# RNAdvisor: a comprehensive benchmarking tool for the measure and prediction of RNA structural model quality

**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

**RNA is a complex macromolecule that plays central roles in the cell. While it is well-known that its structure is directly related to its functions, understanding and predicting RNA structures is challenging. Assessing the real or predictive quality of a structure is also at stake with the complex 3D possible conformations of RNAs. Metrics have been developed to measure model quality while scoring functions aim at assigning quality to guide the discrimination of structures without a known and solved reference. Throughout the years, many metrics and scoring functions have been developed, and no unique assessment is used nowadays. Each developed assessment method has its specificity and might be complementary to understanding structure quality. Therefore, to evaluate RNA 3D structure predictions, it would be important to calculate different metrics and/or scoring functions. For this purpose, we developed RNAdvisor, a comprehensive automated software that integrates and enhances the accessibility of existing metrics and scoring functions. In this paper, we present our RNAdvisor tool, as well as state-of-the-art existing metrics, scoring functions and a set of benchmarks we conducted for evaluating them. Source code is freely available on the EvryRNA platform: https://evryrna.ibisc.univ-evry.fr.**

**RNA 3D structure | metric | scoring function | structure evaluation**

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

## Introduction

The various types of non-coding RNA molecules exert their biological functions either by base-pairing mechanisms or through their three-dimensional structure. As for proteins, experiments for determining the spatial conformation of RNA chains are costly, which has led to the development of computational methods for predicting the biologically active ("native") fold from the sole ribonucleic acid sequence (1–5). However, these methods dedicated to RNA have yet to reach the same accuracy as their protein-specific counterparts, such as AlphaFold2 (6) or ESMFold (7). The fact that the native structures available in the Protein Data Bank (PDB) are far less numerous and diverse for RNA than for protein molecules mainly explains this slower rate of progress in solving the RNA folding problem. Nevertheless, current efforts are put forth to overcome this lack of training data.

The quality of a structural model is defined by its "nativity" or native-like character, i.e. how close it is to the native fold of the same RNA sequence. Therefore, evaluating the performance of a predictive method requires measuring the similarity between the 3D models it generates and the corresponding native RNA structure. For this model quality measure, multiple metrics have been proposed throughout the years. Some were directly transposed from the study of protein structures, such as the root-mean-square deviation of atomic positions (RMSD) or the template modeling score (TM-score) (8). Others, such as interaction network fidelity (INF) (9) or mean of circular quantities (MCQ) (10), have been created to take into account the specificities of RNA 3D structures, in particular their greater flexibility.

For predicting RNA fold from the sequence, algorithms explore the conformational space through different strategies (11–15). This produces a certain number of predicted RNA structures that must be ranked. In a real-case scenario where the native structure is not known, such a ranking requires computing relative quality predictions for all the generated models. For this purpose, different scoring functions, also called model quality assessment programs (MQAP), have been released (16–19). Evaluation of these scoring functions is usually done with near-native structures called decoys, which are disturbed native structures that play the roles of predicted RNA structures for predictive models.

An ideal score for predicting model quality would correlate with the Gibbs free energy change ($\Delta G$) of the RNA folding process, as the native structure is the one with the most negative $\Delta G_{\text{folding}}$. However, calorimetric data are not available for the unfolded states, so the thermodynamic relevance of the MQAP scores cannot be evaluated directly. The predicted quality scores are diverse and computed through different approaches, which raises the question of the equivalence between these metrics for representing the ground truth, i.e. the nativity of the model. In case where they actually represent different aspects of the nativity, a subsequent question regards the dependence of the MQAP's accuracy on these different model-to-native similarity measures. To facilitate the calculation of different metrics and scoring functions for a better evaluation of RNA 3D structure predictions, we developed a computational tool called RNAdvisor. RNAdvisor is an open-source tool that integrates all available codes of state-of-the-art metrics and scoring functions.

In this paper, we bring a comprehensive interpretation

of quality measurement and model quality assessment of RNA 3D structures. We describe our RNAdvisor tool, before presenting the benchmarks we conducted thanks to RNAdvisor. Different benchmarks were carried out, considering three datasets available online. We evaluated the scoring functions, and measured the relationship between the different metrics, and between the scoring functions and metrics. We also measured the running time, as well as the $CO_2$ equivalent consumption. All these benchmarks are reproducible, open-source, and accessible at https://github.com/EvryRNA/rnadvisor_results.

## State-of-the-art metrics

To assess whether a predicted RNA tertiary structure is close to its native fold, multiple metrics have been developed. Quality measurement can be general, telling how well the prediction falls into the global conformation. Other metrics, inspired by protein metrics, consider the alignment of structures to evaluate a predicted structure. Nevertheless, proteins and RNAs have differences that limit the adaptation of protein metrics to RNAs. One of the significant differences lies in folding stabilization, where RNA structure is maintained by base pairing and base stacking, while hydrogen interactions in the skeleton support protein structure. Therefore, metrics have been developed to fit the RNA specificities, considering the different types of interactions.

A summary of the state-of-the-art metrics is provided in Table 1.

### A. General metrics

The general metrics give an overall idea of the quality of a prediction. They are usually based on an overall distance averaged throughout the structure. The most used metric is the RMSD, which gives an overall predicted model evaluation. An improvement of this metric was proposed with $\epsilon$RMSD (19), which incorporates RNA features. On the other hand, the CLASH score (20) assesses the overlaps of atoms and doesn't consider the atom deviations compared to the previously mentioned metrics. The main advantage of general metrics is the quick overview of the nativity of the structure. It gives a unique value, an averaged similarity score over the reference. The RMSD is almost always used as a criterion to assess the quality of a computational approach in a database. Nonetheless, it is limited in explaining the limits of a prediction. A high dissimilarity in a small region would highly bias the RMSD value. The CLASH score is more used as an assessment of possible conformation. An almost native structure would have a very low CLASH score, while a low CLASH score structure doesn't necessarily mean a native structure. Finally, the $\epsilon$RMSD tries to add relative base arrangement to the atomic distance deviation to incorporate RNA structural features.

### B. Protein-inspired metrics

Although proteins and RNAs are different molecules, conformational folding shares few characteristics. A higher proportion of solved protein structures makes developing approaches easier. Consequently, protein metrics have been studied and widely used, especially in the CASP competition. One of the known metrics is the TM-score (8), which adds distance normalization to a classic RMSD. Given aligned structures, the GDT-TS (21) computes superimpositions with different distance cutoffs. Another approach, with CAD-score (22), is using a contact-area function to assess differences. The lDDT (23) score was created to quantify the model quality on the level of the residue's environment, where local atomic interactions are considered to obtain a robust metric. The conception of those metrics is not restricted to proteins and can be adapted to RNA sequences. The proteins-based metrics adapted to RNA molecules can give a general overview of predicted structures. While the TM-score avoids the increase of deviation score if the sequence increases, it is still limited to a general assessment. CAD-score and GDT-TS try to incorporate local superimposition, but it would still suffer from the lack of local information.

### C. RNA-oriented metrics

RNAs are unique molecules with a tertiary conformation maintained by base pairing and base stacking. The torsion angles that describe each nucleotide, such as the approximated pseudo-torsion, can be used to best assess the nativity of a structure compared to a solved structure. RNA also has well-defined pairing patterns, where a base interacts with each other. These interactions are very specific, and general metrics or scores inspired by proteins can not integrate them. That is why multiple metrics like INF (9) (for base pairing patterns) or MCQ (10) (for torsion angles) have been developed to allow the integration of RNA structural specificities. The INF score can be specific to base-pairing interactions ($\text{INF}_{bp}$), the base-stacking interactions ($\text{INF}_{stack}$), or consider both ($\text{INF}_{all}$). To include both RMSD and INF advantages, the deformation index (DI) (9) has been developed as the quotient of RMSD by INF. Another metric is the P-VALUE (24), which assesses the validity of a prediction: it describes if a prediction is better than a random prediction. Metrics specific to RNA have the advantage of considering specificities that are major parts of RNA 3D structure stabilization. The metrics have a more concrete meaning and could help the comprehension of a failing prediction. For instance, a bad $INF_{bp}$ (value near 0) value would mean a failing in base-pairing interactions, whereas a bad RMSD (high value) does not provide this information (and could also be biased by a local misprediction). RNA-oriented metrics remain complementary: INF and MCQ scores describe different structure characteristics. Those RNA-oriented metrics can be added to general and protein-based evaluation metrics for a near-complete assessment of predicted structures.

No unique metric can assess structure quality. Each metric has a different particularity that can complement other metrics. While the RMSD and INF are widely used in the community, their efficiency remains limited with real-world

| Metric | Granularity |
|---|---|
| **General metrics** | |
| RMSD | Atom deviation |
| CLASH (20) | Steric clashes |
| $\epsilon$RMSD (19) | Relative distance and orientation |
| **Protein-inspired** | |
| TM-score (8) | Normalised atom deviation |
| GDT-TS (21) | Count of superimposed residues |
| CAD-score (22) | Contact-area |
| lDDT (23) | Interatomic distance differences |
| **RNA-oriented** | |
| P-VALUE (24) | Non-randomness assessment |
| INF, DI (9) | Key interactions accuracy |
| MCQ (10) | Angles dissimilarity |

**Table 1.** Summary of the state-of-the-art metrics used for assessing the quality of predicted RNA 3D structures compared to a reference.

RNAs.

A complete description of the different metrics is provided in Supplementary Materials. It also details the different implementation languages, as well as the source codes.

## State-of-the-art scoring functions

The nativity of RNA molecules can be computed by dissimilarity metrics but requires having a known solved reference structure. This requirement is challenging as the number of solved structures is low. Furthermore, computational methods usually predict multiple conformations that need to be ranked. The relative quality prediction can not rely on a known solve structure. The adaptation of the free energy of the structure has become a standard in the ranking, filtering and confidence assessment of structures. These predictive quality measurements are knowledge-based approaches that rely on statistical potentials. With the recent success of AlphaFold2 (6), new approaches employ deep learning methods for quality predictions of RNA structures.

A summary of the different state-of-the-art scoring functions is provided in Table 2.

### D. Knowledge-based scoring functions

Prediction-based methods like NAST (11), HiRE-RNA (4) or SimRNA (12) use an adaptation of the free energy in their discriminative phase. A common approach uses knowledge-based statistical potentials considering structures to create a quality measurement score. It has been proven to work well for proteins, such as the one used in AlphaFold 2 (6). These potentials are said to be derived from Boltzmann formulations. They rely on a comparison with non-native base pair interactions, known as a reference state. The reference state should ideally come from a set of non-redundant decoy conformations where no interactions between atoms appear. Unfortunately, no ideal dataset exists (25), but approximations of reference states have been proposed through the years (26–31). Adaptation to RNA has been studied (32) and remains limited by the lack of a large and representative RNA dataset. Most knowledge-based approaches to assess RNA structure nativity employ an all-atom distance potential and use averaging reference states, like 3dRNAScore (33) or RASP (16). The challenge is to find good structural features that consider RNA conformational specificities to distinguish native and non-native folding. Methods like $\epsilon$SCORE (19) or DFIRE-RNA (18) consider relative orientation to incorporate RNA flexibility. Short and long-range interactions are considered differently with different reference states in the new potential rsRNASP (17). The main limitation of knowledge-based scoring functions is the lack of a dataset of reference state decoys.

### E. Deep learning scoring functions

With the recent success of AlphaFold2 (6) and its deep architecture, MQAP scores have been developed like RNA3DCNN (34) or ARES (35). They input different characteristics like chemical type or atom position. They use available native conformations to learn a score without explicitly using a reference state. The objective is an RMSD-like metric, meaning that the network learns atom deviation properties to assess structure predictive quality. The architecture is based on a neural network with either convolutional layers or graph neural networks. They rely on decoy datasets generated by either FARFAR 2 (2) for ARES, or relaxed structures by molecular dynamics from PDB for RNA3DCNN. ARES and RNA3DCNN scoring functions remain limited by the current deep learning drawbacks: the lack of interpretability and the need for large datasets. As the number of solved RNA 3D structures is low, deep learning approaches could easily lack generalization to new unseen structures. Datasets considered are biased by either the chosen model creating decoys or the method to relax structures.

No ideal scoring function exists, and the available scores can also be complementary: one score can weigh more dihedral angles, whereas the other could consider chemical types. As no ideal metric exists, some scoring functions could be more linked to a given metric, making the ranking more difficult.

A complete description of the different scoring functions is available in Supplementary Materials, as well as the different implementation languages and the source codes.

## RNAdvisor tool

As the number of available RNA 3D structures increases, assessing the nativity of predicted structures becomes crucial. Numerous families still have unsolved structures in the PDB, but they might be available in the following years. Assessing and understanding the limits of predicted methods for the available and nearly available RNA 3D structures is essential. As discussed previously, no perfect metrics can discriminate between native-like and wrong-predicted structures. The same goes for the scoring functions. Each metric or scoring function has its specificity and could complement the understanding of RNA conformation. Nonetheless, metrics and

| Score | Granularity |
|---|---|
| **Knowledge-based** | |
| RASP (16) | Pairwise-distances |
| $\epsilon$SCORE (19) | Relative positions |
| 3dRNAScore (33) | Distance and dihedral angles |
| DFIRE-RNA (18) | Pairwise-distances |
| rsRNASP (17) | Short and long pairwise distances |
| **Deep learning** | |
| RNA3DCNN (34) | Atom grid |
| ARES (35) | Atom position and chemical type |

**Table 2.** Summary of the scoring functions used for assessing the nativity confidence of RNA 3D structures.

scoring functions have been developed for years by different researchers in different programming languages, making their use difficult. Some web servers, like RNA-tools (36), can compute RMSD or INF scores. It was introduced after RNAPuzzles (37), a collective challenge to evaluate predicted 3D RNA structures. Web servers might be helpful for discrete tests but can not be used to automate the evaluation process. As the number of RNAs is growing, we can not rely on web servers to check each of the predicted structures. Automating the computation of scoring functions is even more crucial as they are widely used for sampling procedures.

We developed a tool called RNAdvisor, that enables the computation of all the available state-of-the-art metrics and scoring functions in one command line. It integrates eleven metrics and four existing scoring functions from nine standalone codes, as shown in Figure 1. We omitted 3dRNAscore because we could not get the source code. We failed to run the ARES code, and RNA3DCNN had bad results compared to the published ones, so we decided not to include it.

Our tool uses coding best practices like DevOps library, named Docker (38) to emancipate the dependency of OS. All the installation needed by each library is already done and easily accessible.

## Benchmark

To evaluate the performance of a scoring function, a common practice (16, 17, 34, 35) is to compare the rank obtained by the native structure in a set of decoys. In this section, we first describe the three datasets of decoys used for the experimentation, followed by a study of the link between existing metrics. Then, we examine the performance of scoring functions, followed by a study of their correlation with metrics. We finally provide a benchmark of computation time and $CO_2$ emissions for scoring function and metrics.

### F. Datasets

We used three datasets, named Test Set I, Test Set II and Test Set III, to assess the relations between scoring functions and metrics. The first two datasets have decoys generated by two distinct strategies widely used to compare scoring functions (16, 17, 34, 35). The last dataset is a real-case scenario where 3D structures from different model predictions should be ranked by nativity.

**Test Set I**[1] is composed of 85 RNAs with decoys generated by MODELLER (39), a predictive model that is used to create decoys with different set of parameters. It uses Gaussian restraints for atom distances and dihedral angles, leading to 500 decoy structures for each RNA. The decoys are close to the native structures as only minor changes are made in the decoy creation.

**Test Set II**[2] corresponds to the prediction-models (PM) subset from rsRNASP (17). It consists of 20 non-redundant single-stranded RNAs with decoy structures generated by four RNA 3D models (10 per model): FARFAR 2 (2), RNAComposer (40), SimRNA (12) and 3dRNAv2.0 (41). It leads to 20 RNAs with 40 decoy structures for each native RNA. The created decoys are less close to the native structure as they use predicted models to create the decoys.

**Test Set III**[3] is the RNA-Puzzles_standardized dataset. It comes from the competition that reproduces the protein CASP challenge for RNA: RNA-Puzzles (37). It contains 21 RNAs and dozens of decoy structures for each RNA. It is commonly used as the most realistic test set to assess the generalization properties of models. The decoys are not all close to the native structure.

### G. Evaluation metrics

Identifying native structures from non-native or near-native is a property required by scoring functions. To assess the quality of a given scoring function, we used the Pearson correlation coefficient (PCC) and the enrichment score (ES). The PCC is computed between the ranked structures based on scoring functions and structures ranked by metrics and 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) 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 below 1 means a bad score and a value equal to 1 means a random prediction.

[1] http://melolab.org/supmat/RNApot/Sup._Data.html.
[2] https://github.com/Tan-group/rsRNASP.
[3] https://github.com/RNA-Puzzles/standardized_dataset/tree/master

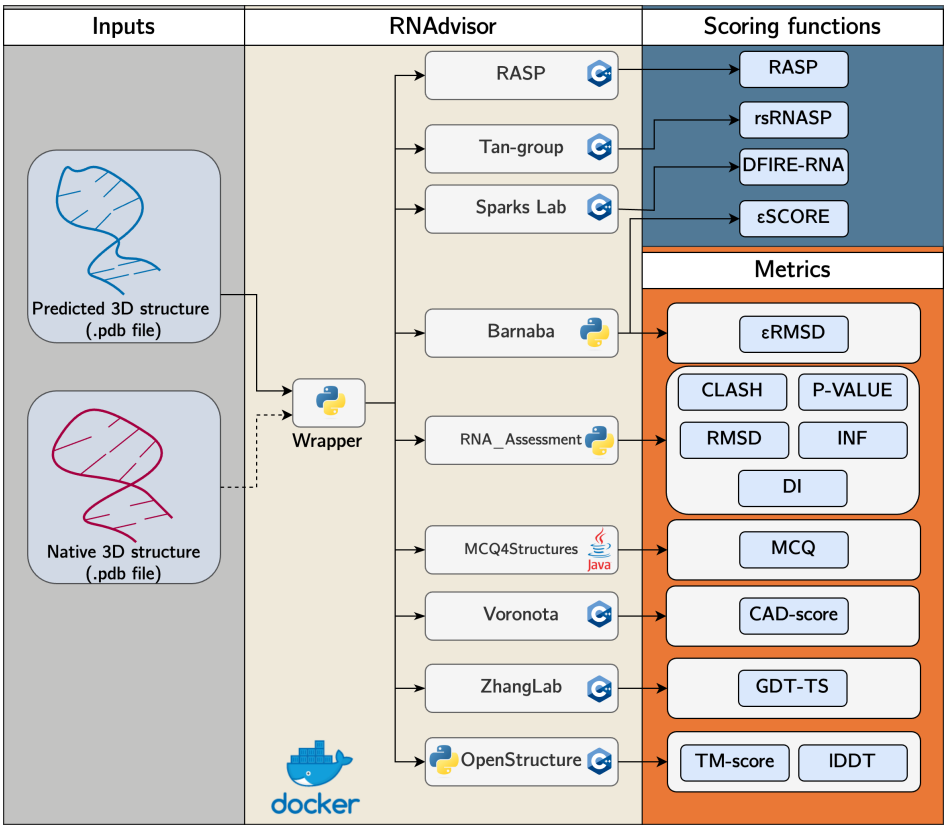


**Figure 1.** Schema of RNAdvisor tool: a wrapper code that gathers open-access libraries for assessment of RNA 3D structures in one interface. It is wrapped in a Docker image to emancipate laborious installation processes. A user can input an RNA 3D structure with the reference structure, and it will compute each metric and scoring function using the different integrated software.

## H. Results

***H.1. Metrics relationship.*** Each metric has its specificity and gives a result based on different assumptions: base interactions, angle conservation, atom distance deviation, etc. Metrics may be redundant with one other. We compared the PCC and ES between each computed metric averaged over the three datasets used. The results are shown in Figure 2. Details for each dataset are also provided in Supplementary file.

RMSD has a high correlation with DI, which is not relevant as DI is composed of both RMSD and INF metrics. RMSD correlates with $\epsilon$RMSD in terms of ES and PCC (6.43 and 0.75, respectively) while being related TM-score (ES of 6.35 and PCC of 0.78). As $\epsilon$RMSD tries to improve the classic RMSD and TM-score adds a normalisation, the correlation makes sense. INF metric highly correlates with ES and PCC with CAD-score and $\epsilon$RMSD (ES of 6.16 and 6.32 and PCC of 0.85 and 0.9, respectively). DI is also linked to $\epsilon$RMSD with an ES of 6.58 and PCC of 0.78. As the $\epsilon$RMSD is an improved RMSD that includes RNA structure specificities, it makes sense that it is correlated to the DI metric as it includes both RMSD and RNA-specific INF metrics. MCQ is the only metric systematically less related to the other metrics. The angle consideration is not mainly included in other metrics computation, which could explain this behaviour. Nonetheless, MCQ has a high correlation for Test Set I with the other metrics (shown in the Supplementary file). It means that for near-native decoys, MCQ behaves like most other metrics, whereas with real-world prediction structures, it is uncorrelated to others. Near-native decoys might keep structural conformations and thus angle conservation, which is not true for the structures from Test Set I and Test Set II. TM-score is connected with another protein-based GDT-TS metric with an ES of 7.14 and PCC of 0.87. Finally, lDDT metric is linked to TM-score (ES of 6.63 and PCC of 0.74), CAD-score (ES of 6.44 and PCC of 0.75) and $\text{INF}_{all}$ (ES of 6.14 and PCC of 0.76). As the lDDT metric incorporates interatomic distance information, this is retrieved in the CAD-score and in the normalised atom deviation of the TM-score.

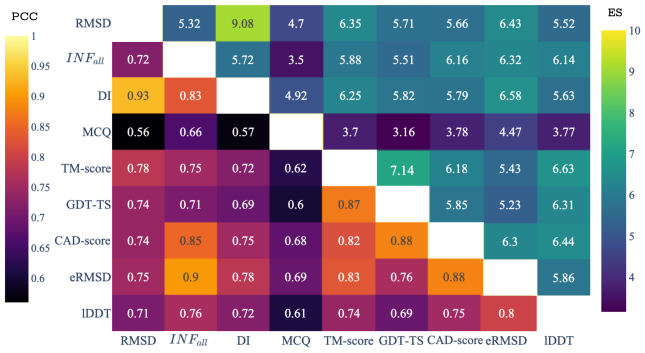


**Figure 2.** ES and PCC scores for each metric averaged over the three test datasets. The lower half of the matrix represents the PCC, while the upper half corresponds to the ES score. The diagonal has a PCC of 1 and ES of 10.

We can conclude that the MCQ metric is highly uncorre-

| | Scoring functions | | | |
|---|---|---|---|---|
| Dataset | RASP | $\epsilon$SCORE | DFIRE-RNA | rsRNASP |
| Test Set I | 83/85 | **85**/85 | 83/85 | 77/85 |
| Test Set II | 2/20 | 9/20 | 10/20 | **16**/20 |
| Test Set III | 2/21 | 5/21 | 10/21 | **18**/21 |
| **Total** | 86/126 | 99/126 | 103/126 | **111**/126 |

**Table 3.** Number of native structures found with the lowest score for each dataset. It corresponds to the number of times the native structure has the lowest scoring function value among the decoys.

lated to the others (for Test Set II and III), while the TM-score and GDT-TS seem very dependent. INF$_{all}$ discriminate decoys with the same behavior as the CAD-score. $\epsilon$RMSD is linked to INF and thus CAD-score (as they are correlated) and DI. Their correlations are not perfect (no ES of 10 or PCC of 1), meaning that every metric can help assess predicted model quality.

***H.2. Scoring function ranking.*** The aptitude of a scoring function to classify native and near-native structures is essential for developing models. Table 3 shows the number of native structures with the lowest scoring function value among the decoys. RASP performs well for near-native structures (Test Set I) by finding 83 out of 85, but fails for the other datasets. $\epsilon$SCORE and DFIRE-RNA have almost identical results, with 99 and 103 found structures out of 126 overall. rsRNASP does not perform as well as the other scoring functions for Test Set I but outperforms them for the two other datasets. It leads the overall native structures found with 111 out of 126. Details of the average rank of the native structure for each dataset are provided in Supplementary file.

rsRNASP seems the best scoring function for ranking and finding the native structure, followed by DFIRE-RNA and $\epsilon$SCORE. rsRNASP is less accurate for close decoys (represented by Test Set I), where RASP discriminates better in this case. Incorporating statistical potentials that weigh differently short, mid-range and long interactions like rsRNASP may not be the best choice for very close decoys.

Results are induced on the four scoring functions that we succeeded in implementing. We can not conclude on 3dRNAscore, RNA3DCNN and ARES performances for ranking native-like structures.

***H.3. Scoring functions and metrics relationship.*** We computed the ES and PCC scores for each data set, each available scoring function and metric. We considered for the metrics the RMSD, INF (also named INF$_{all}$, as we considered the averaged value over base-pairing and base-stacking interactions), DI, MCQ, TM-score, GDT-TS, CAD-score, lDDT and $\epsilon$RMSD. We did not include P-VALUE or CLASH score, as P-VALUE is like a condition-metric to assess the non-randomness of the prediction. The CLASH score computation failed for most RNA molecules, leading to non-reliable results for this metric.

The results for the different datasets are given in Supplementary file. A summary of the most correlated metrics for each scoring function is provided in Table 4, where each best-related metric is counted for all three datasets. It shows that RASP has a high correlation with TM-score and MCQ in terms of ES, and MCQ, RMSD, lDDT and TM-score in terms of PCC. It means that RASP integrates atom deviation well in its statistical potential (as it is related to TM-score, RMSD, lDDT) and favours structures with good angle conservation (linked to MCQ). $\epsilon$SCORE is linked to CAD-score, lDDT and TM-score in terms of ES, and INF$_{all}$, MCQ and CAD-score. The high link with CAD-score in both evaluation criteria means it tends to conserve the RNA interactions (INF) and thus maintains a low contact-area difference (CAD-score). DFIRE-RNA is tied to CAD-score, lDDT and TM-score in terms of ES and MCQ, DI and lDDT for the PCC criteria. It seems to have some RNA interaction properties (INF/DI) while maintaining angle conservation (MCQ) and interatomic distance conservation (lDDT). Finally, rsRNASP is correlated to lDDT, TM-score in terms of ES, and INF$_{all}$, $\epsilon$RMSD, GDT-TS and CAD-score in terms of PCC. As rsRNASP considers low and high-range interactions, it tends to favour structures with good sequence alignment (lDDT, TM-score, $\epsilon$RMSD, GDT-TS, CAD-score) and RNA structural features (INF).

An example of RNA 1ec6D from Test Set I and its decoys is shown in Figure 3. We normalized all the scoring functions and computed the logarithm to plot them on the same scale. Growing scores were reversed to follow the same pattern as the others. DFIRE-RNA has a low slope compared to rsRNASP, RASP and $\epsilon$SCORE. On the other hand, $\epsilon$SCORE has a high slope and tends to increase the gap between near-native decoys. The overall high slope of $\epsilon$SCORE for each metric shows a good discrimination property to divide native from non-native structures. It is supported by the number of native structures founded with the lowest $\epsilon$SCORE for Test Set I: 85 out of 85.

***H.4. Computation time and $CO_2$ emissions.*** Computing a structure's energy is integrated into developing models for predicting RNA 3D structures. Models for predicting 3D structures are usually slow and even slower when the sequence length increases. The computation time of energies shouldn't be a bottleneck for selecting created models' decoys.

We tracked the inference computation time for each energy for RNA of different lengths. We took as a benchmark the chain A with 2878 nucleotides of RNA 3f1hA from Test Set I. We randomly created five substructures for each step of 100 nucleotides from 100 to 2800. We tracked and averaged the time required to compute the scoring functions and metrics. It leads to Figure 4. It highlights the low computation time of DFIRE and $\epsilon$SCORE that doesn't exceed 20 seconds for RNA with a sequence length of less than 2800 nucleotides. The same goes for metrics like $\epsilon$RMSD, GDT-TS, MCQ, DI, RMSD, and INF with a low computation time. On the other hand, RASP takes around 6min48 to compute for a sequence of 2800 nucleotides. rsRNASP has a complexity that almost explodes with the sequence length (10min39 for a sequence of 2800). This computation time is not scalable for the development of high-resolution models. For instance,

| | Test Set I | | Test Set II | | Test Set III | |
|---|---|---|---|---|---|---|
| | ES | PCC | ES | PCC | ES | PCC |
| RASP | TM-score | RMSD/TM-score/lDDT | TM-score | MCQ | MCQ | MCQ |
| $\epsilon$SCORE | TM-score | MCQ | CAD-score | CAD-score | lDDT | $INF_{all}$/CAD-score |
| DFIRE-RNA | TM-score | lDDT | CAD-score | DI | lDDT | MCQ |
| rsRNASP | TM-score | $\epsilon$RMSD | lDDT | $INF_{all}$ | lDDT | CAD-score |

**Table 4.** Summary of the best-correlated metrics for ES and PCC scores for each scoring function for the three test sets.

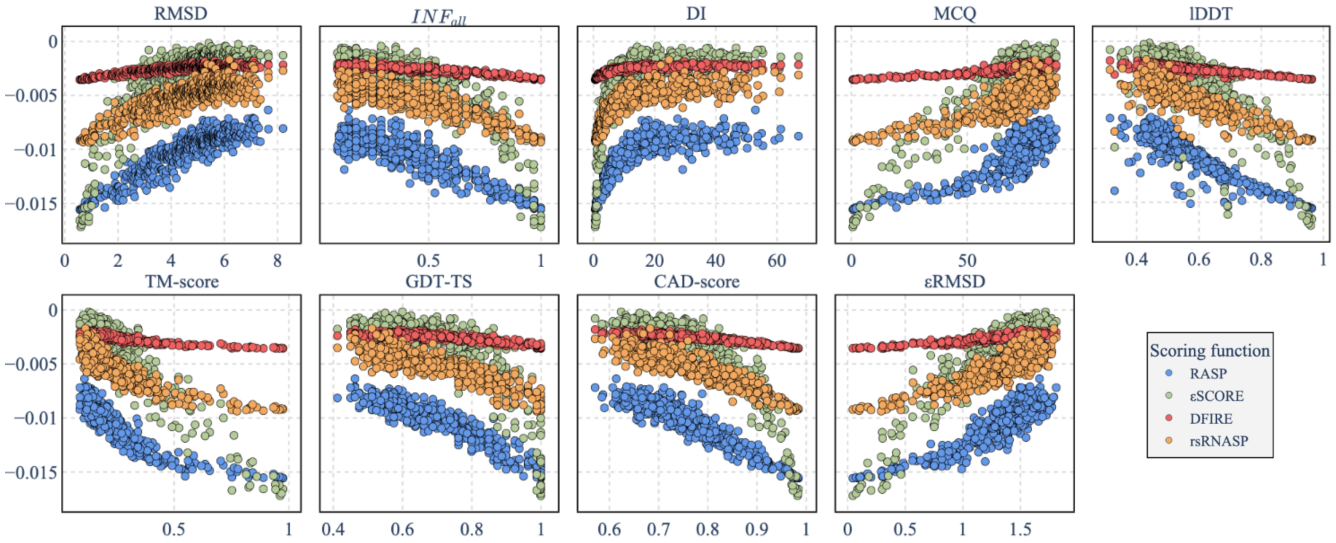


**Figure 3.** Logarithm of the normalized four scoring functions (RASP, $\epsilon$SCORE, DFIRE-RNA and rsRNASP) of RNA 1ec6D and its 500 decoys from Test Set I depending on eight metrics. The increasing scores (like $\epsilon$SCORE) were reversed to follow the same growth pattern as the other scoring functions.

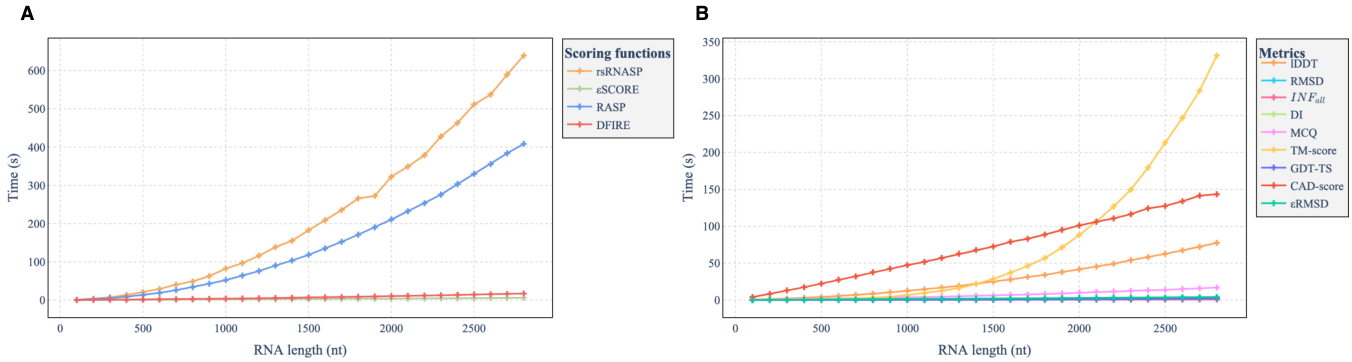


**Figure 4.** Computation time depending on the number of nucleotides in RNA sequences for substructures from RNA 3f1hA (Test Set I). A) Time executions for scoring functions. B) Time executions for metrics.

if a predicted model generates 1000 structures of 2800 nucleotides and then tries to select the best ones with rsRNASP, it will take more than seven days and 9 hours. The MCQ and lDDT have a computation time higher than the other metrics for a sequence of more than a thousand nucleotides. MCQ has a computation time of less than 20 seconds compared to less than 1min17 for RNAs of 2800 nucleotides. CAD-score has a high computation time, with more than 2min20 for RNAs of 2800 nucleotides. Finally, the TM-score has a high computation time of 5min30 for a sequence of 2800 nucleotides.

As computation methods can have environmental impacts, we also included carbon footprints of structural assessment methods. We computed for each dataset the equivalent $CO_2$ measurements for an RNA (averaged over the different decoys), using CodeCarbon tool (42). The results are shown in Figure 5. We observe an overall higher consumption of $CO_2$ for Test Set I, which is explained by the long RNAs in this dataset. rsRNASP and CAD-score have the highest $CO_2$ consumption (with around 0.011 g/$CO_2$ per RNA), followed by RASP and TM-score. CAD-score has a higher $CO_2$ consumption compared to TM-score while calculating quicker. This difference could be explained by better resource management by the TM-score compared to the CAD-score.

## I. Discussion

In these experiments, we found that rsRNASP outperforms the other scoring functions in finding the native structures among the three datasets while being correlated to lDDT, TM-score, INF, $\epsilon$RMSD, GDT-TS and CAD-score. It comes

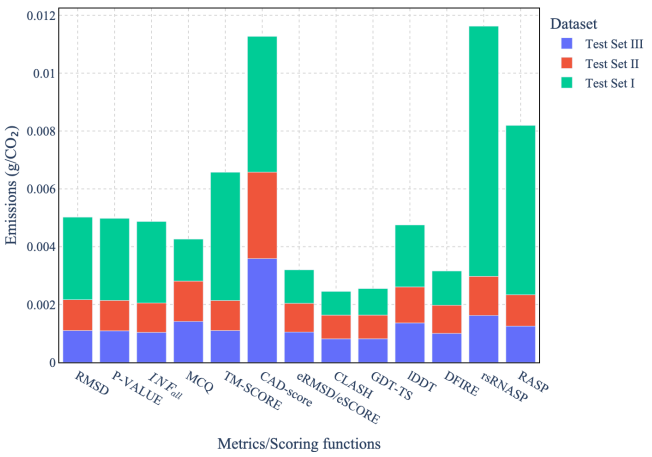


**Figure 5.** The $CO_2$ equivalent measures for each metric and scoring function for each dataset. $CO_2$ emissions

with a price: its computation time highly increases with the number of nucleotides of an RNA sequence. Metrics also correlate between them: INF implies having good interaction accuracy, reducing the contact-area differences and the associated CAD metric. It is also correlated to $\epsilon$RMSD, an improved RMSD considering the differences in the base interaction network. The MCQ metric seems to be the only metric that isn't correlated to the others. Atomic deviation metrics tend also to be correlated: lDDT, TM-score, RMSD and $\epsilon$RMSD.

We advise using MCQ as an evaluation metric to assume RNA nativity with DI (RMSD and INF) and lDDT (or TM-score or GDT-TS, as they are correlated). It provides a complementary set of metrics that assess RNA 3D structure evaluation compared to a reference structure. One should keep in mind the computation time and $CO_2$ consumption that is associated with these metrics. Therefore, we do not recommend the CAD-score or TM-score, which have a high computation time and thus $CO_2$ consumption.

As a scoring function, we suggest using rsRNASP. If the evaluated structures have a long sequence, we suggest using DFIRE-RNA or $\epsilon$SCORE, which has good discriminating properties even if it doesn't outperform rsRNASP. The consumption time and $CO_2$ emissions of rsRNASP prevent using RNA with long sequences.

## Conclusion

In this work, we have presented a general overview of the assessment of the nativity of an RNA 3D structure. One can compare a predicted structure with comparative tools like atom distances, interaction accuracy or angle dissimilarity, given a reference structure. Such metrics can have general assumptions (like RMSD), whereas others tend to target RNA specificities. Protein metrics have also been adapted to be relevant for RNA assumption. Nonetheless, having a known solved structure is a strong condition and impossible when creating a model to predict RNA 3D structure. Instead, statistical potential energies tend to reproduce molecule-free energy: the lowest, the more stable and thus the more native a structure is. We have provided a review of the known RNA scoring functions and an extensive benchmark that is reproducible and open-source.

Each of these metrics and scoring functions results from years of research by different groups of researchers. Each code is written by different authors and is sometimes hard and time-consuming to install locally. We developed a software, named RNAdvisor, that gathers metrics and scoring functions in a unique interface. It provides a documented and wrapped code available in one command line. It helps centralize and automate the computation of metrics and scoring functions to assess RNA 3D structure nativity. RNAdvisor represents an advancement in the automation of RNA 3D structure evaluation. It facilitates the accessibility of existing metrics and scoring functions and thus can help accelerate investigation in RNA 3D structure predictions.

Future works could imply the development of new metrics that consider all the complementary specificities of RNA molecules and current metrics. This development must integrate existing metrics to avoid redundant work. The assessment of RNA nativity with energy score is still an area of research that should be explored. One should integrate the computation time to have a scoring function that could be adapted to long RNAs. Those developments should be guided with easy-to-use code to enable the reproducibility and integration of predicted models.

**ACKNOWLEDGEMENTS**

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

## Bibliography

1. Rhiju Das and David Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America*, 104:14664–9, 10 2007. doi: 10.1073/pnas.0703836104.
2. 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. ISSN 09692126. doi: 10.1016/j.str.2020.05.011.
3. Petr Sulc, Flavio Romano, Thomas Ouldridge, Jonathan Doye, and Ard Louis. A nucleotide-level coarse-grained model of rna. *The Journal of chemical physics*, 140, 03 2014. doi: 10.1063/1.4881424.
4. Tristan Cragnolini, Yoann Laurin, Philippe Derreumaux, and Samuela Pasquali. Coarse-Grained HiRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings. *Journal of Chemical Theory and Computation*, 11(7):3510–3522, 2015. doi: 10.1021/acs.jctc.5b00200. PMID: 26575783.
5. Andrey Krokhotin, Kevin Houlihan, and Nikolay V. Dokholyan. iFoldRNA v2: folding RNA with constraints. *Bioinformatics*, 31:2891–2893, 9 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv221.
6. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 8 2021. ISSN 0028-0836. doi: 10.1038/s41586-021-03819-2.
7. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.
8. Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702—710, December 2004. ISSN 0887-3585. doi: 10.1002/prot.20264.
9. Marc Parisien, José Cruz, Eric Westhof, and François Major. New metrics for comparing

and assessing discrepancies between RNA 3D structures and models. *RNA (New York, N.Y.)*, 15:1875–85, 09 2009. doi: 10.1261/rna.1700409.

10. Tomasz Zok, Mariusz Popenda, and Marta Szachniuk. MCQ4Structures to compute similarity of molecule structures. *Central European Journal of Operations Research*, 22, 04 2013. doi: 10.1007/s10100-013-0296-5.
11. Magdalena A. Jonikas, Randall J. Radmer, Alain Laederach, Rhiju Das, Samuel Pearlman, Daniel Herschlag, and Russ B. Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, 15:189–199, 2 2009. ISSN 1355-8382. doi: 10.1261/rna.1270809.
12. Michal J. Boniecki, Grzegorz Lach, Wayne K. Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M. Rother, and Janusz M. Bujnicki. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research*, 44:e63–e63, 4 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1479.
13. 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. ISSN 1549-9618. doi: 10.1021/acs.jctc.0c01148.
14. Dong Zhang, Shi jie Chen, and Ruhong Zhou. Modeling Noncanonical RNA Base Pairs by a Coarse-Grained IsRNA2 Model. *The journal of physical chemistry. B*, 2021.
15. 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. ISSN 0305-1048. doi: 10.1093/nar/gkad122.
16. Emidio Capriotti, Tomas Norambuena, Marc A. Marti-Renom, and Francisco Melo. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, 27:1086–1093, 4 2011. ISSN 1460-2059. doi: 10.1093/bioinformatics/btr093.
17. Ya-Lan Tan, Xunxun Wang, Ya-Zhou Shi, Wenbing Zhang, and Zhi-Jie Tan. rsRNASP: A residue-separation-based statistical potential for RNA 3D structure evaluation. *Biophysical Journal*, 121:142–156, 1 2022. ISSN 00063495. doi: 10.1016/j.bpj.2021.11.016.
18. Emidio Capriotti, Tomas Norambuena, Marc A. Marti-Renom, and Francisco Melo. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, 27(8):1086–1093, 02 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr093.
19. Sandro Bottaro, Francesco Di Palma, and Giovanni Bussi. The Role of Nucleobase Interactions in RNA Structure and Dynamics. *Nucleic acids research*, 42, 10 2014. doi: 10.1093/nar/gku972.
20. I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, and D. C. Richardson. Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*, 35:W375–W383, 5 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm216.
21. Adam Zemla, Česlovas Venclovas, John Moult, and Krzysztof Fidelis. Processing and analysis of CASP3 protein structure predictions. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):22–29, 1999. doi: https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<22::AID-PROT5>3.0.CO;2-W.
22. Kliment Olechnovič, Eleonora Kulberkytė, and Česlovas Venclovas. CAD-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins*, 81, 01 2013. doi: 10.1002/prot.24172.
23. Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics (Oxford, England)*, 29(21):2722–2728, 2013. doi: 10.1093/bioinformatics/btt473.
24. Christine Hajdin, Feng Ding, Nikolay Dokholyan, and Kevin Weeks. On the significance of an RNA tertiary structure prediction. *RNA (New York, N.Y.)*, 16:1340–9, 07 2010. doi: 10.1261/rna.1837410.
25. Hong Deng, Yanqing Jia, Yangjun Wei, and Yang Zhang. What is the best reference state for designing statistical atomic potentials in protein structure prediction? *Proteins: Structure, Function, and Bioinformatics*, 80:2311–2322, 2012. doi: 10.1002/prot.24121.
26. Ram Samudrala and John Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, 275: 895–916, 1998. doi: 10.1006/jmbi.1997.1479.
27. Hua Lu and Jeffrey Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Structure, Function, and Genetics*, 44:223–232, 2001. doi: 10.1002/prot.1087.
28. Dmitry Rykunov and Andras Fiser. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. *Proteins: Structure, Function, and Bioinformatics*, 67:559–568, 2007. doi: 10.1002/prot.21279.
29. Huan-Xiang Zhou and Yaoqi Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, 11:2714–2726, 2002. doi: 10.1110/ps.0217002.
30. Min-Yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15:2507–2524, 2006. doi: 10.1110/ps.062416606.
31. Jian Zhang and Yang Zhang. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS One*, 5:e15386, 2010. doi: 10.1371/journal.pone.0015386.
32. Ya-Lan Tan, Chen-Jie Feng, Lei Jin, Ya-Zhou Shi, Wenbing Zhang, and Zhi-Jie Tan. What is the best reference state for building statistical potentials in rna 3d structure evaluation? *RNA*, 25:793–812, 7 2019. ISSN 1355-8382. doi: 10.1261/rna.069872.118.
33. Jian Wang, Yunjie Zhao, Chunyan Zhu, and Yi Xiao. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Research*, 43: e63–e63, 5 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv141.
34. Jun Li, Wei Zhu, Jun Wang, Wenfei Li, Sheng Gong, Jian Zhang, and Wei Wang. 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. doi: 10.1371/journal.pcbi.1006514.
35. Raphael J. L. Townshend, Stephan Eismann, Andrew M. Watkins, Ramya Rangan, Maria Karelina, Rhiju Das, and Ron O. Dror. Geometric deep learning of RNA structure. *Science*, 373:1047–1051, 8 2021. ISSN 0036-8075. doi: 10.1126/science.abe5650.
36. Marcin Magnus. rna-tools.online: a swiss army knife for rna 3d structure modeling workflow. *Nucleic Acids Research*, 50(W1):W657–W662, 2022. doi: 10.1093/nar/gkac372.
37. José Cruz, Marc-Frédérick Blanchet, Michal Boniecki, Janusz Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay Dokholyan, Samuel Flores, Lili Huang, Christopher Lavender, Veronique Lisi, François Major, Katarzyna Mikolajczak, Dinshaw Patel, Anna Philips, Tomasz Puton, John Santalucia, and Eric Westhof. RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA (New York, N.Y.)*, 18:610–25, 02 2012. doi: 10.1261/rna.031054.111.
38. Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.
39. Andrej Sali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234:779–815, 1993.
40. Mariusz Popenda, Marta Szachniuk, Maciej Antczak, Katarzyna J. Purzycka, Piotr Lukasiak, Natalia Bartol, Jacek Blazewicz, and Ryszard W. Adamiak. Automated 3D structure composition for large RNAs. *Nucleic Acids Research*, 40:e112–e112, 8 2012. ISSN 1362-4962. doi: 10.1093/nar/gks339.
41. Jun Wang, Jian Wang, Yanzhao Huang, and Yi Xiao. 3dRNA v2.0: An Updated Web Server for RNA 3D Structure Prediction. *International Journal of Molecular Sciences*, 20:4116, 8 2019. ISSN 1422-0067. doi: 10.3390/ijms20174116.
42. Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. In *Climate Change workshop, NeurIPS 2019*, 2019.