

cgRNASP: coarse-grained statistical potentials with residue separation for RNA structure evaluation

Ya-Lan Tan^{1,2,†}, Xunxun Wang^{2,†}, Shixiong Yu², Bengong Zhang^{1,*} and Zhi-Jie Tan^{1,2,*}

¹Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430073, China and ²Department of Physics and Key Laboratory of Artificial Micro & Nano-structures of Education, School of Physics and Technology, Wuhan University, Wuhan 430072, China

Received September 21, 2022; Revised January 21, 2023; Editorial Decision January 27, 2023; Accepted February 03, 2023

ABSTRACT

Knowledge-based statistical potentials are very important for RNA 3-dimensional (3D) structure prediction and evaluation. In recent years, various coarse-grained (CG) and all-atom models have been developed for predicting RNA 3D structures, while there is still lack of reliable CG statistical potentials not only for CG structure evaluation but also for all-atom structure evaluation at high efficiency. In this work, we have developed a series of residue-separation-based CG statistical potentials at different CG levels for RNA 3D structure evaluation, namely cgRNASP, which is composed of long-ranged and short-ranged interactions by residue separation. Compared with the newly developed all-atom rsRNASP, the short-ranged interaction in cgRNASP was involved more subtly and completely. Our examinations show that, the performance of cgRNASP varies with CG levels and compared with rsRNASP, cgRNASP has similarly good performance for extensive types of test datasets and can have slightly better performance for the realistic dataset—RNA-Puzzles dataset. Furthermore, cgRNASP is strikingly more efficient than all-atom statistical potentials/scoring functions, and can be apparently superior to other all-atom statistical potentials and scoring functions trained from neural networks for the RNA-Puzzles dataset. cgRNASP is available at <https://github.com/Tan-group/cgRNASP>.

INTRODUCTION

RNAs have critical biological functions such as gene regulations and catalysis (1,2), and their functions are generally coupled to their structures (3,4). Consequently, the knowledge of RNA structures, especially 3-dimensional

(3D) structures, is crucial for understanding RNA biological functions (5,6). Due to the huge cost of experimental measurements, high-resolution 3D structures of RNAs deposited in Protein Data Bank (PDB) database are still limited (7). Complementary to experiments, various computational models have been developed to predict RNA 3D structures *in silico* (8,9), and these models can be roughly classified into fragment-assembly-based ones and physics-based ones (10–46). A computational model for 3D structure prediction generally requires a reliable energy function for evaluating/assessing predicted structure candidates (47,48). Therefore, a reliable energy function or a reliable structure evaluation is very important for a computational model of RNA 3D structure prediction.

For proteins, knowledge-based statistical potentials derived from experimental structures deposited in the PDB database have been shown to be rather efficient and effective for the quality evaluation/assessment of protein 3D structures, protein-protein, and protein-ligand docking (49–69). For RNA 3D structure evaluation, several statistical potentials have been developed based on different modeled reference states (70–74), since the ideal reference state containing all possible non-redundant decoy conformations without inter-atom interactions in phase space is generally inaccessible in practical modeling (65,74). Bernauer *et al.* have derived differentiable statistical potentials (KB) at both all-atom and coarse-grained levels based on the quasi-chemical approximation reference state (70). Capriotti *et al.* have built all-atom and coarse-grained statistical potentials (RASP) based on the averaging reference state (71). Wang *et al.* have derived a combined distance- and torsion angle-dependent all-atom statistical potential (3dRNAscore) based on the averaging reference state (72). Zhang *et al.* have proposed an all-atom statistical potential based on the finite-ideal-gas reference state (DFIRE-RNA) (73). Recently, we have made a comprehensive survey on the six existing reference states widely used for proteins through building six statistical potentials based on the same training dataset, and found that the finite-

*To whom correspondence should be addressed. Tel: +86 027 687524811; Email: zjtian@whu.edu.cn

Correspondence may also be addressed to Bengong Zhang. Email: bgzhang@wtu.edu.cn

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Table 1. Heavy atoms for CG beads used in cgRNASP and in existing CG-based RNA 3D structure models

Heavy atoms for CG beads	No. of CG beads	CG-based structure prediction models
P, C4', N1 (or N9) ^a	3	SimRNA [43], Vfold [23,25], Shapiro's model [80]; A CG model with salt effect [33,37]
P, C4'	2	SimRNA [43], Vfold [23,25], Shapiro's model [80]; HiRE-RNA [40,41], IsRNA [26,27], RACER [45,46], a <i>k</i> -state virtual bond model [14], A CG model with salt effect [33,37]
C4'	1	SimRNA [43], Vfold [23,25], Shapiro's model [80]; HiRE-RNA [40,41], IsRNA [26,27], RACER [45,46], a <i>k</i> -state virtual bond model [14], A CG model with salt effect [33,37]

^aP, C4', and N1 are for pyrimidine, and P, C4' and N9 are for purine.

ideal-gas and random-walk-chain reference states are modestly better than other ones in identifying native structures and ranking decoy structures (74). However, the existing traditional statistical potentials only achieve a poor performance for the realistic test dataset—RNA-Puzzles dataset (75,76). Beyond traditional statistical potentials, machine/deep learning approaches have been employed for RNA 3D structure evaluation (77,78). Despite of ‘black-box’ training/learning process, RNA3DCNN built through 3D convolutional neural networks exhibits remarkably improved performance in identifying native structures for the RNA-Puzzles dataset compared with the traditional statistical potentials (77), and the newly developed scoring function of ARES from deep neural network based on training data from FARFAR2 showed rather good performance for evaluating structures from FARFAR2 (78). Very recently, we have developed an all-atom statistical potential of rsRNASP by distinguishing short- and long-ranged interactions at residue separation level, and for the RNA-Puzzles dataset, rsRNASP has visibly improved performance than existing statistical potentials and scoring functions from neural networks (47,79).

Very importantly, for reducing conformational space and improving computational efficiency, almost all existing physics-based models for RNA 3D structure prediction are based on different-level CG representations rather than the all-atom representation, including SimRNA (43,44), iFold (29,30), NAST (31), IsRNA (26,27), Vfold (23,25), HiRE-RNA (40,41), oxRNA (42), RACER (45,46) and our CG model with salt effect (33,37). Consequently, a reliable CG statistical potential is crucially important for a CG-based 3D structure prediction model rather than an all-atom one. However, existing CG statistical potentials (e.g. CG KB (70) and CG RASP (71)) are still with very low performance, and recently developed statistical potentials/scoring functions with relatively good performance (e.g. rsRNASP (79), RNA3DCNN (77), ARES (78) and DFIRE-RNA (73)) are all based on the all-atom representation. Therefore, until now, reliable CG statistical potentials are still highly required at different CG levels. Such CG statistical potentials can be very useful not only for CG structure evaluation but also for all-atom structure evaluation at high efficiency.

In this work, we have developed a series of distance-dependent CG statistical potentials based on residue separation for RNA 3D structure evaluation at different CG levels, namely cgRNASP, which is composed of short- and long-ranged potentials distinguished by residue separation. In cgRNASP, beyond the newly developed all-atom rsRNASP, the interactions between nearest-neighbor

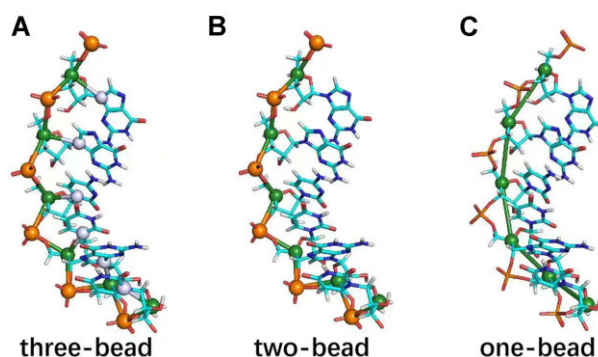


Figure 1. Illustration for different coarse-grained (CG) representations used in developing cgRNASP for RNA 3D structure evaluation. (A) 3 CG beads at heavy atoms of P, C4', and N1 for pyrimidine (or P, C4', and N9 for purine); (B) 2 CG beads at heavy atoms of P and C4' and (C) 1 CG bead at heavy atom of C4'. Please see also Table 1.

residues and between next-nearest ones were explicitly added in the short-ranged interaction. The performance of cgRNASP can have slightly better performance than the all-atom rsRNASP for RNA-Puzzles dataset (a realistic dataset) and strikingly higher evaluation efficiency than all-atom statistical potentials/scoring functions. Moreover, for the RNA-Puzzles dataset, cgRNASP is apparently superior to other all-atom traditional statistical potentials and scoring functions trained from neural networks.

MATERIALS AND METHODS

Coarse-grained representations

First, a survey was made on existing physics-based models for RNA 3D structure prediction to figure out which heavy atoms were used for representing CG atoms in these models, and for convenience, we only considered heavy atoms rather than dummy atoms although some CG-based models involved the (mass) centers of certain atom groups as dummy CG atoms. As shown in Table 1, P and C4' atoms were used most widely to describe RNA backbones, and N9 atom for purine or N1 atom for pyrimidine was used frequently to (partially) describe bases. According to Table 1, we developed our residue-separation-based CG statistical potentials (cgRNASP) at several CG levels: (i) three CG beads at P, C4', and N9 atoms for purine (or N1 atom for pyrimidine); (ii) two CG beads at P and C4' atoms and (iii) one CG bead on C4' atom, which were also illustrated in Figure 1. Correspondingly, for simplicity, our CG statistical potentials are named as cgRNASP, cgRNASP-PC and

cgRNASP-C respectively, and cgRNASP of 3-bead representation is regarded as a representative one of the proposed CG statistical potentials in this work.

Residue-separation-based CG statistical potentials

Since RNA folding is generally hierarchical (81), the different residue-separation-ranged interactions may play different roles in stabilizing RNA 3D structures (49). Here, residue separation is described by $k = |m - n|$, where m and n stand for the observed nucleotide indices of a pair of atoms along an RNA sequence, and a residue separation threshold k_0 is applied to distinguish short- and long-ranged interactions. In analogy to the newly developed all-atom rsRNASP (79), in cgRNASP, the total energy for an RNA conformation C of a given sequence S is composed of short-ranged and long-ranged contributions:

$$\Delta E(S, C) = \Delta E_{\text{short}} + \omega \Delta E_{\text{long}} \quad (1)$$

where ω is a weight to balance the two contributions.

The long-ranged energy ΔE_{long} can be given by (79)

$$\Delta E_{\text{long}} = \sum \Delta E_{\text{long}}(i, j, r) \quad (2)$$

and

$$\Delta E_{\text{long}}(i, j, r) = -k_B T \ln \frac{P_{k \geq k_0}^{\text{obs}}(i, j, r)}{P_{k \geq k_0}^{\text{ref}}(i, j, r)} \quad (3)$$

where $P_{k \geq k_0}^{\text{obs}}(i, j, r)$ and $P_{k \geq k_0}^{\text{ref}}(i, j, r)$ are the probabilities of the atom pairs of types i and j with residue separation $k \geq k_0$ located in the spatial distance interval $[r, r + dr]$ for the native and reference states, respectively. Here, r is the spatial distance between the atom pairs of types i and j , and dr represents the distance bin width.

The short-ranged energy ΔE_{short} in cgRNASP is treated in a more complete and subtle way than the all-atom rsRNASP (79) and is given by

$$\Delta E_{\text{short}} = \sum \Delta E_{\text{short}, k=1}(i, j, r) + \alpha \sum \Delta E_{\text{short}, k=2}(i, j, r) + \beta \sum \Delta E_{\text{short}, 3 \leq k < k_0}(i, j, r) \quad (4)$$

and

$$\Delta E_{\text{short}, k \in \text{range}}(i, j, r) = -k_B T \ln \frac{P_{k \in \text{range}}^{\text{obs}}(i, j, r)}{P_{k \in \text{range}}^{\text{ref}}(i, j, r)} \quad (5)$$

where $k \in \text{range}$ means that residue separation k should be in the corresponding ranges in Eq. (4). $P_{k \in \text{range}}^{\text{obs}}(i, j, r)$ and $P_{k \in \text{range}}^{\text{ref}}(i, j, r)$ are the probabilities of the atom pairs of types i and j with residue separation $k \in \text{range}$ (in Eq. (4)) located in spatial distance interval $[r, r + dr]$ for the native and reference states, respectively. In addition, α and β are the weights to balance the contributions of energy terms in Eq. (4). It is noted that beyond the all-atom rsRNASP, the interactions between nearest-neighbor residues and between next-nearest neighbor ones were explicitly added in the short-ranged interaction in cgRNASP.

In analogy to the all-atom rsRNASP (79), the average reference state was used for short-ranged energy ΔE_{short} , and

for long-ranged energy ΔE_{long} , the finite-ideal-gas reference state (61) was used in cgRNASP instead of the random-walk-chain reference state in rsRNASP since the former one also has relatively better performance and has been widely used in protein structure evaluation (48,74). Please see the Supplementary Materials for the details of using the reference states for deriving the short- and long-ranged potentials, and see Refs (54,61,74) for the detailed description of the existing reference states.

Training set and parameters

In cgRNASP, we used the same non-redundant training native set recently used in deriving the all-atom rsRNASP and the dataset is available at <https://github.com/Tan-group/rsRNASP> (79). The dataset was produced based on the RNA 3D Hub non-redundant set (82) (Release 3.102), which can be found at <http://rna.bgsu.edu/rna3dhub/nrlist>. The training native set contains 191 RNA structures with chain length >10 nt, sequence identity $\leq 80\%$, coverage $\leq 80\%$ and X-ray resolution <3.5 Å, and excludes those RNAs complexed with proteins or DNAs (79). It is noted that some continuously updated databases have recently been developed for the convenience of RNA structure data-driven usages, such as RNANet (83) containing the sequences and structures of RNA homologs, and RNA-solo (84) containing non-redundant RNA-only structures in PDB format instead of the IDs from BGSU RNA site (82). The PDB IDs of these 191 RNAs were also listed in Table S1 in the Supplementary Material. It is noted that a few of RNAs in the test sets have sequence identity $>80\%$ and coverage $>80\%$ with the RNAs in the training native set, and for maintaining the complete structure spectrum, we still kept these RNAs in the training native set (71). For these RNAs in test datasets, the leave-one-out/jackknife method was used in examining the performance of cgRNASP (71,79).

To optimize the weights in cgRNASP, we built a training decoy dataset through four RNA 3D structure prediction models (FARFAR2 (12), RNAComposer (17), SimRNA (43) and 3dRNA v2.0 (21)) with input secondary structures. Beyond the training decoy dataset for rsRNASP (79), the input secondary structures include not only native ones (85,86), but also those from secondary predictions models of RNAfold (87) and 2dRNA (88) for each RNA. Thus, the built decoy training set contains 35 single-stranded RNAs with a wide structure spectrum at both 3D level and secondary level. Please see Section S2 in the Supplementary Material for the details of building the training decoy dataset, and the dataset can be found at <https://github.com/Tan-group/cgRNASP>. In cgRNASP, an RNA length N -dependent function $f(N)$ was also involved to normalize the N -dependent CG atom-pair number of the long-ranged interactions due to the large residue-separation range and the consequent N -dependent CG atom-pair number (79), and consequently, ω in Eq. (1) is equal to $\omega = \omega_0/f(N)$. According to the all-atom rsRNASP, k_0 was taken as 5 (79).

In cgRNASP, the distance bin width was taken as 0.3 Å (72,74,79), and the distance cutoffs for the potentials in k

ranges of 1, 2, 3–4 and $k \geq 5$ were set to the values according to the distance distributions between CG beads in the respective residue separation ranges; see the Supplementary Material for details. For the situation where some atom pairs were not observed within a certain bin width, the potentials were set to the highest potential value in the whole range for corresponding CG atom pair types and $k_B T$ was taken as the unit of potential energy. Based on the examinations on the training decoy dataset, the parameters were determined for cgRNASP, cgRNASP-PC and cgRNASP-C, respectively. Please see Section S2 and Figures S1 and 2 in the Supplementary Material for the details of $f(N)$ and the determination of parameters.

Test datasets

To examine the performance of cgRNASP (at different CG levels) and make comparisons with other existing statistical potentials/scoring functions, extensive typical test datasets were used including MD, NM, PM, and RNA-Puzzles datasets generated by different methods.

Test set MD consisting of five RNAs with about 3500 decoy structures for each RNA was generated by Bernauer *et al.* (70) through replica-exchange molecular dynamics simulations with constraints and the RMSDs of the decoy structures are mainly distributed in a narrow range of 0–3 Å (74); test set NM composed of 15 RNAs with 500 decoy structures for each RNA was also generated by Bernauer *et al.* (70) through normal mode perturbation method and the RMSDs of the decoy structures are also mainly distributed in a narrow range of 1–5 Å (74); test set PM consisting of 20 RNAs with about 40 decoy structures for each RNA was generated recently by us through four RNA 3D structure prediction models (FARFAR2 (12), RNA-Composer (17), SimRNA (43) and 3dRNA v2.0 (21)) with given native secondary structures parsed by X3DNA-DSSR (85,86) from corresponding native 3D structures, and test set PM can be downloaded from <https://github.com/Tang-group/rsRNASP> (79); test set RNA-Puzzles containing both 22 target RNA structures and dozens of predicted (decoy) structures from different RNA structure prediction models for the 22 target RNA structures, which come from the CASP-like competition of blind 3D RNA structure predictions and can be downloaded from https://github.com/RNA-Puzzles/standardized_dataset (76). Here, test sets PM and RNA-Puzzles were called ‘realistic datasets’ to distinguish them from those obtained by perturbation methods, since they are composed of decoy structures of large RNAs generated from different 3D structure prediction models and the RMSDs of decoy structures are distributed in a wide range of ~2–34 Å (74). The RNA-Puzzles dataset is of particular importance since it was generated from the blind CASP-like 3D RNA structure predictions from various top research groups with given sequences (76).

Measuring RNA structure similarity and evaluation metrics

DI (deformation index), a combined metric between RMSD and INF for describing base interactions, was used to measure the structural similarity between two all-atom

RNA structures, and DI is defined as (89):

$$DI(A, B) = \frac{RMSD(A, B)}{INF(A, B)} \quad (6)$$

where $RMSD(A, B)$ and $INF(A, B)$ reflects the difference in geometry and topology between structures A and B, respectively. $RMSD(A, B)$ is the root-mean-square-deviation (RMSD) between structures A and B (90). $INF(A, B)$ is the interaction network fidelity between structures A and B, and is measured by Matthews correlation coefficient of base-pairing and base-stacking interactions (91). It should be noted that the INF used in calculating DI was obtained by RNA-Puzzles toolkit for all interactions, including base stacking, canonical and non-canonical base pairing interactions in all-atom RNA 3D structures (92). The INF from an all-atom RNA structure provides a unified metric for the performances of coarse-grained and all-atom statistical potentials. The tools for calculating DI and INF can be downloaded from <https://github.com/RNA-Puzzles/BasicAssessMetrics> (92).

To describe the performance of cgRNASP, in analogy to previous works (73,74,79), we used the number of native structures with the lowest energy, the DI of lowest-energy structure (including the native one), and the DI of lowest-energy decoy structure (excluding native one) as evaluation metrics for identifying native/near-native structures. Moreover, we used Pearson correlation coefficient (PCC) (71,79) as an evaluation metric for ranking decoy structures, and PCC for decoy structures of an RNA can be calculated as (71):

$$PCC = \frac{\sum_{n=1}^N (E_n - \bar{E})(R_n - \bar{R})}{\sqrt{\sum_{n=1}^N (E_n - \bar{E})^2} \sqrt{\sum_{n=1}^N (R_n - \bar{R})^2}} \quad (7)$$

where E_n and R_n are the energy and DI of the n th decoy for the RNA, respectively. \bar{E} and \bar{R} are the average energy and DI of all decoys for the RNA, respectively. The value of PCC ranges from 0 to 1, and PCC of 1 represents a perfect performance of the statistical potential. In determining the parameters of cgRNASP (and cgRNASP-PC and cgRNASP-C), we maximized an integrated metric $N\% \times PCC$ (percentage $N\%$ of identified native structures multiplied by PCC) against the training decoy dataset to obtain the final parameters; see Section S2 and Figure S3 in the Supplementary Material for the details.

RESULTS AND DISCUSSION

In the following, we examined the performance of cgRNASP against extensive types of test datasets. As described in the subsection of test datasets, test sets MD and NM were mainly generated from perturbation methods, while test sets PM and RNA-Puzzles were generated from various 3D structure prediction models and the RNA-Puzzles dataset from the blind CASP-like RNA structure predictions is of particular importance. Therefore, we first examined the overall performance of 3-bead cgRNASP against all the test sets, and afterward focused on the performance of cgRNASP at different CG levels and the computation efficiency of cgRNASP against the realistic dataset, RNA-Puzzles.

Overall performance of cgRNASP on test datasets

Overall performance of cgRNASP on all the test datasets. As shown in Figure 2A and Table S2 in the Supplementary Material, for all the test sets including the test sets MD and NM with small RNAs from perturbation methods and the test sets PM and RNA-Puzzles with large RNAs from various 3D structure prediction models, cgRNASP identifies ~77% native structures, i.e. 48 native structures out of decoy ones of 62 RNAs. Such values are identical to those of the newly developed all-atom rsRNASP (~77% and 48 out of 62 RNAs), and appear higher than the all-atom statistical potentials/scoring functions of RNA3DCNN (~74% and 46 out of 62), ARES (~11% and 7 out of 62), DFIRE-RNA (~56% and 35 out of 62), 3dRNAscore (~34% and 21 out of 62), and RASP (~26% and 16 out of 62). This indicates that cgRNASP identifies the similar number of native structures with rsRNASP while more native ones than other all-atom statistical potentials/scoring functions for all the test datasets. Furthermore, we calculated the mean DI of lowest-energy structures including native ones and excluding native ones. As shown in Figures 2B and C and Tables S3 and S4 in the Supplementary Material, the mean DI of lowest-energy structures and that excluding native structures calculated from cgRNASP are 2.1 and 6.9 Å, which are identical to those from the all-atom rsRNASP (2.1 and 6.9 Å). However, such two values from cgRNASP are both visibly smaller than those from other all-atom statistical potential/scoring functions of RNA3DCNN (3.1 and 8.3 Å), ARES (9.3 and 9.5 Å), DFIRE-RNA (4.8 and 7.8 Å), 3dRNAscore (8.3 and 9.6 Å), and RASP (9.1 and 10.4 Å), suggesting that cgRNASP identifies the structures more similar to native ones.

To examine the performance of cgRNASP in ranking decoy structures, we calculated the PCC values between energies and DIs for all the test datasets by cgRNASP and other all-atom statistical potentials/scoring functions; see Eq. (7). As shown in Figure 2D and Table S5 in the Supplementary Material for all the test sets, cgRNASP has an overall comparable performance to the all-atom rsRNASP in ranking decoy structures since the average PCC value (0.71) from cgRNASP is slightly smaller than that (0.73) from all-atom rsRNASP. Furthermore, the PCC value from cgRNASP is slightly higher than that (PCC ~ 0.70) from the all-atom DFIRE-RNA while is apparently larger than other all-atom statistical potential/scoring function of RNA3DCNN (PCC ~ 0.61), ARES (PCC ~ 0.60), 3dRNAscore (PCC ~ 0.55) and RASP (PCC ~ 0.53), respectively. Namely, for all the test datasets, cgRNASP has similar performance to the all-atom rsRNASP and DFIRE-RNA while apparently better performance than other all-atom statistical potentials in ranking decoy structures. The RMSD-energy scatter-plots for all RNAs in the MD, NM, PM and RNA-Puzzles datasets by cgRNASP can be found in Figures S4–S6 in the Supplemental Material. The detailed comparison between cgRNASP and existing all-atom statistical potentials/scoring functions for all the test sets can also be illustrated by violin plots in Figure S7 in the Supplementary Material, respectively.

Therefore, for all the test datasets, the present (coarse-grained) cgRNASP has an overall similar performance

with the newly developed all-atom rsRNASP while is overall superior to other existing all-atom statistical potential/scoring functions in identifying RNA native/near-native structures and in ranking decoy structures.

Performance of cgRNASP on realistic datasets—PM and RNA-puzzles. Test datasets PM and RNA-Puzzles generated from various RNA 3D structure prediction models contain large RNA decoy structures with a large RMSD range, and consequently can serve as realistic test sets for a statistical potential beyond test sets MD and NM composed of small RNAs and near-native decoy structures from perturbation methods; see the subsection of test datasets.

As shown in Figures 2E–G and Tables S2–S4 in the Supplementary Material, for the PM and RNA-Puzzles datasets, cgRNASP identified 79% native structures (33 out of 42), and the mean DI of the lowest-energy structures and that excluding native ones from cgRNASP are 3.9 and 12.7 Å, respectively. Such values are very slightly superior to those from the all-atom rsRNASP (~76%, 3.9 Å and 12.8 Å), suggesting a very slightly better performance of cgRNASP than the all-atom rsRNASP in identifying native/near-native structures for test datasets PM and RNA-Puzzles. However, the comparison of the three metrics values with other all-atom statistical potentials/scoring functions indicates that cgRNASP has apparently better performance than RNA3DCNN (~64%, 6.1 Å and 15.5 Å), ARES (~5%, 16.7 Å and 17.1 Å), DFIRE-RNA (~48%, 8.9 Å and 13.7 Å), 3dRNAscore (~10%, 16.5 Å and 18.3 Å) and RASP (~10%, 17.4 Å and 19.4 Å).

Moreover, as shown in Figure 2H and Table S5 for the PM and RNA-Puzzles datasets, the PCC value (0.60) from cgRNASP is very close to that (0.61) from rsRNASP while is apparently higher than other all-atom statistical potentials/scoring functions of RNA3DCNN (0.41), ARES (0.38), DFIRE-RNA (0.53), 3dRNAscore (0.27) and RASP (0.26). This suggests that cgRNASP has the same performance as the all-atom rsRNASP and apparently better performance than other all-atom statistical potentials/scoring functions in ranking decoy structure for the PM and RNA-Puzzles datasets. Besides, the detailed comparison between cgRNASP and existing all-atom statistical potentials/scoring functions for the PM and RNA-Puzzles datasets can also be shown by violin plots in Figure S8 in the Supplementary Material.

Therefore, overall, for realistic datasets including PM and RNA-Puzzles, the present coarse-grained cgRNASP has very slightly better performance than the newly developed all-atom rsRNASP in identifying native/near-native structures, while appears apparently superior to other all-atom statistical potentials/scoring functions in identifying native/near-native structures and in ranking decoy ones.

Performance of cgRNASP on RNA-puzzles dataset. Test dataset RNA-Puzzles composed of 22 RNAs has been widely considered a realistic test dataset and consequently is of particular importance since it was generated from the blind CASP-like 3D RNA structure predictions from various top research groups with given sequences (74).

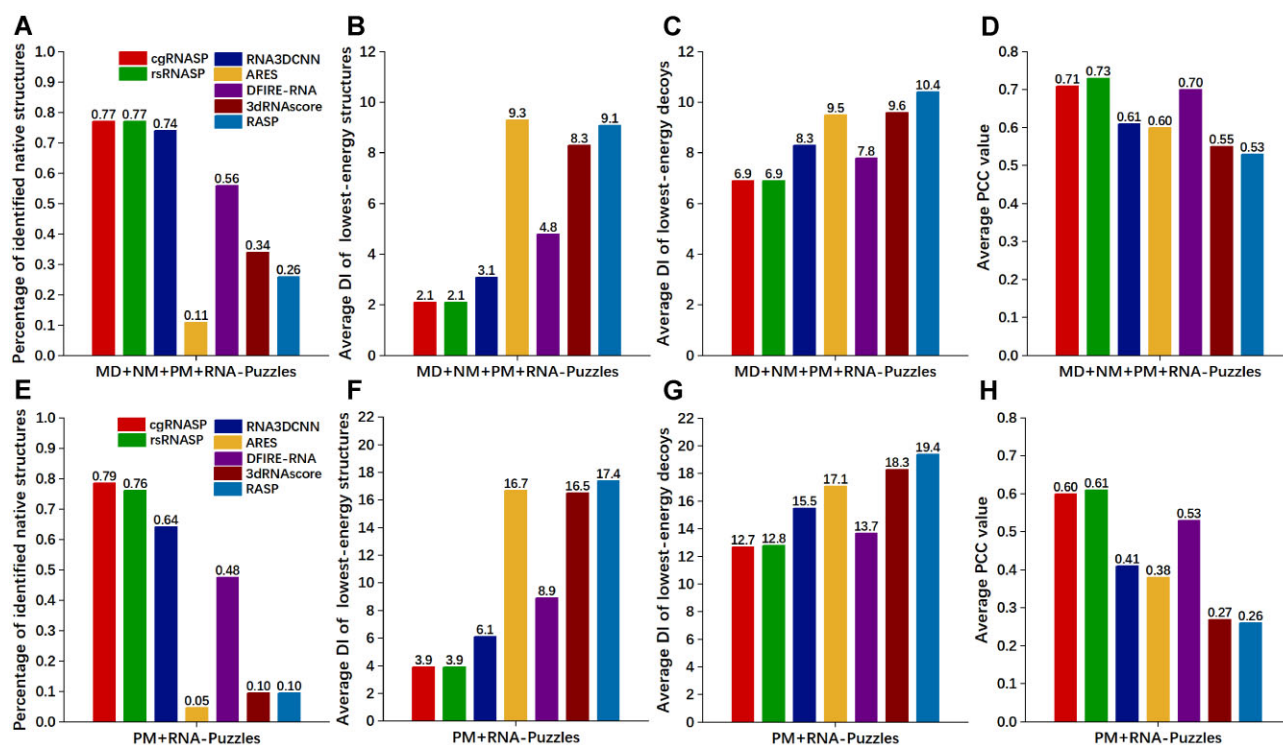


Figure 2. (A, E) Percentages of the number of identified native structures, (B, F) average DI values of structures with the lowest energy (including native ones), (C, G) average DI values of decoy structures with the lowest energy (excluding native ones), and (D, H) average PCC values between DIs and energies by coarse-grained cgRNASP and other all-atom statistical potentials. Panels (A–D) are for all the test sets (MD + NM + PM + RNA-Puzzles), and (E–H) are for the realistic test sets (PM + RNA-Puzzles) from various 3D structure prediction models. Here, the PCC values for all the test sets (MD + NM + PM + RNA-Puzzles) and for the realistic sets (PM + RNA-Puzzles) were averaged over the mean values of respective test sets since decoys in a test set have similar structure features.

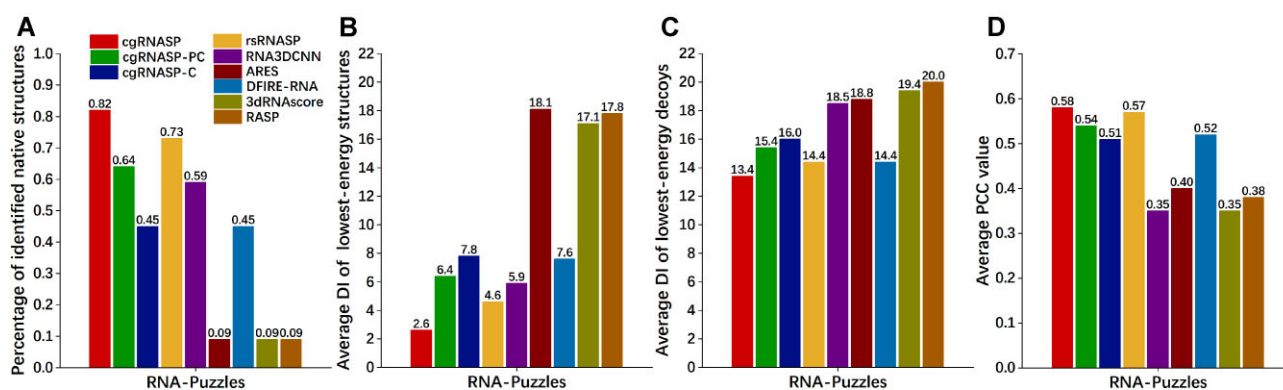


Figure 3. (A) Percentages of the number of identified native structures, (B) average DI values of structures with the lowest energy (including native ones), (C) average DI values of decoy structures with the lowest energy (excluding native ones) and (D) average values of PCCs between DIs and energies by cgRNASP at different CG levels (cgRNASP, cgRNASP-PC and cgRNASP-C) and other all-atom statistical potentials for the RNA-Puzzles dataset.

As shown in Figures 3A–C and Tables S2–S4 in the Supplementary Material, for the RNA-Puzzles dataset, cgRNASP identifies 18 native structures out from decoys of 22 RNAs, and the mean DI of lowest-energy structures and that excluding native ones from cgRNASP are 2.6 and 13.4 Å. In contrast, rsRNASP identifies 16 native ones out from decoys of 22 RNAs and the mean DI of lowest-energy structures and that excluding native ones from rsRNASP are 4.6 and 14.4 Å, respectively. Moreover, the three metrics values from cgRNASP appear apparently better than

those from other statistical potential/scoring functions of RNA3DCNN (13, 5.9 Å and 18.5 Å), ARES (2, 18.1 Å and 18.8 Å), DFIRE-RNA (10, 7.6 Å and 14.4 Å), 3dRNAscore (2, 17.1 Å and 19.4 Å) and RASP (2, 17.8 Å and 20.0 Å). Thus, for the RNA-Puzzles dataset, cgRNASP is slightly superior to rsRNASP and has apparently better performance than other all-atom statistical potentials/scoring functions in identifying native/near-native structures.

Furthermore, as shown in Figure 3D and Table S5 for the RNA-Puzzles dataset, the PCC between DIs and en-

ergies calculated from cgRNASP is 0.58, which is slightly higher than that from rsRNASP (0.57). Moreover, such value of cgRNASP is apparently higher than those from other all-atom statistical potentials/scoring functions of RNA3DCNN (0.35), ARES (0.40), DFIRE-RNA (0.52), 3dRNAscore (0.35) and RASP (0.38). Thus, for the RNA-Puzzles dataset, cgRNASP is slightly better than rsRNASP and is apparently superior to other all-atom statistical potentials/scoring functions in ranking decoy structures. Moreover, the detailed comparison between cgRNASP and existing all-atom statistical potentials/scoring functions for the RNA-Puzzles dataset can be illustrated by violin plots in Figure S9 in the Supplementary Material.

Additionally, since the non-canonical base pairs are important structural elements in RNAs, we examined the performance of cgRNASP for RNAs with abundant non-canonical base pairs. Nine RNAs (rp05, rp06, rp09, rp10, rp12, rp14_free, rp14_bound, rp17 and rp21) from RNA-Puzzles dataset were selected to calculate the PCCs between DIs and energies by cgRNASP and other all-atom statistical potentials, and for the native structures of these RNAs, the percentage of non-canonical base pairs over total base pairs resolved by X3DNA-DSSR (85,86) is more than 35%. For the decoy sets of these nine RNAs with native non-canonical base pairs, cgRNASP also has good performance, compared with other all-atom statistical potentials; see Table S6 in the Supplementary Material. Furthermore, we examined the non-Watson-Crick (non-canonical) interactions by calculating INFs for the RNAs in the RNA-Puzzles dataset. The calculated INFs for the RNA-Puzzles dataset have been shown in Figure S10 and Table S7 in the Supplementary Material including the INFs for all interactions (INF_all) and non-Watson-Crick (non-canonical) interactions (INF_nwc), as well as the INFs of structures with lowest energies by cgRNASP. It is shown that the INF_nwc distribution highlights the insufficient predictions for non-Watson-Crick interactions from various models involved in the RNA-Puzzles dataset (92). Moreover, as shown in Table S7 in the Supplementary Material, the average INF_all and INF_nwc of the lowest-energy structures identified by cgRNASP for the RNA-Puzzles dataset including native structures are ~ 0.96 and ~ 0.93 , which are both close to the maximum INF value (1.0). However, for the RNA-Puzzles dataset excluding native structures, the average INF_all (~ 0.78) of the lowest-energy structures by cgRNASP is also close to the maximum INF_all (~ 0.83), while the average INF_nwc (~ 0.34) is visibly lower than maximum INF_nwc (~ 0.52). This indicates that cgRNASP has better performance in identifying the whole interactions than in identifying the non-canonical interactions. Thus, it is still necessary to develop high-performance statistical potentials/scoring functions and structure prediction models with good ability of capturing non-canonical base-pair interactions.

Afterwards, to examine the performance of cgRNASP against the predicted structures from different structure prediction groups, we separated the predicted structures from Bujnicki, Das and Chen groups in the RNA-Puzzles dataset. As shown in Table S8 in the Supplementary Material, the performances of cgRNASP against the structures predicted from these three groups follow the order

of Das group > Chen group > Bujnicki group in identifying near-native structures and follow the order of Bujnicki group > Chen group > Das group in ranking structures. It is also noted that, cgRNASP performs excellently in identifying near-native structures for the structures predicted from Das group and in ranking the structures predicted from Bujnicki group.

Therefore, for the RNA-Puzzles dataset from CASP-like blind 3D structure prediction competition, our present coarse-grained cgRNASP is slightly better than the recently developed all-atom rsRNASP and is apparently superior to the other all-atom statistical potentials/scoring functions in identifying native/near-native structures and in ranking decoy ones.

Performance of cgRNASP at different CG levels on RNA-puzzles dataset

In this subsection, we would examine the relative performance of cgRNASP at different CG levels on the RNA-Puzzles dataset, i.e. 3-bead cgRNASP, 2-bead cgRNASP-PC, and 1-bead cgRNASP-C; see Materials and Methods.

As shown in Figures 3A-C, cgRNASP, cgRNASP-PC and cgRNASP-C identify 82% (18 out of 22), 64% (14 out of 22) and 45% (10 out of 22) native structures in the RNA-Puzzles dataset, respectively. The mean DI with the lowest-energy structures and that excluding native ones are 2.6 and 13.4 Å for cgRNASP, 6.4 and 15.4 Å for cgRNASP-PC, and 7.8 and 16.0 Å for cgRNASP-C, respectively. Thus, in identifying native/near-native structures, the performance of cgRNASP at different CG levels follows the order of cgRNASP > cgRNASP-PC > cgRNASP-C for the RNA-Puzzles dataset, which is consistent with the number of CG beads used in the respective potentials. Furthermore, Figure 3D shows that the PCC values between energies and DIs of decoys are 0.58, 0.54 and 0.51 for cgRNASP, cgRNASP-PC and cgRNASP-C, respectively. Thus, the performance in ranking decoy structures in the RNA-Puzzles dataset follows the order of cgRNASP > cgRNASP-PC > cgRNASP-C, which is identical to the above order in identifying native/near-native structures. It is understandable that the performance of cgRNASP follows the order of cgRNASP > cgRNASP-PC > cgRNASP-C, since more CG beads would generally involve more geometric information and constraints for a structure and consequently would be more effective for structure evaluation. The detailed comparison among cgRNASP at different CG levels for the RNA-Puzzles dataset can be illustrated by violin plots in Figure S9 in the Supplementary Material.

Overall, compared with other all-atom statistical potentials/scoring functions, for the realistic dataset, RNA-Puzzles dataset, cgRNASP has slightly better performance with mean DI of lowest-energy structures ~ 2.6 Å and PCC ~ 0.58 than the all-atom rsRNASP (with the two values of 4.6 Å and 0.57). Moreover, cgRNASP-PC has also an acceptable performance with mean DI of lowest-energy structures ~ 6.4 Å and PCC ~ 0.54 overall beyond existing all-atom statistical potentials/scoring functions except for rsRNASP. Additionally, it is encouraging that cgRNASP-C based on an extremely reduced 1-bead

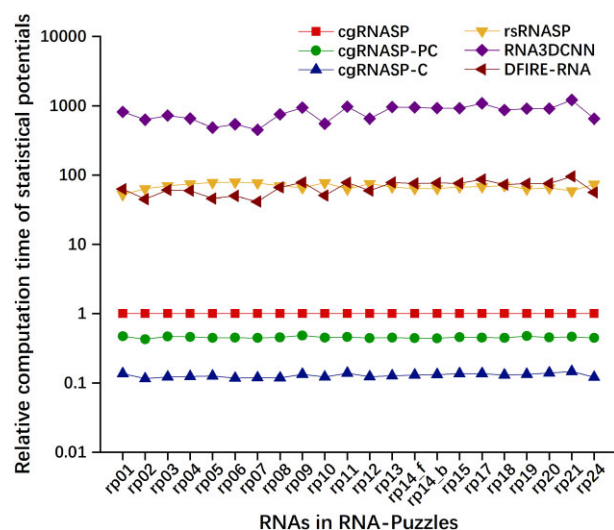


Figure 4. Computation times of cgRNASP at different CG levels (cgRNASP, cgRNASP-PC and cgRNASP-P) and other top all-atom statistical potentials for the RNAs in the RNA-Puzzles dataset, relative to that of 3-bead cgRNASP. Here, rp14-f and rp14-b stand for rp14-free and rp14-bound, respectively.

representation of C4' atom has a similar performance to the all-atom DIFRE-RNA and appears superior to the all-atom RNA3DCNN in ranking decoys.

Evaluation efficiency of cgRNASP on RNA-puzzles dataset

A reliable CG-based statistical potential can be used not only for evaluating CG structures but also for evaluating all-atom structures at high efficiency due to the greatly reduced atom representation. Here, we would examine the computation efficiency of cgRNASP on the RNA-Puzzles dataset, in a comparison with existing top all-atom statistical potentials/scoring functions.

As shown in Figure 4, for evaluating decoys of RNAs in the RNA-Puzzles dataset, cgRNASP is strikingly more efficient than the all-atom statistical potentials/scoring functions of rsRNASP, RNA3DCNN and DFIRE-RNA. Specifically, for the RNA-Puzzles dataset, the mean computation time of (3-bead) cgRNASP is only $\sim 1/65$ of that of rsRNASP, while rsRNASP has a comparable computation time with DFIRE-RNA and is ~ 10 times more efficient than RNA3DCNN. Furthermore, the computation time of cgRNASP slightly decreases with the decrease in the number of CG beads used in cgRNASP. For example, cgRNASP is ~ 2 times less efficient than cgRNASP-PC and is ~ 9 times less efficient than cgRNASP-C. It is understandable that cgRNASP is strikingly more efficient than all-atom statistical potentials/scoring functions and cgRNASP with fewer CG beads is more efficient since the computation time of a statistical potential should be approximately proportional to the square of the number of (CG) atoms per nucleotide involved in the potential.

The strikingly higher efficiency of cgRNASP with good performance would be very beneficial to RNA 3D structure evaluation through either greatly saving evaluation time or evaluating much more structure candidates within a given time.

How can cgRNASP with coarse-grained representation have good performance?

As shown above, the present (3-bead) cgRNASP with coarse-grained representation can have similar performance for all the test datasets and even slightly better performance for the RNA-Puzzles dataset, compared with the newly developed all-atom rsRNASP. Then how can cgRNASP have good performance with highly reduced (coarse-grained) representation?

First, we examined the individual contributions of short- and long-ranged interactions against the RNA-Puzzles dataset. As shown in Figures 5A–D, in identifying native structures and in ranking decoy ones, the long-ranged contribution in cgRNASP performs comparably similarly to that in rsRNASP which identified seven native structures and yielded an average PCC value of 0.53 (79), while the short-ranged contribution with $3 \leq k < 5$ in cgRNASP performs worse than that in rsRNASP which identified 11 native structures and yielded an average PCC value of 0.45 (79); see also Figure 3 in (79). However, the explicit involvement of the potentials with $k = 1$ and $k = 2$ causes that the short-ranged potential of cgRNASP has a comparable performance to that of rsRNASP without involving the potentials with $k = 1$ and $k = 2$ (79). Therefore, it is not very strange that cgRNASP is comparable and even slightly superior to rsRNASP for the RNA-Puzzles dataset, which is attributed to the more complete/subtle involvement of short-ranged potentials at different residue separations in cgRNASP. In addition, the good performance of cgRNASP may also be attributed to the 3-bead CG representation used in cgRNASP which can well describe both of backbone and bases for RNAs despite of the 3-bead CG approximation.

Furthermore, we examined the microscopic interactions captured in cgRNASP, comparatively with those in rsRNASP (79). As shown in Figure 5E and F for the potentials between N9 atom of adenine (A) and N1 atom of uracil (U) as a paradigm, the long-ranged potential in cgRNASP captures similar energy minimums with that in rsRNASP while the short-ranged potentials in cgRNASP can capture more energy minimums than that in rsRNASP; see also Figures 4AB in (79). This is consistent with the above-discussed performances of short- and long-ranged contributions in cgRNASP. Compared with rsRNASP (79), the explicit and complete involvement of short-ranged potentials at different residue separations in cgRNASP leads to more detailed inter-bead (atom) interactions captured in the short-ranged potential including nearest-neighbor base stacking (at ~ 4.6 Å), next-nearest-neighbor U-turns (at ~ 4.3 and ~ 5.7 Å), and next-nearest-neighbor base stacking (at ~ 8.9 Å). Furthermore, the short-ranged potentials in cgRNASP also capture the similar interactions with rsRNASP such as canonical Watson-Crick base pairing (at ~ 8.9 Å), the intra-loop base interactions (at ~ 5.7 Å), and the reverse Hoogsteen base-pairing interaction (at ~ 7.1 Å) (79). In addition, similar to rsRNASP (79), the long-ranged potential in cgRNASP also captures some tertiary base-stacking interactions including base-stacking between two residues in two adjacent branches of RNA 3D structures (e.g. tetraloop receptor), coaxial-stacking at junction region and base-stacking in the triple-helical region; see Figure S11 in the Supplementary Material and see also (79). Addition-

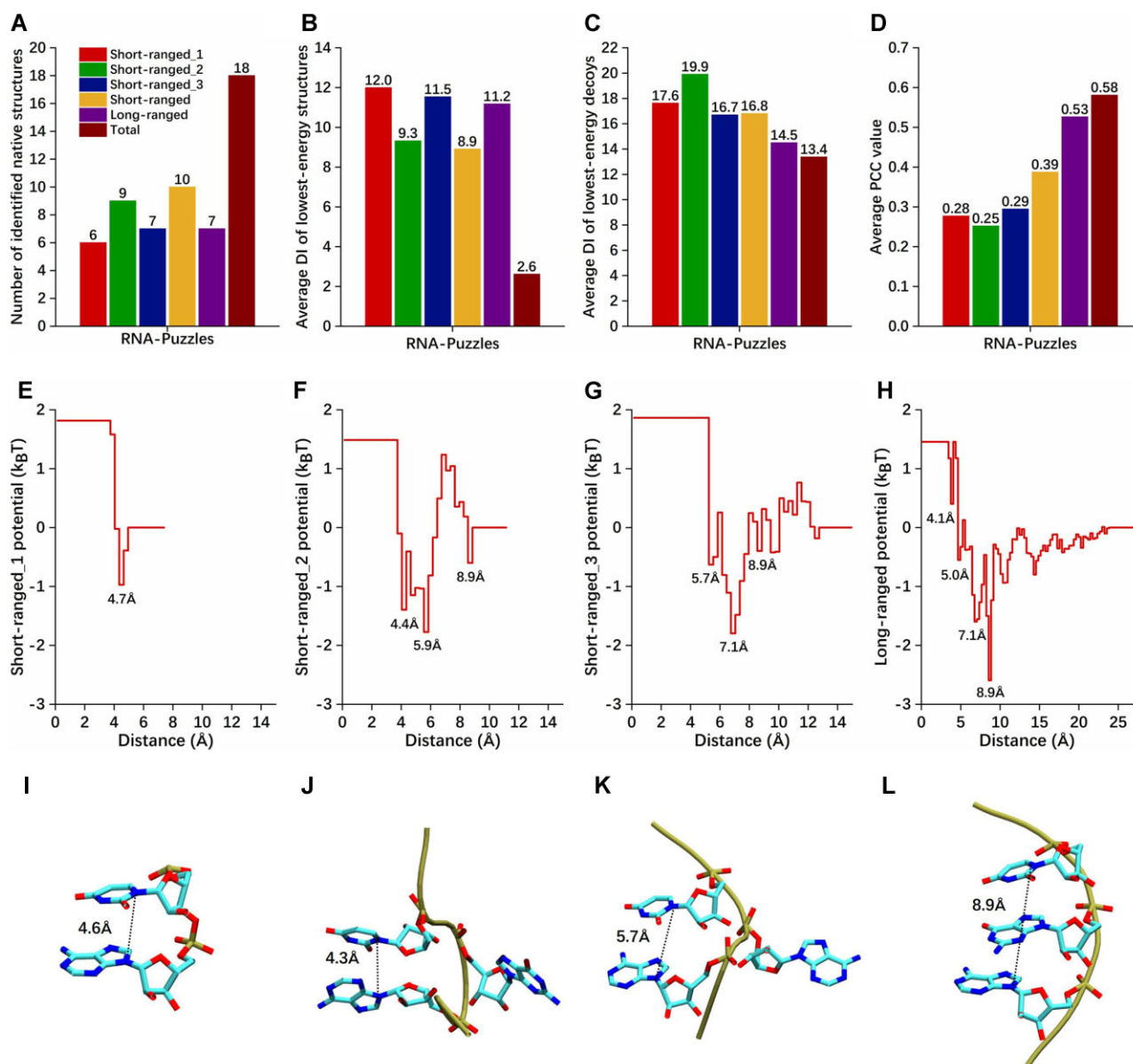


Figure 5. (A) Percentages of the number of identified native structures, (B) average DI values of structures with the lowest energy (including native ones), (C) average DI values of decoy structures with the lowest energy (excluding native ones) and (D) average values of PCCs on DIs by short-ranged and long-ranged energies from cgRNASP. (E–G) Short-ranged and (H) long-ranged potentials between AN9 and UN1 in cgRNASP. In panels (A–H), short-ranged_1, short-ranged_2 and short-ranged_3 potentials denote those at residue separations of $k = 1$, $k = 2$ and $3 \leq k < 5$, respectively. (I, J) Representative distances between AN9 and UN1 for nearest-neighbor base stacking (I), next-nearest neighbor U-turns (J and K), and next-nearest neighbor base-stacking (L) captured in the short-ranged potentials. Here, we did not show other representative distances for reverse Hoogsteen (~ 7.1 Å) and Watson–Crick base pairing (~ 8.9 Å) in the short-ranged_3 potential, as well as those for base stacking between adjacent branches (~ 4 Å), base stacking at triplex (~ 5 Å) and coaxial-stacking at junctions (~ 5 Å) in the long-ranged potential, respectively, which have been illustrated in the Supplementary Material and (79).

ally, we examined the performance of cgRNASP against the RNAs with abundant non-canonical base pairs in the RNA-Puzzles dataset, and found that cgRNASP has a better average performance for these RNAs than other statistical potentials; see Table S6 in the Supplementary Material. As a typical example, the potential of a non-canonical base pair between N9 atom of adenine (A) and N9 atom of guanine (G) in cgRNASP has also been shown in Figure S12 in the Supplementary Material, indicating a non-canonical base pairing interaction. These indicate that the

non-canonical base pair interactions may be partially captured in cgRNASP in an implicit way. However, it should be noted that the simplified representation of RNA structures in cgRNASP would definitely lead to the accuracy loss for describing the non-canonical base pairs which could be better described by the all-atom representation. Moreover, the existing RNA structures deposited in the Protein Data Bank are still too limited and it is difficult to include all types of non-canonical base pairs in the training set due to the diverse types of non-canonical base pairs (Shi-Jie Chen,

private communication). Consequently, the present 3-bead cgRNASP should not well discriminate all types of non-canonical base pairs due to the coarse-graining simplification and the limited RNA structures in the Protein Data Bank.

Therefore, the combination of a more detailed short-ranged potential and long-ranged one contribute to the good performance of cgRNASP even with highly reduced (coarse-grained) representation.

CONCLUSION

In this work, we have developed a series of residue-separation-based statistical potentials at different CG levels (cgRNASP) for RNA 3D structure evaluation. The CG potentials of cgRNASP were built through distinguishing short- and long-ranged interactions, and compared with the all-atom rsRNASP, more complete and subtle treatment on the short-ranged interactions was involved in cgRNASP. The tests against extensive types of datasets indicate that, overall, cgRNASP has varying performance with different CG levels, and can have similarly good or even slightly better performance for the realistic dataset—RNA-Puzzles dataset than the all-atom rsRNASP. Furthermore, cgRNASP has strikingly higher computation efficiency than existing all-atom statistical potentials/scoring functions, and has better performance than other existing traditional statistical potentials and scoring functions from neural networks for the realistic test datasets including the RNA-Puzzles dataset. Thus, our cgRNASP can be directly used for evaluating CG structure candidates generated from related CG-based structure prediction models, and can be also very useful for evaluating all-atom structure candidates at strikingly higher efficiency than existing all-atom statistical potentials.

The present CG statistical potentials of cgRNASP can be furtherly improved for more applicable versions to different CG structure prediction models and a more accurate evaluation on RNA 3D structures. First, cgRNASP potentials at different CG levels only involve those heavy atoms rather than dummy atoms (e.g. mass centers of atom groups). For more a specific application, the present cgRNASP can be extended to include those dummy atoms used in some specific CG structure prediction models, and such specifically improved potentials can be more useful for those corresponding CG structure prediction models. Second, some other geometric parameters such as torsional angle and orientation can be involved in cgRNASP to enhance the description of the relative relationship between atoms or atom groups beyond pair atom-atom distance (48,72,93). Third, a multi-body (three- or four-body) potential can be explicitly supplemented to cgRNASP which would improve the description of correlated atom-atom distributions (48,94). Fourth, the reference states in cgRNASP can be possibly circumvented through some treatments such as the iterative technique since the used reference states may still deviate from the ideal one (67,68). Fifth, due to the currently limited RNA native structures in the PDB database (7), cgRNASP can be continuously improved with the increase of the number of experimentally determined RNA native structures. Additionally, as suggested by Carrasco *et al.*, the stereo-

chemistry of RNA structures may be critical to distinguish native structures and decoy ones from structure prediction models (95). Thus, the involvement of stereochemistry in a statistical potential/scoring functions would possibly improve the performance for RNA 3D structure evaluation. Finally, some underlying geometric information may be captured from training through neural network over a high-quality training dataset, and the captured information may be combined to cgRNASP to derive a new CG statistical potential/scoring function with higher performance.

Nevertheless, the present CG statistical potentials of cgRNASP would be definitely beneficial to related CG-based 3D structure evaluations, as well as to all-atom-based 3D structure evaluations at strikingly high computation efficiency.

DATA AVAILABILITY

All relevant data are within the paper and its Supporting Information files. The potentials of cgRNASP at different CG levels and the related datasets are available at website <https://github.com/Tan-group/cgRNASP>.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We are grateful to Profs Shi-Jie Chen (University of Missouri) and Jian Zhang (Nanjing University) for valuable discussions. The numerical calculations in this work were performed on the super computing system in the Super Computing Center of Wuhan University.

Author contributions: Z.J.T., B.Z. and Y.L.T. designed the research. Y.L.T., X.W. and S.X.Y. performed the research. Z.J.T., Y.L.T., X.W. and B.Z. analyzed the data. Y.L.T., X.W., B.Z. and Z.J.T. wrote the manuscript.

FUNDING

National Science Foundation of China [12075171, 11774272, 11971367].

Conflict of interest statement. None declared.

REFERENCES

- Gesteland, R. F., Cech, T. R. and Atkins, J. F. (eds). (2006) In: *The RNA World*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Dethoff, E.A., Chugh, J., Mustoe, A.M. and Al-Hashimi, H.M. (2012) Functional complexity and regulation through RNA dynamics. *Nature*, **482**, 322–330.
- Zhang, Y., Zhang, J. and Wang, W. (2011) Atomistic analysis of pseudoknotted RNA unfolding. *J. Am. Chem. Soc.*, **133**, 6882–6885.
- Wu, Y.Y., Zhang, Z.L., Zhang, J.S., Zhu, X.L. and Tan, Z.J. (2015) Multivalent ion-mediated nucleic acid helix-helix interactions: RNA versus DNA. *Nucleic Acids Res.*, **43**, 6156–6165.
- Doherty, E.A. and Doudna, J.A. (2001) Ribozyme structures and mechanisms. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 457–475.
- Edwards, T.E., Klein, D.J. and Ferré-D'Amaré, A.R. (2007) Riboswitches: small-molecule recognition by gene regulatory rnas. *Curr. Opin. Struct. Biol.*, **17**, 273–279.

7. Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.
8. Sun, L.Z., Zhang, D. and Chen, S.J. (2017) Theory and modeling of RNA structure and interactions with metal ions and small molecules. *Annu. Rev. Biophys.*, **46**, 227–246.
9. Miao, Z. and Westhof, E. (2017) RNA structure: advances and assessment of 3D structure prediction. *Annu. Rev. Biophys.*, **46**, 483–503.
10. Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci.*, **104**, 14664–14669.
11. Das, R., Karanicolas, J. and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.
12. Watkins, A.M., Rangan, R. and Das, R. (2020) FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure*, **28**, 963–976.
13. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
14. Zhang, J., Lin, M., Chen, R., Wang, W. and Liang, J. (2008) Discrete state model and accurate estimation of loop entropy of RNA secondary structures. *J. Chem. Phys.*, **128**, 03B624.
15. Zhang, J., Dundas, J., Lin, M., Chen, R., Wang, W. and Liang, J. (2009) Prediction of geometrically feasible three-dimensional structures of pseudoknotted RNA through free energy estimation. *RNA*, **15**, 2248–2263.
16. Zhang, J., Bian, Y., Lin, H. and Wang, W. (2012) RNA fragment modeling with a nucleobase discrete-state model. *Phys. Rev. E*, **85**, 021909.
17. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K.J., Lukasiak, P., Bartol, N., Blazewicz, J. and Adamiak, R.W. (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res.*, **40**, e112.
18. Sim, A.Y.L., Levitt, M. and Minary, P. (2012) Modeling and design by hierarchical natural moves. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 2890–2895.
19. Wang, J., Mao, K., Zhao, Y., Zeng, C., Xiang, J., Zhang, Y. and Xiao, Y. (2017) Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. *Nucleic Acids Res.*, **45**, 6299–6309.
20. Zhao, Y., Huang, Y., Gong, Z., Wang, Y., Man, J. and Xiao, Y. (2012) Automated and fast building of three-dimensional RNA structures. *Sci. Rep.*, **2**, 734.
21. Wang, J., Wang, J., Huang, Y. and Xiao, Y. (2019) 3dRNA v2.0: an updated web server for RNA 3D structure prediction. *Int. J. Mol. Sci.*, **20**, 4116.
22. Bida, J.P. and Maher III, L.J. (2012) Improved prediction of RNA tertiary structure with insights into native state dynamics. *RNA*, **18**, 385–393.
23. Cao, S. and Chen, S.J. (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*, **11**, 1884–1897.
24. Cao, S. and Chen, S.J. (2011) Physics-based de novo prediction of RNA 3D structures. *J. Phys. Chem. B*, **115**, 4216–4226.
25. Xu, X., Zhao, P. and Chen, S.J. (2014) Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One*, **9**, e107504.
26. Zhang, D. and Chen, S.J. (2018) IsRNA: an iterative simulated reference state approach to modeling correlated interactions in RNA folding. *J. Chem. Theory Comput.*, **14**, 2230–2239.
27. Zhang, D., Li, J. and Chen, S.J. (2021) IsRNA1: de novo prediction and blind screening of RNA 3D structures. *J. Chem. Theory Comput.*, **17**, 1842–1857.
28. Tan, R.K., Petrov, A.S. and Harvey, S.C. (2006) YUP: a molecular simulation program for coarse-grained and multiscaled models. *J. Chem. Theory Comput.*, **2**, 529–540.
29. Ding, F., Sharma, S., Chalasani, P., Demidov, V.V., Broude, N.E. and Dokholyan, N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
30. Sharma, S., Ding, F. and Dokholyan, N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.
31. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. and Altman, R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
32. Mustoe, A.M., Al-Hashimi, H.M. and Brooks, C.L. (2014) Coarse grained models reveal essential contributions of topological constraints to the conformational free energy of RNA bulges. *J. Phys. Chem. B*, **118**, 2615–2627.
33. Shi, Y.Z., Wang, F.H., Wu, Y.Y. and Tan, Z.J. (2014) A coarse-grained model with implicit salt for RNAs: predicting 3D structure, stability and salt effect. *J. Chem. Phys.*, **141**, 105102.
34. Shi, Y.Z., Jin, L., Wang, F.H., Zhu, X.L. and Tan, Z.J. (2015) Predicting 3D structure, flexibility, and stability of RNA hairpins in monovalent and divalent ion solutions. *Biophys. J.*, **109**, 2654–2665.
35. Jin, L., Shi, Y.Z., Feng, C.J., Tan, Y.L. and Tan, Z.J. (2018) Modeling structure, stability, and flexibility of double-stranded RNAs in salt solutions. *Biophys. J.*, **115**, 1403–1416.
36. Shi, Y.Z., Jin, L., Feng, C.J., Tan, Y.L. and Tan, Z.J. (2018) Predicting 3D structure and stability of RNA pseudoknots in monovalent and divalent ion solutions. *PLoS Comput. Biol.*, **14**, e1006222.
37. Jin, L., Tan, Y.L., Wu, Y.Y., Wang, X. and Tan, Z.J. (2019) Structure folding of RNA kissing complexes in salt solutions: predicting 3D structure, stability, and folding pathway. *RNA*, **25**, 71119–71662.
38. Zhang, B., Qiu, H., Jiang, J., Liu, J. and Shi, Y.Z. (2019) 3D structure stability of the HIV-1 TAR RNA in ion solutions: a coarse-grained model study. *J. Chem. Phys.*, **151**, 165101.
39. Feng, C., Tan, Y.L., Chen, Y.X., Shi, Y.Z. and Tan, Z.J. (2021) Salt-dependent RNA pseudoknot stability: effect of spatial confinement. *Front. Mol. Biosci.*, **8**, 666369.
40. Pasquali, S. and Derreumaux, P. (2010) HiRE-RNA: a high resolution coarse-grained energy model for RNA. *J. Phys. Chem. B*, **114**, 11957–11966.
41. Cragolini, T., Derreumaux, P. and Pasquali, S. (2013) Coarse-grained simulations of RNA and DNA duplexes. *J. Phys. Chem. B*, **117**, 8047–8060.
42. Sulc, P., Romano, F., Ouldridge, T.E., Doye, J.P. and Louis, A.A. (2014) A nucleotide-level coarse-grained model of RNA. *J. Chem. Phys.*, **140**, 235102.
43. Boniecki, M.J., Lach, G., Dawson, W.K., Tomala, K., Lukasz, P., Soltysinski, T., Rother, K.M. and Bujnicki, J.M. (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.*, **44**, e63.
44. Magnus, M., Boniecki, M.J., Dawson, W. and Bujnicki, J.M. (2016) SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res.*, **44**, W315–W319.
45. Xia, Z., Bell, D.R., Shi, Y. and Ren, P. (2013) RNA 3D structure prediction by using a coarse-grained model and experimental data. *J. Phys. Chem. B*, **117**, 3135–3144.
46. Bell, D.R., Cheng, S.Y., Salazar, H. and Ren, P. (2017) Capturing RNA folding free energy with coarse-grained molecular dynamics simulations. *Sci. Rep.*, **7**, 45812.
47. Wienecke, A. and Laederach, A. (2022) A novel algorithm for ranking RNA structure candidates. *Biophys. J.*, **121**, 7–10.
48. Tan, Y.L., Feng, C.J., Wang, X., Zhang, W. and Tan, Z.J. (2021) Statistical potentials for 3D structure evaluation: from proteins to RNAs. *Chin. Phys. B*, **30**, 028705.
49. Tanaka, S. and Scheraga, H.A. (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**, 945–950.
50. Miyazawa, S. and Jernigan, R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
51. Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.
52. Thomas, P.D. and Dill, K.A. (1996) Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.*, **257**, 457–469.
53. Melo, F. and Feytmans, E. (1997) Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, **267**, 207–222.
54. Samudrala, R. and Moult, J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, **275**, 895–916.

55. Betancourt, M.R. and Thirumalai, D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.*, **8**, 361–369.
56. Skolnick, J., Kolinski, A. and Ortiz, A. (2000) Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins*, **38**, 3–16.
57. Buchete, N.V., Straub, J.E. and Thirumalai, D. (2004) Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.*, **13**, 862–874.
58. Buchete, N.V., Straub, J.E. and Thirumalai, D. (2004) Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.*, **14**, 225–232.
59. Gromiha, M.M. and Selvaraj, S. (2004) Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.*, **86**, 235–277.
60. Lu, H. and Skolnick, J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, **44**, 223–232.
61. Zhou, H. and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.
62. Shen, M.Y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein*, **15**, 2507–2524.
63. Rykunov, D. and Fiser, A. (2010) New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinf.*, **11**, 128.
64. Zhang, J. and Zhang, Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One*, **5**, e15386.
65. Deng, H., Jia, Y., Wei, Y. and Zhang, Y. (2012) What is the best reference state for designing statistical atomic potentials in protein structure prediction? *Proteins*, **80**, 2311–2322.
66. Huang, S.Y. and Zou, X. (2011) Statistical mechanics-based method to extract atomic distance-dependent potentials from protein structures. *Proteins*, **79**, 2648–2661.
67. Huang, S.Y. and Zou, X. (2008) An iterative knowledge-based scoring function for protein-protein recognition. *Proteins*, **72**, 557–579.
68. Feng, Y. and Huang, S.Y. (2020) ITScore-NL: an iterative knowledge-based scoring function for nucleic acid-ligand interactions. *J. Chem. Inf. Model.*, **60**, 6698–6708.
69. Ma, Z. and Zou, X. (2021) MDock: a suite for molecular inverse docking and target prediction. In: *Protein-Ligand Interactions and Drug Design*. pp. 313–322. Humana, NY.
70. Bernauer, J., Huang, X., Sim, A.Y.L. and Levitt, M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066–1075.
71. Capriotti, E., Norambuena, T., Marti-Renom, M.A. and Melo, F. (2011) All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, **27**, 1086–1093.
72. Wang, J., Zhao, Y., Zhu, C. and Xiao, Y. (2015) 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Res.*, **43**, e63.
73. Zhang, T., Hu, G., Yang, Y., Wang, J. and Zhou, Y. (2019) All-atom knowledge-based potential for RNA structure discrimination based on the distance-scaled finite ideal-gas reference state. *J. Comput. Biol.*, **27**, 856–867.
74. Tan, Y.L., Feng, C.J., Jin, L., Shi, Y.Z., Zhang, W. and Tan, Z.J. (2019) What is the best reference state for building statistical potentials in RNA 3D structure evaluation? *RNA*, **25**, 793–812.
75. Cruz, J.A., Blanchet, M.F., Boniecki, M., Bujnicki, J.M., Chen, S.J., Cao, S., Das, R., Ding, F., Dokholyan, N.V., Flores, S.C. et al. (2012) RNA-puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **18**, 610–625.
76. Miao, Z., Adamiak, R.W., Antczak, M., Boniecki, M.J., Bujnicki, J., Chen, S.J., Cheng, C.Y., Cheng, Y., Chou, F.C., Das, R. et al. (2020) RNA-puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, **26**, 75120–75341.
77. Li, J., Zhu, W., Wang, J., Li, W., Gong, S., Zhang, J. and Wang, W. (2018) RNA3DCNN: local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS Comput. Biol.*, **14**, e1006514.
78. Townshend, R.J.L., Eismann, S., Watkins, A.M., Rangan, R., Karelina, M., Das, R. and Dror, R.O. (2021) Geometric deep learning of RNA structure. *Science*, **373**, 1047–1051.
79. Tan, Y.L., Wang, X., Shi, Y.Z., Zhang, W. and Tan, Z.J. (2022) rsRNASP: a residue-separation-based statistical potential for RNA 3D structure evaluation. *Biophys. J.*, **121**, 142–156.
80. Paliy, M., Melnik, R. and Shapiro, B.A. (2010) Coarse-graining RNA nanostructures for molecular dynamics simulations. *Phys. Biol.*, **24**, 036001.
81. Brion, P. and Westhof, E. (1997) Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 113–137.
82. Leontis, N. and Zirbel, C. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: Leontis, N. and Westhof, E. (eds). *RNA 3D Structure Analysis and Prediction*. Springer Berlin, Heidelberg, Vol. **27**, pp. 281–298.
83. Becquey, L., Angel, E. and Tah, F. (2021) RNANet: an automatically built dual-source dataset integrating homologous sequences and RNA structures. *Bioinformatics*, **37**, 1218–1224.
84. Adamczyk, B., Antczak, M. and Szachniuk, M. (2022) RNAsolo: a repository of cleaned PDB-derived RNA 3D structures. *Bioinformatics*, **38**, 3668–3670.
85. Lu, X.J., Bussemaker, H.J. and Olson, W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.
86. Li, S., Olson, W.K. and Lu, X.J. (2019) Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.*, **47**, W26–W34.
87. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R. and Hofacker, I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.
88. Mao, K., Wang, J. and Xiao, Y. (2020) Prediction of RNA secondary structure with pseudoknots using coupled deep neural networks. *Biophys. Rep.*, **6**, 146–154.
89. Parisien, M., Cruz, J.A., Westhof, E. and Major, F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
90. Kabsch, W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **34**, 827–828.
91. Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
92. Magnus, M., Antczak, M., Zok, T., Wiedemann, J., Lukasiak, P., Cao, Y., Bujnicki, J.M., Westhof, E., Szachniuk, M. and Miao, Z. (2020) RNA-puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Res.*, **48**, 576–588.
93. Xiong, P., Wu, R., Zhan, J. and Zhou, Y. (2021) Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement. *Nat. Commun.*, **12**, 2777.
94. Masso, M. (2018) All-atom four-body knowledge-based statistical potential to distinguish native tertiary RNA structures from nonnative folds. *J. Theor. Biol.*, **453**, 58–67.
95. Carrascoza, F., Antczak, M., Miao, Z., Westhof, E. and Szachniuk, M. (2022) Evaluation of the stereochemical quality of predicted RNA 3D models in the RNA-Puzzles submissions. *RNA*, **28**, 250–262.