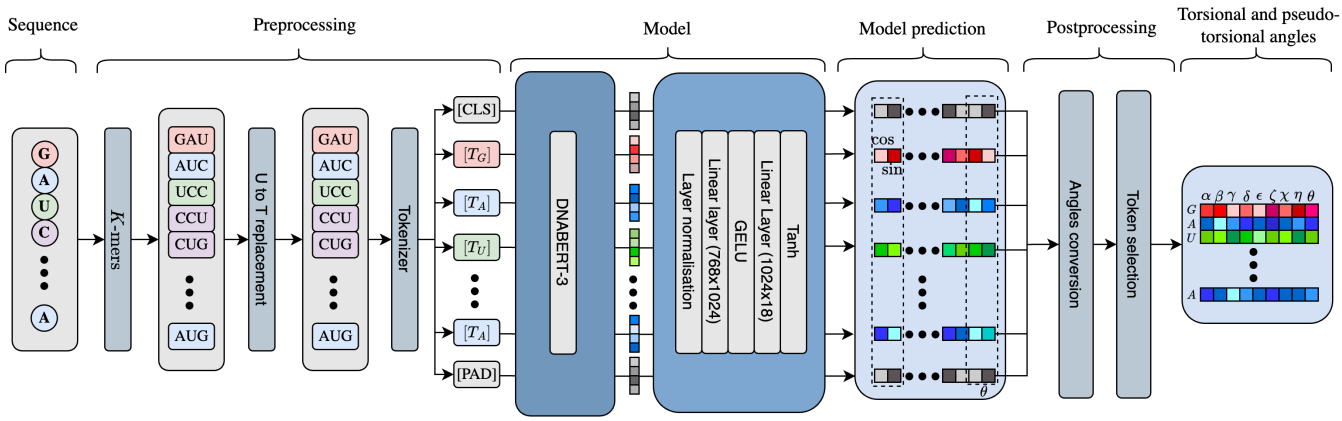


**Figure 2.** Schema of the proposed language model architecture for torsional and pseudo-torsional angles prediction. Given an RNA sequence, preprocessing is applied to map each sequence's nucleotide into a token with pre-trained embeddings from the language model. *CLS* and *PAD* tokens are added to the sequence tokens. We convert the Uracil (U) with its equivalent in DNA: Thymine (T) for DNABERT model. Then, the language model (DNABERT or RNABERT) will output hidden states with a representation of each token. This is fed into extra layers before entering a last Tanh activation to have the cosine and sine per angle. A postprocessing is required to convert back to angles from the sine and cosine.

to be 768 by the original DNABERT architecture. We then apply extra layers to map the hidden state outputs to the desired final output dimension (Lx18). These extra layers comprise layer normalisation, a linear layer (from 768 to 1024), a GELU activation, another linear layer (1024 to 18), and a Tanh final activation. The final output layer is of size 18 because it outputs a sine and a cosine for the seven and two pseudo-torsional angles. It allows the relief of the periodicity of the different angles. The Tanh activation maps the outputs to the cosine and sine range (between -1 and +1), which is then converted into angle predictions using the formula $\alpha = \tan^{-1}(\frac{\sin(\alpha)}{\cos(\alpha)})$ (adaptable for the other angles). Details on the training process are in the Supplementary file.

## B. Model quality assessment based on torsional angles

***B.1. RNA-Torsion-A approach.*** Existing metrics have been developed to assess the quality of predicted RNA 3D structures with access to a reference. The most famous one is the RMSD (root-mean-square deviation), which assesses the general folding of structures like $\epsilon$RMSD (24). $\epsilon$RMSD adds key RNA specifities to the RMSD computation. Protein-based approaches have also been adapted to RNA, like the TM-score (25), which computes alignment and normalized distances. The GDT-TS (26) counts the number of overlaps for different thresholds, while the lDDT (27) assesses the interatomic distances. The CAD-score (28) evaluates the contact-area differences between two structures. More RNA-specific metrics have been introduced, like $\text{INF}_{all}$ (29) (interaction network fidelity) that counts the number of key RNA interactions (Watson-Crick, non-Watson-Crick, stacking). It has been extended to the DI (29) (deformation index) with the incorporation of the RMSD. Finally, a torsional-angles-based metric, named MCQ (30) (mean of circular quantities), has been introduced. The MCQ computes the deviation in angle space with the superposition of structures and complements other existing metrics.

The metric assessment of RNA 3D structures requires two structures, with one being the experimentally solved structure. Having this reference is a strong asset that can not be available when developing prediction models. To rank near-native structures without a reference, scoring functions have been developed throughout the years. They usually adapt free-energy like RASP (31), $\epsilon$SCORE (24), DFIRE-RNA (32), 3dRNAScore (21), or recently rsRNASP (22). Other methods employ deep learning approaches like RNA3DCNN (33) or ARES (34).

To discriminate near-native structures in the torsional and pseudo-torsional space, we have derived a scoring function from our RNA-TorsionBERT model. We first have proven that the evaluation criteria used in RNA-TorsionBERT, the MAE (mean average error), has a high correlation with existing metrics. Then, we replaced the true torsional angles with the predicted angles from our model to compute the mean average error (MAE) over the near-native structure. Therefore, the MAE computation compares the prediction of our model angles with the angles from the predicted non-native structures. We named this scoring function RNA-Torsion-A (RNA torsional assessment).

***B.2. RNA-Torsion-A architecture.*** In SPOT-RNA-1D (20), the authors introduced the mean-average error (MAE) metric to assess the performance of their method SPOT-RNA-1D in the prediction of torsional and pseudo-torsional angles. Given a sequence $S$ of $L$ residues and one angle $z \in Z = \{\alpha, \beta, \gamma, \delta, \zeta, \chi, \theta, \eta\}$, the MAE is computed as

$$\text{MAE(z)} = \min(d_z, 360 - d_z)$$

with:

$$d_z = \frac{1}{L} \sum_{res \in S} |z_{res}^{pred} - z_{res}^{true}|$$

The angle $z_{res}^{pred}$ defines the predicted angle computed for the nucleotide $res$ of the sequence $S$. The angle $z_{res}^{true}$ corresponds to the angle of nucleotide $res$ experimentally determined. While it has only been considered angle per angle in SPOT-RNA-1D, we can extend this MAE by averaging the

value over the different torsional and pseudo-torsional angles.

$$\mathbf{MAE} = \frac{1}{|Z|} \sum_{z \in Z} \text{MAE(z)}$$

This formulation enables the computation of a unique value for two structures given their torsional angles. As it provides a quality score for two structures, it can be seen as a metric for RNA 3D structure quality assessment.

To evaluate the quality of a structure without a reference, we can only use features inferred from the structure. Therefore, if we use the inferred angles from the candidate structure, we can compute the MAE with the predicted angles from our RNA-TorsionBERT model (given the sequence of the structure). This MAE now becomes a scoring function, as it only takes as input a structure without any known native structure. Figure 3 shows the architecture of our scoring function. Given a structure, the DSSR (35) outputs the inferred torsional angles and the sequence. The sequence is then pre-processed by RNA-TorsionBERT, and the inference gives predicted angles. Then, we compute the MAE to finally output a structural quality measure for an RNA 3D structure. Schema of the MAE as a metric is displayed in Figure S3 of the Supplementary file.

## 3. Results and Discussion

### A. Results on Torsional Angles prediction

We present here the different experiments for the torsional angles prediction task. We used the MAE presented above as a criterion to assess our model performances.

***A.1. Datasets.*** To validate the performances of torsional angle prediction models, we used different datasets of native structures:

**Pre-training**: we downloaded each available PDB structure and removed the structures from the non-redundant Training, Validation and Test sets presented below. We also ensure the structures from this dataset have a sequence similarity below 80% compared to the other used datasets. This dataset aims to give an overall idea of the native structures and is then used to pre-train our language models. We considered only the structures of a maximum sequence length of 512 (DNABERT can only input 512 tokens). The final set is composed of 4,267 structures with sequences from 11 to 508 nucleotides.

**Training**: we used the training set from rsRNASP (22). It is composed of 185 3D structures with sequences between 17 and 395 nucleotides.

**Validation**: we used the validation structures from SPOT-RNA-1D (20). It contains 29 structures with sequences between 33 and 288 nucleotides.

**Test**: we combined two well-known test sets: RNA-Puzzles (36–39) and CASP-RNA (4). We combined both of these datasets as a whole Test set to assess the robustness of our model. It leads to a Test set of 52 structures (40 from RNA-Puzzles and 12 from CASP-RNA), with sequences from 27 nucleotides to 374 (we remove the RNA of PDB ID R1138 as its length exceeds 512 nucleotides).

**Table 1.** MAE (mean average error) for the Validation set averaged over the nine angles for DNABERT (3,4, 5 or 6-mers) and RNABERT.

| | DNABERT | | | | RNABERT |
|---|---|---|---|---|---|
| | 3-mer | 4-mer | 5-mer | 6-mer | 1-mer |
| MAE | **27.7** | 28.5 | 28 | 29 | 28.2 |

The distribution of the seven torsional angles and the two pseudo-torsional angles is given in Figure 4. These distributions are similar for the four datasets, meaning the learned distribution from the training set could allow good generalisation for the model.

***A.2. Language model selection.*** Multiple language-based models exist to be fine-tuned to our torsional angles prediction task. As each model has a different format of inputs ($K$-mers), we had to select the best tokenization of our RNA sequences. We also had to decide which pre-trained model was the best for our task. Therefore, we trained the same DNABERT model with the different values of K (3, 4, 5 or 6) and RNABERT on the Training set and observed the performances on the Validation set. We did not use DNABERT 2 because the inputs are encoded with byte pair encoding. It loses the nucleotide level representation, and thus, we can not predict torsional angles for each nucleotide.

The results are shown in Table 1 on the validation set. In terms of $K$-mers, DNABERT trained on 3-mers has better results (averaged MAE of 27.7°) compared to the other $K$-mers and RNABERT. We tried to improve these results by pre-training the best model, DNABERT with 3-mers, on the Pre-trained dataset before fine-tuning it on the Training set. It led to an MAE of 27.1, a mean improvement of 0.6°.

DNABERT with 3-mers outperforms RNABERT, which does not input $K$-mers. This result remains surprising as we could have thought that RNABERT, as pre-trained specifically on RNA data, could have done better than the DNABERT model. This difference might be explained by the $K$-mers representation that is used by DNABERT compared to RNABERT, where the size of the vocabulary is extended, and thus a finer representation of the inputs is embedded.

For the rest of the article, we name the DNABERT with 3-mers firstly trained on the Pre-training dataset and then fine-tuned on the Training dataset RNA-TorsionBERT (for RNA torsional BERT).

***A.3. Performances.*** We present here the prediction results obtained by our method RNA-TorsionBERT on the Test set (presented above) compared to the state-of-the-art approach SPOT-RNA-1D (20), which is the only method that predicts

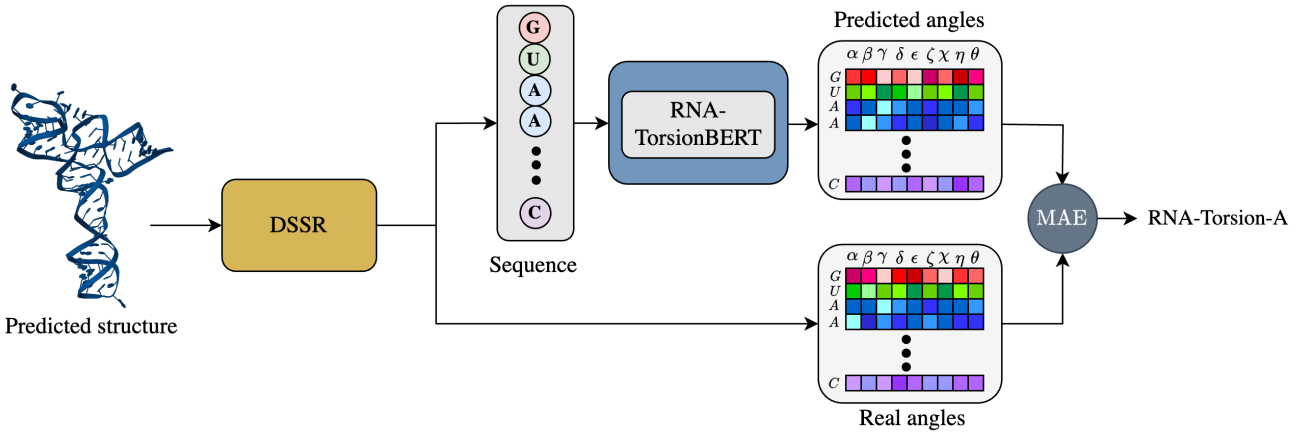


**Figure 3.** RNA-Torsion-A scoring function computation. Given a predicted RNA 3D structure, the DSSR (35) software gives the sequence and calculates angles. The sequence is fed to our RNA-TorsionBERT model to predict the different angles. The scoring function is then computed by taking the MAE between the predicted angles and the angles read from the RNA 3D structure.

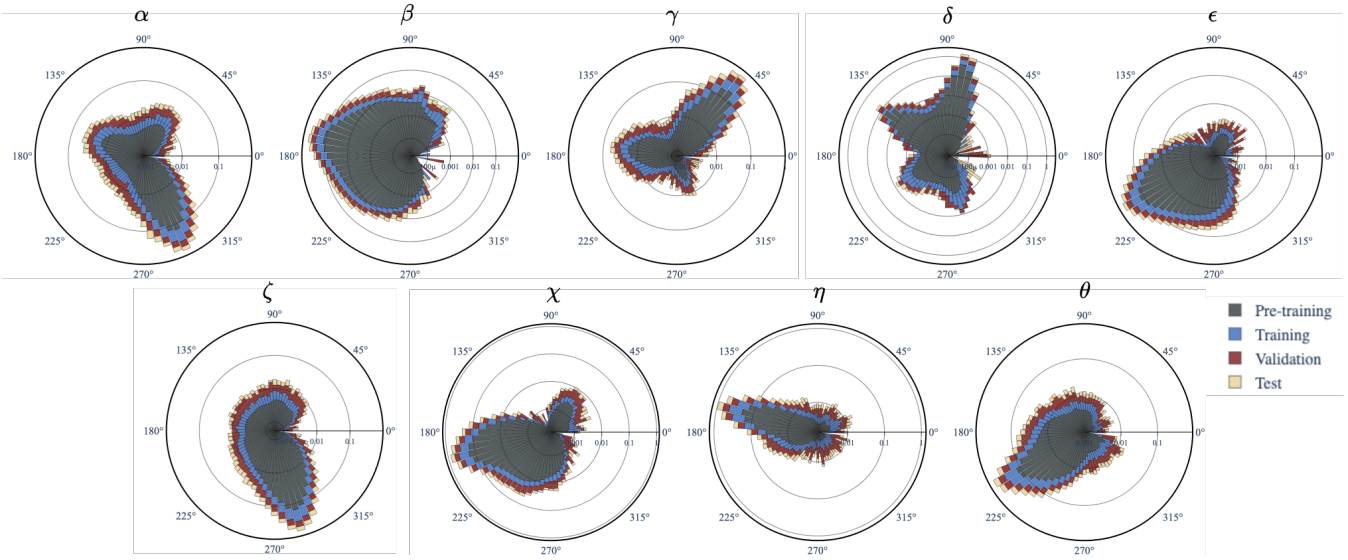


**Figure 4.** Polar distribution of the seven torsional angles ($\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$, $\zeta$, $\chi$) and the two pseudo-torsional angles ($\eta$ and $\theta$) for the Pre-training, Training, Validation and Test datasets. For each angle, the logarithm of the normalized count is depicted.

**Table 2.** MAE per angle averaged over the Test set for RNA-TorsionBERT compared to the state-of-the-art method SPOT-RNA-1D (20). We also include inferred torsional angles from state-of-the-art methods that predict RNA 3D structures. The MAE was computed for those methods only on single-stranded structures, as in State-of-the-RNArt (2). Methods are sorted by mean MAE.

| | Method type | $\alpha$ | $\beta$ | $\gamma$ | $\delta$ | $\epsilon$ | $\zeta$ | $\chi$ | $\eta$ | $\theta$ | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RNA-TorsionBERT | Deep Learning | **37.3** | **19.6** | **29.4** | **13.6** | **16.6** | **26.6** | **14.7** | **20.1** | **25.4** | **22.6** |
| SPOT-RNA-1D (20) | Deep Learning | 50.7 | 30.1 | 35.8 | 28.3 | 21.9 | 29.8 | 20.9 | 25.0 | 27.5 | 30.0 |
| IsRNA1 (40) | *Ab initio* | 53.1 | 26.2 | 39.2 | 14.6 | 25.9 | 38.4 | 19.0 | 27.6 | 31.7 | 30.6 |
| AlphaFold 3 (11) | Deep Learning | 51.2 | 26.5 | 38.0 | 17.0 | 21.3 | 38.3 | 20.5 | 30.2 | 36.3 | 31.0 |
| RNAJP (41) | *Ab initio* | 56.8 | 31.6 | 41.9 | 18.6 | 27.7 | 42.1 | 21.1 | 29.8 | 33.3 | 33.7 |
| Vfold-Pipeline (42) | Template-based | 55.4 | 28.1 | 38.8 | 20.1 | 32.2 | 43.1 | 25.0 | 34.2 | 41.5 | 35.4 |
| RNAComposer (43) | Template-based | 58.9 | 31.5 | 48.2 | 17.9 | 27.1 | 40.0 | 24.8 | 32.8 | 38.6 | 35.5 |
| MC-Sym (44) | Template-based | 72.2 | 31.2 | 61.5 | 31.8 | 25.6 | 49.4 | 29.6 | 41.2 | 44.6 | 43.0 |
| 3dRNA (45) | Template-based | 65.7 | 40.0 | 53.6 | 33.7 | 40.3 | 52.7 | 27.2 | 40.2 | 45.6 | 44.3 |
| trRosettaRNA (8) | Deep Learning | 67.6 | 36.8 | 67.0 | 26.4 | 32.9 | 54.3 | 58.3 | 38.2 | 48.3 | 47.8 |
| Rhofold (46) | Deep Learning | 94.3 | 66.0 | 76.2 | 60.6 | 52.8 | 72.6 | 32.4 | 42.0 | 46.6 | 60.4 |

RNA torsional angles from the sequence only. We reproduced the architecture of SPOT-RNA-1D (because we only had the code to do inference) and trained it with the exact same data as RNA-TorsionBERT. We also included the results for the inferred angles from methods benchmarked in State-of-the-RNArt (2). The methods included either *ab initio* with IsRNA (40) and RNAJP (41), or template-based with RNAComposer (43), Vfold-Pipeline (42), MC-Sym (44) and 3dRNA (45). We also include three deep learning methods: trRosettaRNA (8) and RhoFold (46) and the newly AlphaFold 3 (11).

We report the averaged angle MAE on the Test Set in Table 2. Details of the MAE for each angle and each structure are reported in Figure S1 of the Supplementary file. Our RNA-TorsionBERT model has better performances than SPOT-RNA-1D for every angle. It has an average MAE on all angles of 22.6° compared to 30° for SPOT-RNA-1D. The MAE improvement ranges between 2.1° (for $\theta$) and 14.7° (for $\delta$). It also outperforms the angles inferred from state-of-the-art methods for RNA 3D structure prediction, including the last published method, AlphaFold 3 (11). trRosettaRNA and RhoFold, two deep learning methods, have the worst MAE compared to *ab initio* and template-based approaches. It can be explained by the use of physics on *ab initio* and template-based methods that are inferred in the torsional angles. The use of deep learning approaches might have the counterpart to not include physics enough, except for AlphaFold 3. Deep learning methods, while having the best overall results, as shown in the benchmark done in State-of-the-RNArt (2), remain limited in torsional angle predictions. AlphaFold 3 outperforms the other deep learning approaches for the inferred torsional angles but is beaten by an *ab initio* method (IsRNA1), SPOT-RNA-1D and our RNA-TorsionBERT approaches. Removing torsional angle predictions in their architecture, which should have been adapted if used for RNAs, might explain the current performances.

Figure 5 reports the MAE averaged over the nine angles obtained by our method and by SPOT-RNA-1D on each RNA of the Test set. We can see our method achieves more frequently MAE below 20° than SPOT-RNA-1D, and sometimes achieves MAE below 10° (which is not the case for SPOT-RNA-1D). For sequences higher than 230 nucleotides, our method seems similar to SPOT-RNA-1D.

Details on examples of structures with good and bad performances in terms of MAE from RNA-TorsionBERT are shown in Figure S1 of the Supplementary file. Specific MAE values for each RNA structure are displayed in Figure S2 of the Supplementary file.

## B. Model quality assessment based on torsional angles

In this part, we describe datasets of decoys (near-native structures) used to compare the correlation between the existing metrics with the MAE. Then, we discuss the correlation measures used to link the metrics and the scoring functions. Finally, we present the results of our scoring function, RNA-Torsion-A, compared to existing scoring functions on the decoy test sets.

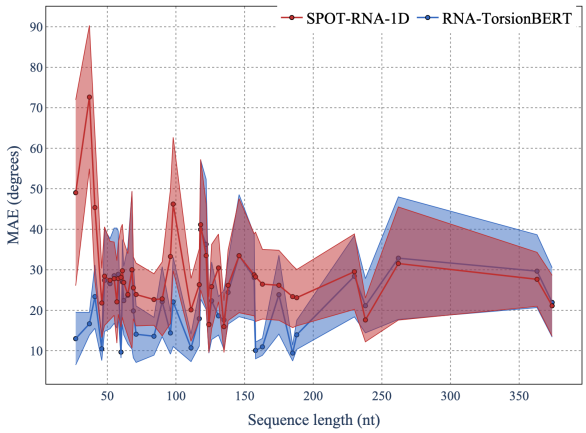


**Figure 5.** Mean MAE and standard deviation averaged over the angles depending on the sequence length of the RNAs from the Test set for RNA-TorsionBERT and SPOT-RNA-1D.

***B.1. Datasets.*** Datasets of near-native structures (or decoys) are necessary to compare model quality assessment metrics. Indeed, scoring functions are used to discriminate between near-native structures, meaning that we need to have non-native structures to evaluate the quality of our scoring function.

We used three different datasets with different strategies of structure generation:

**Decoy Test Set I** is from RASP (31), composed of 85 native RNAs with decoys generated with MODELLER (31). MODELLER is a predictive model that is used to create decoys with different sets of Gaussian restraint parameters. Each RNA has 500 decoys, which are close to the native structure. We only kept 83 RNAs and removed the two RNAs that have sequence lengths higher than 512 nucleotides (PDB ID: 3df3A and 3f1hA).

**Decoy Test Set II** corresponds to the prediction-models (PM) subset from rsRNASP (22). It has 20 non-redundant single-stranded RNAs. For each RNA, 40 decoys are generated with four RNA 3D structure prediction models: FARFAR 2 (47), RNAComposer (43), SimRNA (48) and 3dRNAv2.0 (45). The decoys are not as near to native structures as with the Decoy Test Set I.

**Decoy Test Set III** is the RNA-Puzzles standardized dataset (49). This dataset comprises 21 RNAs and dozens of decoy structures for each RNA. The decoys are not all close to the native structures.

***B.2. Evaluation measures.*** Scoring functions aim to discriminate near-native structures. The Pearson correlation coefficient (PCC) and the enrichment score (ES) are used to assess the correctness of a given scoring function. They assess the link between a scoring function and a given metric.

The Pearson coefficient correlation (PCC) is computed between the ranked structures based on scoring functions and

structures ranked by metrics. It is defined as:

$$PCC = \frac{\sum_{i=1}^{N_{decoys}} (E_n - \bar{E})(R_n - \bar{R})}{\sqrt{\sum_{n=1}^{N_{decoys}} (E_n - \bar{E})^2} \sqrt{\sum_{n=1}^{N_{decoys}} (R_n - \bar{R})^2}}$$

where $E_n$ is the energy of the $n$th structure, and $R_n$ the metric of the $n$th structure. PCC ranges from 0 to 1, where a PCC of 1 means the relationship between metric and energy is completely linear.

The enrichment score (ES) considers top-ranked structures from both scoring function and metric. It is defined as:

$$ES = 100 \times \frac{|E_{top10\%} \cap R_{top10\%}|}{N_{decoys}}$$

where $|E_{top10\%} \cap R_{top10\%}|$ is the number of common structures from the top 10% of structures (measured by the metric) and the top 10% of structures with the lowest scoring function. ES ranges between 0 and 10 (perfect scoring). An enrichment score of 1 means a random prediction, whereas below 1 means a bad score.

***B.3. MAE as model quality assessment.*** As recently done in RNAdvisor (50), where we benchmarked the existing scoring functions and metrics, we computed for the three Decoys sets the PCC and ES between nine existing metrics. To depict the legitimacy of the MAE as a model quality assessment, we added it as a metric to the previous ones. Results are shown in Figure 6. We observe a correlation of 0.96 in terms of PCC between MAE and the MCQ, and an ES of 6.77. It means the MAE has a high correlation with the MCQ while having a lower correlation to other existing metrics (PCC below 0.7 and ES below 4.8 for the other metrics). It is the best PCC correlation over the existing metrics and one of the best in terms of ES. This means that the MAE could be a valuable metric compared to MCQ, and thus, it could be used to assess the quality of RNA 3D structures in terms of angle deviation and inspiring a scoring function.

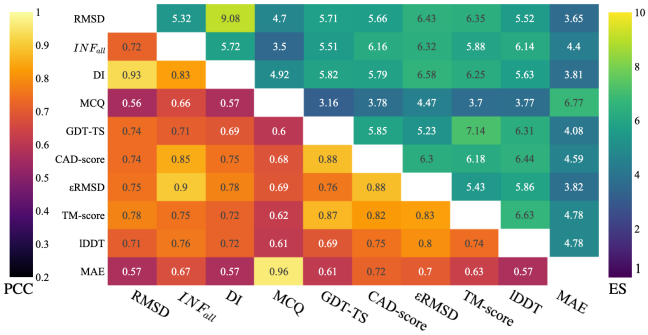


**Figure 6.** PCC (left) and ES (right) scores for ten metrics averaged over the three Decoys sets. The lower half of the matrix represents the PCC, while the upper half corresponds to the ES score. The diagonal has a PCC of 1 and ES of 10. The lighter the colour, the better the correlation.

***B.4. RNA-Torsion-A as scoring function.*** To assess the validity of our scoring function, we computed with RNAdvisor (50) the available scoring functions RASP (31), $\epsilon$SCORE (24), DFIRE-RNA (32) and rsRNASP (22) for the three different Decoys test sets.

We compared RNA-Torsion-A with the state-of-the-art scoring functions using PCC and ES over nine metrics and the MAE. The averaged values are shown in Figure 7. RNA-Torsion-A has a high correlation with MCQ and MAE, the highest compared to the other existing metrics (PCC/ES of 0.77/3.73 and 0.78/4.5, respectively, for MCQ and MAE). These results show that the RNA-Torsion-A scoring function correlates highly with MCQ, outperforming the other scoring functions for reproducing quality assessment based on torsional and pseudo-torsional angles. It outperforms rsRNASP, which is the only state-of-the-art scoring function that includes RNA torsional angles, in terms of correlation to the MCQ (PCC of 0.77 for RNA-Torsion-A compared to 0.66 for rsRNASP). Nonetheless, RNA-Torsion-A is not as performant as the other scoring functions compared to the other distance-based existing metrics (see more details in Figure S2 of the Supplementary file). Indeed, as it only considers the torsional features, our scoring function allows the comparison in angle space but does not enable atom distance consideration.

## 4. Conclusion

In this work, we have developed a language-based model, RNA-TorsionBERT, to predict RNA torsional and pseudo-torsional angles from the sequence. With a pre-trained model, the learned embeddings have been used as a starting point to infer structural features from the sequence. We have achieved improvement compared to SPOT-RNA-1D, the only tool for RNA torsional angle prediction from the raw sequence.

Through an extensive benchmark of state-of-the-art methods, we have outperformed the angles inferred from the predictive models. We have also included in the benchmark the new release of AlphaFold, named AlphaFold 3, which has competitive results to *ab initio* solutions in terms of MAE on inferred angles. Our method, RNA-TorsionBERT, remains better for the prediction of RNA torsional angles with only the sequence as input.

Through the evaluation of our model for backbone torsional angles prediction, we have extended this evaluation as a model quality assessment for RNA 3D structures. Then, we have inferred a scoring function named RNA-Torsion-A. This scoring function could help the selection of near-native structures in terms of angle deviation. It is also specific to torsional angles and, thus, is more related to the MCQ (mean of circular quantities) than the other atom-deviation-based metrics. RNA-Torsion-A is, therefore, the best scoring function to reproduce the MCQ metric, but it should not be chosen to replicate distance-based metrics. The quality of the scoring function could be enhanced by incorporating distance atom features, or directly by improving the prediction of torsional angles itself.

The predictions of torsional and pseudo-torsional angles to help reconstruct the RNA 3D structure directly remain unsolved for the moment. The RNA 3D structure is very sensi-

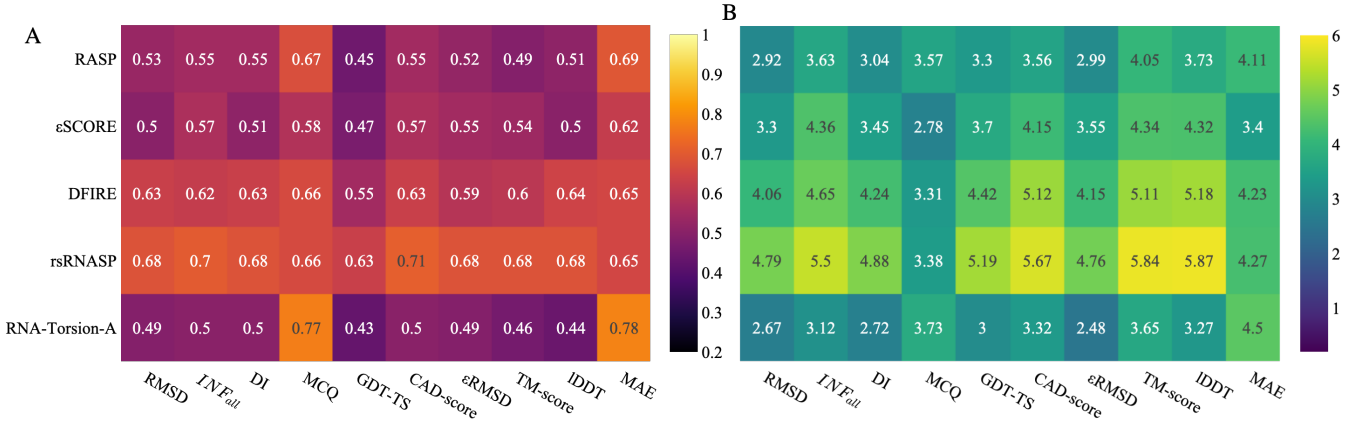


**Figure 7.** PCC (A) and ES (B) between five different scoring functions, RASP, $\epsilon$SCORE, DIFRE-RNA, rsRNASP and our scoring function RNA-Torsion-A with ten metrics, RMSD, INF$_{all}$, DI, MCQ, GDT-TS, CAD-score, $\epsilon$RMSD, TM-score, lDDT and MAE. Values are averaged over the three decoy test sets. The lighter the colour, the better the correlation.

tive to small angle deviations, which means our model does not have enough accuracy to help guide the reconstruction. It can nonetheless be used as a starting point for the reconstruction of RNA 3D structures, with *ab initio* methods, for instance, that include molecular dynamics to relax the structure. It could also be used as a feature in a bigger network to predict RNA 3D conformation. Sugar conformation could also be added to improve the structural features predicted.

**ACKNOWLEDGEMENTS**

This work is supported in part by UDOPIA-ANR-20-THIA-0013, Labex DigiCosme (project ANR11LABEX0045DIGICOSME), performed using HPC resources from GENCI/IDRIS (grant AD011014250), and operated by ANR as part of the program "Investissement d'Avenir" Idex ParisSaclay (ANR11IDEX000302).

*Conflict of Interest*: None is declared.

## Bibliography

1. Kailong Zhang, Shuai Li, Kristina Kappel, et al. Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution. *Nature Communications*, 10:5511, 2019.
2. Clément Bernard, Guillaume Postic, Sahar Ghannay, and Fariza Tahi. State-of-the-RNArt: benchmarking current methods for RNA 3D structure prediction. *NAR Genomics and Bioinformatics*, 6(2):lqae048, June 2024. doi: 10.1093/nargab/lqae048.
3. John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 8 2021.
4. R Das, RC Kretsch, AJ Simpkin, et al. Assessment of three-dimensional RNA structure prediction in CASP15. *Proteins*, 91(12):1747–1770, 2023.
5. Bohdan Schneider, Blake Alexander Sweeney, Alex Bateman, Jiri Cerny, Tomasz Zok, and Marta Szachniuk. When will RNA get its AlphaFold moment? *Nucleic Acids Research*, 51 (18):9522–9532, 2023. doi: 10.1093/nar/gkad726.
6. Robin Pearce, Gilbert S. Omenn, and Yang Zhang. De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. *bioRxiv*, May 2022.
7. Y. Kagaya, Z. Zhang, N. Ibtehaz, X. Wang, T. Nakamura, D. Huang, and D. Kihara. NuFold: A Novel Tertiary RNA Structure Prediction Method Using Deep Learning with Flexible Nucleobase Center Representation. *bioRxiv*, September 2023. doi: 10.1101/2023.09.20.558715. Preprint.
8. W. Wang, C. Feng, R. Han, et al. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nat Commun*, 14:7266, 2023.
9. Minkyung Baek, Reid McHugh, Ivan Anishchenko, et al. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, 21:117–121, 2024. doi: 10.1038/s41592-023-02086-5.
10. Y. Li, C. Zhang, C. Feng, et al. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nature Communications*, 14:5745, 2023.
11. J. Abramson, J. Adler, J. Dunger, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024. doi: 10.1038/s41586-024-07487-w.
12. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
13. Yanrong Ji, Zhihan Zhou, Han Liu, et al. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37: 2112–2120, 8 2021.
14. Zhihan Zhou, Yanrong Ji, Weijian Li, et al. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome, 2023.
15. Manato Akiyama and Yasubumi Sakakibara. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics*, 4(1):lqac012, 02 2022.
16. Zeming Lin and et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023.
17. Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. 2022.
18. Xi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, and Han Wen. UNI-RNA: UNIVERSAL PRE-TRAINED MODELS REVOLUTIONIZE RNA RESEARCH. *bioRxiv*, 2023. doi: 10.1101/2023.07.11.548588.
19. Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, Yonghong Tian, Jian Zhan, Jie Chen, and Yaoqi Zhou. Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3, 2024. doi: 10.1093/nar/gkad1031.
20. Jaswinder Singh, Kuldip Paliwal, Jaspreet Singh, et al. RNA Backbone Torsion and Pseudo-torsion Angle Prediction Using Dilated Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, 61:2610–2622, 6 2021.
21. Jian Wang, Yunjie Zhao, Chunyan Zhu, et al. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Research*, 43:e63–e63, 5 2015.
22. Ya-Lan Tan, Xunxun Wang, Ya-Zhou Shi, et al. rsRNASP: A residue-separation-based statistical potential for RNA 3D structure evaluation. *Biophysical Journal*, 121:142–156, 1 2022.
23. Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
24. Sandro Bottaro, Francesco Di Palma, and Giovanni Bussi. The Role of Nucleobase Interactions in RNA Structure and Dynamics. *Nucleic acids research*, 42, 10 2014.
25. Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702—710, December 2004.
26. Adam Zemla, Ceslovas Venclovas, John Moult, et al. Processing and analysis of CASP3 protein structure predictions. *Proteins: Structure, Function, and Bioinformatics*, 37(S3): 22–29, 1999.
27. Valerio Mariani, Marco Biasini, Alessandro Barbato, et al. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics (Oxford, England)*, 29(21):2722–2728, 2013.
28. Kliment Olechnovic, Eleonora Kulberkyte, and Ceslovas Venclovas. CAD-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins*, 81, 01 2013.
29. Marc Parisien, José Cruz, Eric Westhof, et al. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA (New York, N.Y.)*, 15:1875–85, 09 2009.
30. Tomasz Zok, Mariusz Popenda, and Marta Szachniuk. MCQ4Structures to compute similarity of molecule structures. *Central European Journal of Operations Research*, 22, 04 2013.
31. Emidio Capriotti, Tomas Norambuena, Marc A. Marti-Renom, et al. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, 27:1086–1093, 4 2011.
32. Emidio Capriotti, Tomas Norambuena, Marc A. Marti-Renom, et al. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, 27(8):1086–1093, 02 2011.
33. Jun Li, Wei Zhu, Jun Wang, et al. RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLOS Computational Biology*, 14(11):1–18, 11 2018.
34. Raphael J. L. Townshend, Stephan Eismann, Andrew M. Watkins, et al. Geometric deep learning of RNA structure. *Science*, 373:1047–1051, 8 2021.
35. Andrew Colasanti, Xiang-Jun Lu, and Wilma K. Olson. Analyzing and building nucleic acid

structures with 3DNA. *Journal of Visualized Experiments*, 74:e4401, 2013.

36. José Almeida Cruz, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay V. Dokholyan, Samuel Coulbourn Flores, Lili Huang, Christopher A. Lavender, Véronique Lisi, François Major, Katarzyna Mikolajczak, Dinshaw J. Patel, Anna Philips, Tomasz Puton, John Santalucia, Fredrick Sijenyi, Thomas Hermann, Kristian Rother, Magdalena Rother, Alexander Serganov, Marcin Skorupski, Tomasz Soltysinski, Parin Sripakdeevong, Irina Tuszynska, Kevin M. Weeks, Christina Waldsich, Michael Wildauer, Neocles B. Leontis, and Eric Westhof. Rna-puzzles : A casp-like evaluation of rna three-dimensional structure prediction. *RNA*, 18:610–625, 4 2012. ISSN 1355-8382. doi: 10.1261/rna.031054.111.
37. Zhichao Miao, Ryszard W Adamiak, Marc F Blanchet, Michal Boniecki, Janusz M Bujnicki, Shi-Jie Chen, Clarence Cheng, Grzegorz Chojnowski, Fang-Chieh Chou, Pablo Cordero, Jose Almeida Cruz, Adrian R Ferré-D'Amaré, Rhiju Das, Feng Ding, Nikolay V Dokholyan, Stanislaw Dunin-Horkawicz, Wipapat Kladwang, Andrey Krokhotin, Grzegorz Lach, Marcin Magnus, and et al. Rna-puzzles round ii: assessment of rna structure prediction programs applied to three large rna structures. *RNA*, 21(6):1066–1084, 2015. doi: 10.1261/rna.049502.114.
38. Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Robert T. Batey, Alexander J. Becka, Marcin Biesiada, Michał J. Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Fang-Chieh Chou, Adrian R. Ferré-D'Amaré, Rhiju Das, Wayne K. Dawson, Feng Ding, Nikolay V. Dokholyan, Stanisław Dunin-Horkawicz, Caleb Geniesse, Kalli Kappel, Wipapat Kladwang, Andrey Krokhotin, Grzegorz E. Łach, François Major, Thomas H. Mann, Marcin Magnus, Katarzyna Pachulska-Wieczorek, Dinshaw J. Patel, Joseph A. Piccirilli, Mariusz Popenda, Katarzyna J. Purzycka, Aiming Ren, Greggory M. Rice, John Santalucia, Joanna Sarzynska, Marta Szachniuk, Arpit Tandon, Jeremiah J. Trausch, Siqi Tian, Jian Wang, Kevin M. Weeks, Benfeard Williams, Yi Xiao, Xiaojun Xu, Dong Zhang, Tomasz Zok, and Eric Westhof. Rna-puzzles round iii: 3d rna structure prediction of five riboswitches and one ribozyme. *RNA*, 23:655–672, 5 2017. ISSN 1355-8382. doi: 10.1261/rna.060368.116.
39. Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Michał J. Boniecki, Janusz Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Yi Cheng, Fang-Chieh Chou, Rhiju Das, Nikolay V. Dokholyan, Feng Ding, Caleb Geniesse, Yangwei Jiang, Astha Joshi, Andrey Krokhotin, Marcin Magnus, Olivier Mailhot, Francois Major, Thomas H. Mann, Paweł Piątkowski, Radoslaw Pluta, Mariusz Popenda, Joanna Sarzynska, Lizhen Sun, Marta Szachniuk, Siqi Tian, Jian Wang, Jun Wang, Andrew M. Watkins, Jakub Wiedemann, Yi Xiao, Xiaojun Xu, Joseph D. Yesselman, Dong Zhang, Yi Zhang, Zhenzhen Zhang, Chenhan Zhao, Peinan Zhao, Yuanzhe Zhou, Tomasz Zok, Adriana Żyła, Aiming Ren, Robert T. Batey, Barbara L. Golden, Lin Huang, David M. Lilley, Yijin Liu, Dinshaw J. Patel, and Eric Westhof. Rna-puzzles round iv: 3d structure predictions of four ribozymes and two aptamers. *RNA*, 26: 982–995, 8 2020. ISSN 1355-8382. doi: 10.1261/rna.075341.120.
40. Dong Zhang, Jun Li, and Shi-Jie Chen. IsRNA1: De Novo Prediction and Blind Screening of RNA 3D Structures. *Journal of Chemical Theory and Computation*, 17:1842–1857, 3 2021.
41. Jun Li and Shi-Jie Chen. RNAJP: enhanced RNA 3D structure predictions with non-canonical interactions and global topology sampling. *Nucleic Acids Research*, 51:3341–3356, 4 2023.
42. Jiajie Li, Shanshan Zhang, Dan Zhang, et al. Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics*, 38(16):4042–4043, 2022.
43. Mariusz Popenda, Marta Szachniuk, Maciej Antczak, et al. Automated 3D structure composition for large RNAs. *Nucleic Acids Research*, 40:e112–e112, 8 2012.
44. Marc Parisien and François Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–55, 3 2008.
45. Jun Wang, Jian Wang, Yanzhao Huang, et al. 3dRNA v2.0: An Updated Web Server for RNA 3D Structure Prediction. *International Journal of Molecular Sciences*, 20:4116, 8 2019.
46. Tao Shen, Zhihang Hu, Zhangzhi Peng, et al. E2Efold-3D: End-to-End Deep Learning Method for Accurate de Novo RNA 3D Structure Prediction. *arXiv preprint arXiv:2207.01586*, 2022.
47. Andrew Martin Watkins, Ramya Rangan, and Rhiju Das. FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure*, 28:963–976.e6, 8 2020.
48. Michal J. Boniecki, Grzegorz Lach, Wayne K. Dawson, et al. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research*, 44:e63–e63, 4 2016.
49. Marcin Magnus, Maciej Antczak, Tomasz Zok, et al. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Research*, 48(2):576–588, 12 2019.
50. Clement Bernard, Guillaume Postic, Sahar Ghannay, and Fariza Tahi. RNAdvisor: a comprehensive benchmarking tool for the measure and prediction of RNA structural model quality. *Briefings in Bioinformatics*, 25:bbae064, March 2024.