

Robust RNA structure refinement by a nucleobase-centric sampling algorithm coupled with a backbone rotameric and quantum-mechanical-energy-scaled base-base knowledge-based potential.

Yaoqi Zhou (✉ yaoqi.zhou@griffith.edu.au)

Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Dr. Southport, QLD 4222 <https://orcid.org/0000-0002-9958-5699>

Peng Xiong

Griffith University

Ruibo Wu

Sun Yat-Sen University

Jian Zhan

Griffith University

Article

Keywords: Molecular Biology, Knowledge-based Energy Function, Base-base Statistical Potentials, Indirect Interactions, Near-native RNA Models

Posted Date: January 6th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-136920/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Robust RNA structure refinement by a nucleobase-centric sampling algorithm coupled with a backbone rotameric and quantum-mechanical-energy-scaled base-base knowledge-based potential.

Peng Xiong^{1,2}, Ruibo Wu³, Jian Zhan^{1§} and Yaoqi Zhou^{1,2§}

¹Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Drive, Southport, QLD 4222, Australia

²Institute for Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China

³School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou 510006, China

§Co-corresponding authors (yaoqi.zhou@griffith.edu.au, j.zhan@griffith.edu.au)

ABSTRACT

Refining modelled structures to approach experimental accuracy is one of the most challenging problems in molecular biology. Despite many years' efforts, the progress in protein or RNA structure refinement has been slow because native structures are often not the global minimum of existing approximate energy scores. Here, we propose a fully knowledge-based energy function that captures the full orientation dependence of base-base, base-oxygen and oxygen-oxygen interactions with the RNA backbone modelled by rotameric states and internal energies. A total of 4,000 quantum-mechanical calculations were performed to reweight base-base statistical potentials for minimizing possible effects of indirect interactions. The resulting BRiQ knowledge-based potential, equipped with a nucleobase-centric sampling algorithm, provides a robust improvement in refining near-native RNA models generated by a wide variety of modelling techniques.

Introduction

Three-dimensional structures of RNAs offer the best clues for their functions and yet only a few thousand structures have been determined ¹, compared to 24 million of non-coding RNA sequences collected in RNACentral (as of October 25 2020) ². This huge gap continues to expand exponentially because sequencing entire transcriptomes is only a fraction of the cost and time required for structure determination of a single RNA by the techniques such as X-ray crystallography, nucleic magnetic resonance, and cryo-electron microscopy.

One possible solution to this growing problem is to predict RNA structures from their sequences by computational template-based homology modelling and *de novo* folding^{3,4}. Homology modelling such as RNABuilder⁵ and ModeRNA⁶ attempts to predict the structure of a target RNA by mapping the query sequence onto the template structure(s) of its homologs. However, most RNAs do not have their respective homologous templates. Thus, method development was mainly devoted to predicting structures by *de novo* folding from sequences through fragment or motif assembling. Examples are FARNA⁷, MC-fold/MC-Sym⁸, Rosetta-SWM⁹, ifoldRNA^{10,11}, Vfold^{12,13}, 3dRNA^{14,15}, and SimRNA¹⁶. Evaluation of these methods by RNAPuzzles¹⁷⁻¹⁹ indicates that reasonable predictions are possible only for RNAs with either simple topology or existing homologous templates. Moreover, reasonable

near-native predictions were often ranked poorly by the predictors. Clearly both homology and *de novo* models would greatly benefit from a structure-refinement technique that can bring the models closer to native structures locally and globally and improve near-native ranking.

Structure refinement, however, is challenging with few methods available for RNAs. Even for protein structure refinement, despite a long history of method development, only a moderate local improvement can be made by combining molecular mechanics force fields and knowledge-based potentials with pre-defined restraints to prevent large deviation away from the native basin^{20,21}. For RNAs, most all-atom refinements were performed after coarse-grained sampling for removing steric clashes and non-ideal bond lengths, bond angles, or torsional angles. For example, the QRNAS refinement²² utilizes a modified AMBER force field with enforced base-pair planarity, explicit hydrogen bonds, and backbone regularization. FARFAR^{23,24} improves model accuracy through all-atom refinement after FARNA coarse-grained sampling. It employed a Rosetta all-atom force field that mixes physical and knowledge-based scores with RNA-specific terms for Watson-Crick base pairing, base stacking, and torsional potentials²⁵.

Here, we build a knowledge-centric refinement energy score. Atomic-level knowledge-based energy functions, derived from known three-dimensional structures, have traditionally focused on distance dependence in protein structures²⁶ with similar statistical potentials developed for RNA²⁷⁻²⁹. Unlike these atomic knowledge-based scores and previous all-atom RNA refinement energy scores^{22,23}, the statistical potential in this work is tailored specifically to RNA interactions that were dominant by orientation-dependent base-pairing and stacking interactions³⁰ with rotameric backbone³¹, in contrast to dominant distance-dependent hydrophobic interactions³² with rotameric sidechains³³ in protein folding. Base-base interactions were described by fine grids in a six-dimensional orientational space and scaled by quantum mechanical calculations allowing better capture of both local and global interactions. This Backbone Rotameric and Quantum-mechanical-energy-scaled base-base knowledge-based potential (BRiQ) is integrated with a new nucleobase-centric tree algorithm that samples backbone conformations around predicted or known base pairs. The resulting refinement technique can consistently improve model structures for the majority of near-native structural models as demonstrated by refining Rosetta-SWM motif³⁴, RNA puzzles¹⁷⁻¹⁹, and FARFAR2²⁴ models.

Results

BRiQ refinement energy score. The BRiQ refinement energy score is designed to capture stably stacked bases linked by a more flexible ribose and phosphate backbone. In particular, the interactions associated with bases and oxygen atoms in backbones are strongly orientation-dependent. To capture this unique structural property of RNA chains, the energy score (E) is separated into six terms: orientation-dependent interactions between two bases (E_{bb}), between a base and a main-chain oxygen atom (E_{bo}), and between two main-chain oxygen atoms (E_{oo}), backbone rotameric energy (E_{rot}), internal energy ($E_{internal}$), and atomic clash energy (E_{clash}). That is,

$$E = E_{bb} + E_{bo} + E_{oo} + E_{rot} + E_{internal} + E_{clash} \quad (1)$$

where we employed five oxygen atom types including OP for OP1 and OP2 and O2', O3', O4', and O5' in ribose. To capture the orientation dependence, the whole nucleobase (A, C, G, and U) is treated as a single rigid group with a local coordinate system and the relative position between two bases is described by the distance vectors and their orientations (Figure 1A). The densities in the six-dimensional space were derived from known RNA structures using kernel density estimations, rather than histogram statistics (See Methods). This allows a smoother energy landscape (Figure 1B) and the near-native local minima are closer to the corresponding native structure in a fine orientational space. To remove the impact of indirect base-base interactions, statistical energy scores are scaled by quantum mechanical calculations according to the correlation coefficient between representative hydrogen-bonded base pairs (Figures 1B and 1C). Similarly, kernel density estimations were also employed to obtain base-oxygen (Figure 1D) and oxygen-oxygen interactions whereas statistical rotameric states were obtained for ribose backbones (Figure 1E). Empirical internal bond angle, bond length, and atom-crash energies were also developed to ensure appropriate bond geometry and packing density (Methods).

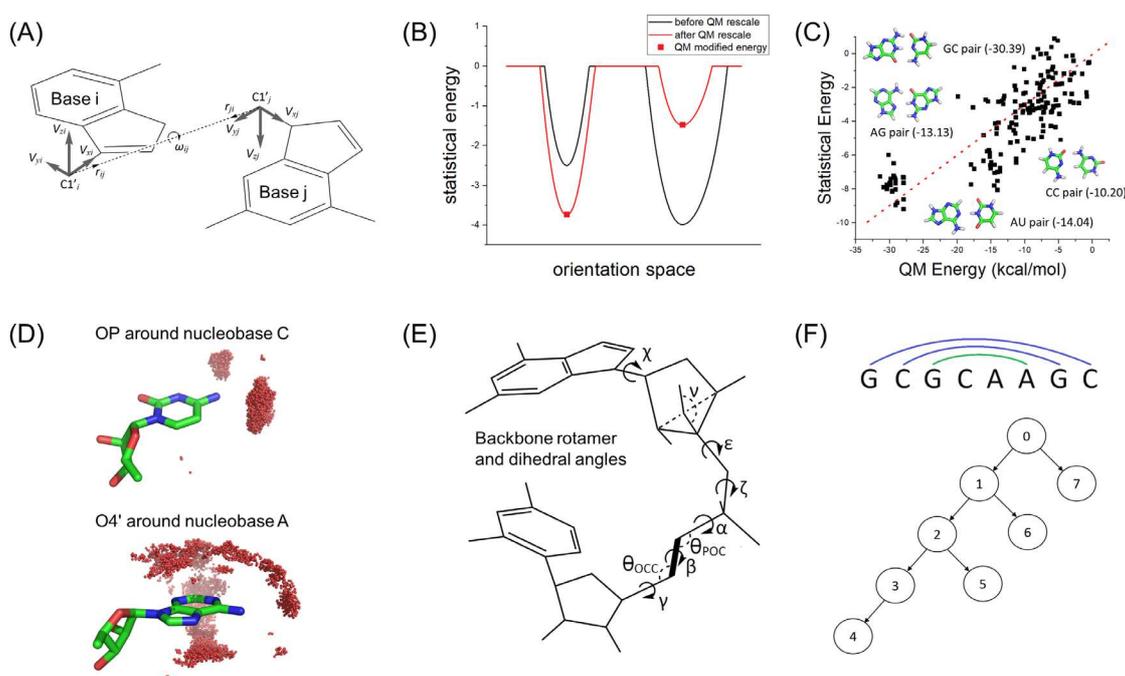


Figure 1. (A) Six-dimensional base-base statistical potentials with relative positions defined by r_{ij} and r_{ji} vectors along with the rotational angle ω_{ij} . (B) A schematic illustration of the orientation dependence before and after Quantum Mechanical (QM) energy scaling. (C) This QM scaling is based on the correlation between the statistical energy scores of hydrogen-bonded base pairs ($-\ln P_{hb}$) and QM calculations. P_{hb} is the probability of hydrogen-bonded base pairs. (D) The distribution of OP around nucleobase C and the distribution of O4' around the nucleobase A as labelled. (E) Backbone rotamers defined according to various torsion and improper angles that control the ribose (χ, ν) and phosphate (ϵ, ζ) backbone. Dihedral angles α, β, γ , bond angles $\theta_{POC}, \theta_{OCC}$ and bond length of C5'O5' were required to

calculate the internal energy. (F). Nucleobase-centric fold-tree (NuTree) algorithm for refinement by defining bases as the nodes and locally or globally connected bases as the edges, illustrated by the GCAA tetraloop.

NuTree sampling algorithm. The above nucleobase-centric energy score is coupled with a nucleobase-centric tree (NuTree) algorithm for conformational sampling, in which each node is a base and each edge represents the relative position between two bases (Figure 1F). Two connected bases in the folding tree could be sequential neighbours or hydrogen-bonded base pairs. Instead of sampling in backbone-dihedral angle space, conformations are sampled according to the pre-defined orientation space of each edge type because the orientations around more stable bases are the determinants of backbone orientations. For the same reason, coordinates of sidechains were built prior to those of the backbones. Only local moves were allowed for each node whereas both local and global moves were allowed for each edge, making both local and global structural improvements possible. In the process of Monte Carlo sampling, we only need to calculate the energy change upon a random structure move, without knowing the first or second derivative of the energy function. Therefore, this sampling algorithm could be used for optimizing an RNA structure with a non-differentiable energy function in atomic accuracy.

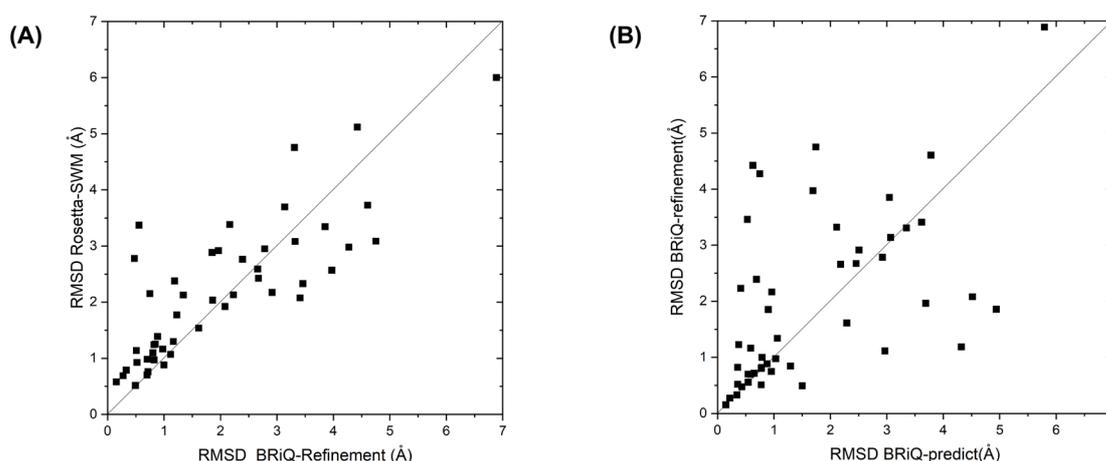


Figure 2. (A) Refining Rosetta-SWM motif models by BRiQ improves the majority of those model structures with $\text{RMSD} < 2\text{\AA}$ as demonstrated by RMSD comparison (Lowest RMSD of top 1% before refinement on Y-axis versus after refinement on X-axis). (B) Fixing all native base pairs with random initial conformations for all other regions and then folding the remaining structure leads to more accurate motif models for the majority of the motifs than refining Rosetta-SWM models that have pre-assigned, partially fixed base pairs (Lowest RMSD of top 1% BRiQ-refined models on Y-axis versus lowest RMSD of top 1% BRiQ models with fixed native base pairs on X-axis).

Refining Rosetta-SWM conformations for RNA motifs: We first test our method using a benchmark of 48 motifs. This benchmark³⁴ was developed to examine the ability of Rosetta-SWM to recover the loop region given some of the base pairs and environment contacts fixed (See Methods and Supplementary Table S1). We test our refinement method by refining all 88,352 conformations generated by Rosetta-SWM with the NuTree algorithm and the BRiQ energy score. The NuTree was constructed with the base-pairing information extracted from Rosetta-SWM model, and the BRiQ energy was optimized by Monte Carlo sampling at low

temperature. The lowest RMSD values of top 1% predicted models by the Rosetta and BRiQ energies are compared in Figure 2A. The majority of predicted motif conformations (31/48, 64.5%) have an improved RMSD after BRiQ refinement with a median reduction of 0.2Å RMSD change among all 48 motifs. The improvement is more impressive on the refinement from near-native models (RMSD of Rosetta-SWM model less than 2Å), 81% (17/21) predicted conformations have an improved RMSD after BRiQ refinement. The RMSDs of the top 1% predicted models for the remaining four motifs are essentially unchanged.

Generate motif structure with base-pairing information: As the NuTree sampling algorithm is a refinement protocol, near-native conformations will be difficult to obtain if initial conformations contain false base pairs generated by other methods. Here, we further test whether or not near-native structures can be generated from the NuTree algorithm if all motif base pairs are known. Starting from a random backbone conformation with base pairs fixed, we run Monte Carlo sampling at different temperatures (Methods) to optimize the BRiQ energy score. The lowest RMSD of top 1% predicted models by the BRiQ score are compared in Figure 2B. The majority of predicted conformations (33/48, 68.8%) have an improved RMSD with correct and complete base-pairing information, compared to partial fixation by Rosetta-SWM. The median improvement in the best RMSD value within top 1% predicted models is 0.29Å.

Four representative motifs were chosen for illustration in Figure 3. They are a standard helix made of GC pairs (Figure 3A), two tetraloops (GCAA_tetra loop, Figure 3B, UUCG_tetra loop, Figure 3C) commonly used for testing molecular mechanics force fields³⁵, and an example of refinement leading to worse structures (j55a_P4P6_fixed, Figure 3D). In these figures, the top to bottom panels represents the results from Rosetta-SWM, refinement of Rosetta-SWM conformations by BRiQ and the structure prediction by BRiQ with native base-pairing information. For the GC helix (Figure 3A), minimum RMSD sampled by Rosetta-SWM is 0.58Å, compared to 0.14Å by BRiQ. For GCAA_tetra loop (Figure 3B), near-native local minimal improved from around 1.5Å RMSD by Rosetta-SWM to about 1Å by BRiQ. UUCG tetraloop (Figure 3C) is an example that <1Å near-native structure is the global minimum in BRiQ, whereas the lowest energy conformation in Rosetta-SWM is around 3.5 Å. For all three motifs, BRiQ refinement of Rosetta-SWM models leads to an improved backbone fitness to the native backbone. Motif j55a_P4P6_fixed (Figure 3D) is an example of worse RMSD values after refining Rosetta-SWM conformations. We noticed that this native structure has one base protruded into the solvent, which may be the reason for the failure of BRiQ as the solvation effect is only implicitly accounted for in a statistical potential. We further found that if native conformations are directly refined by BRiQ, the lowest energy conformations of the most motifs (45/48, 94%) remain <2.5 Å away from the native structure. Refining models from the native structures and the Rosetta-SWM models finds the same minimum for most motifs with similar energy. As a result, most motifs have high-quality near-native conformations as the global minimum.

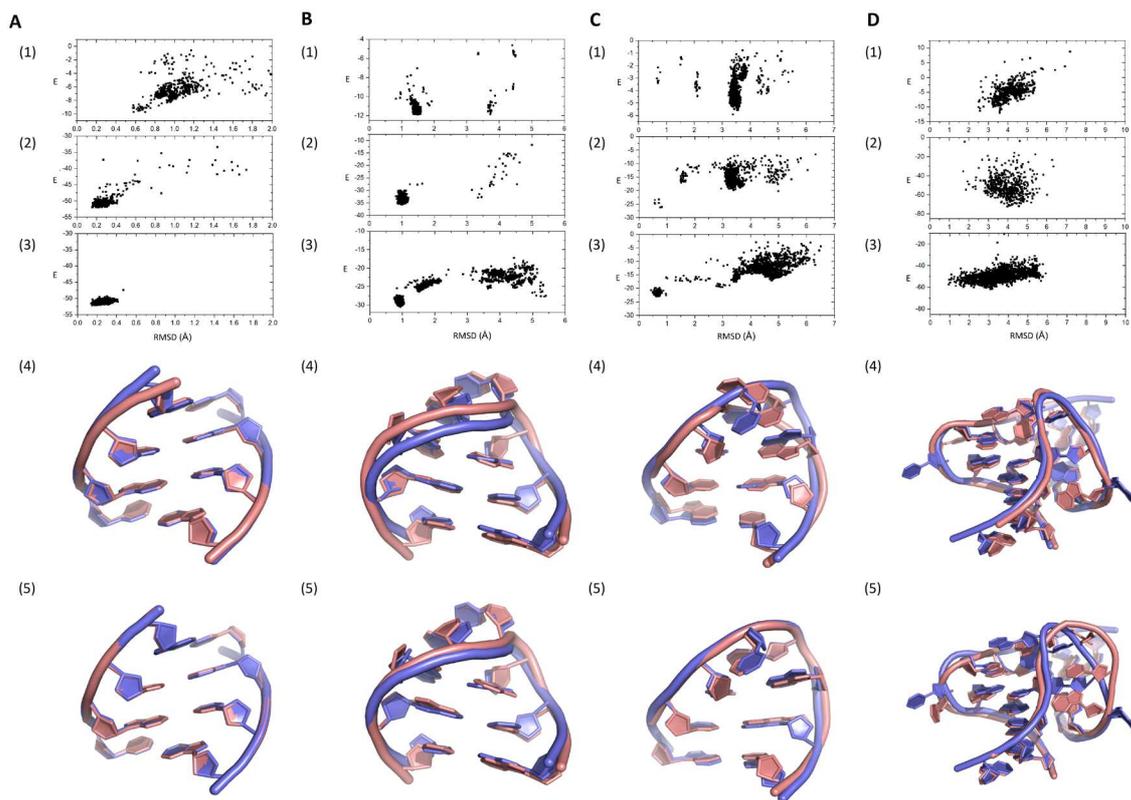


Figure 3. Energy versus RMSD values of the conformations sampled by (1) Rosetta-SWM, (2) refinement of Rosetta-SWM conformations by BRiQ, (3) structure prediction by BRiQ with all native base pairs fixed for (A) CG-helix, (B) GCAA_tetra loop, (C) UUCG_tetra loop, and (D) j55a_P4P6_fixed. Structure alignment of native (blue) to the best in top 1% predicted models by Rosetta-SWM (panel 4) and BRiQ refinement (panel 5) in the bottom panel.

RNA puzzle refinement. A real-world test of any refinement techniques is to refine previously predicted models. Here we refined all submitted models in 24 RNA puzzle experiments (PZ1-PZ25, Supplementary Table S2). For each submitted model, we generated 20 refinement models. Then, the lowest energy model within 20 refinement models is treated as the BRiQ predicted model. The lowest RMSD model from all BRiQ predicted models for a given RNA puzzle is compared to the lowest RMSD value among all submitted models in Figure 4A. As the figure shown, if initial models contain models with $\text{RMSD} < 4\text{\AA}$, refinement always leads to improvement, with the largest RMSD reduction of 1.5\AA . Figure 4B further compares the lowest RMSD value among all submitted models to those lowest RMSD values among all BRiQ refined models. Essentially all RNAs have better models sampled after refinement with the largest RMSD reduction at 2.2\AA .

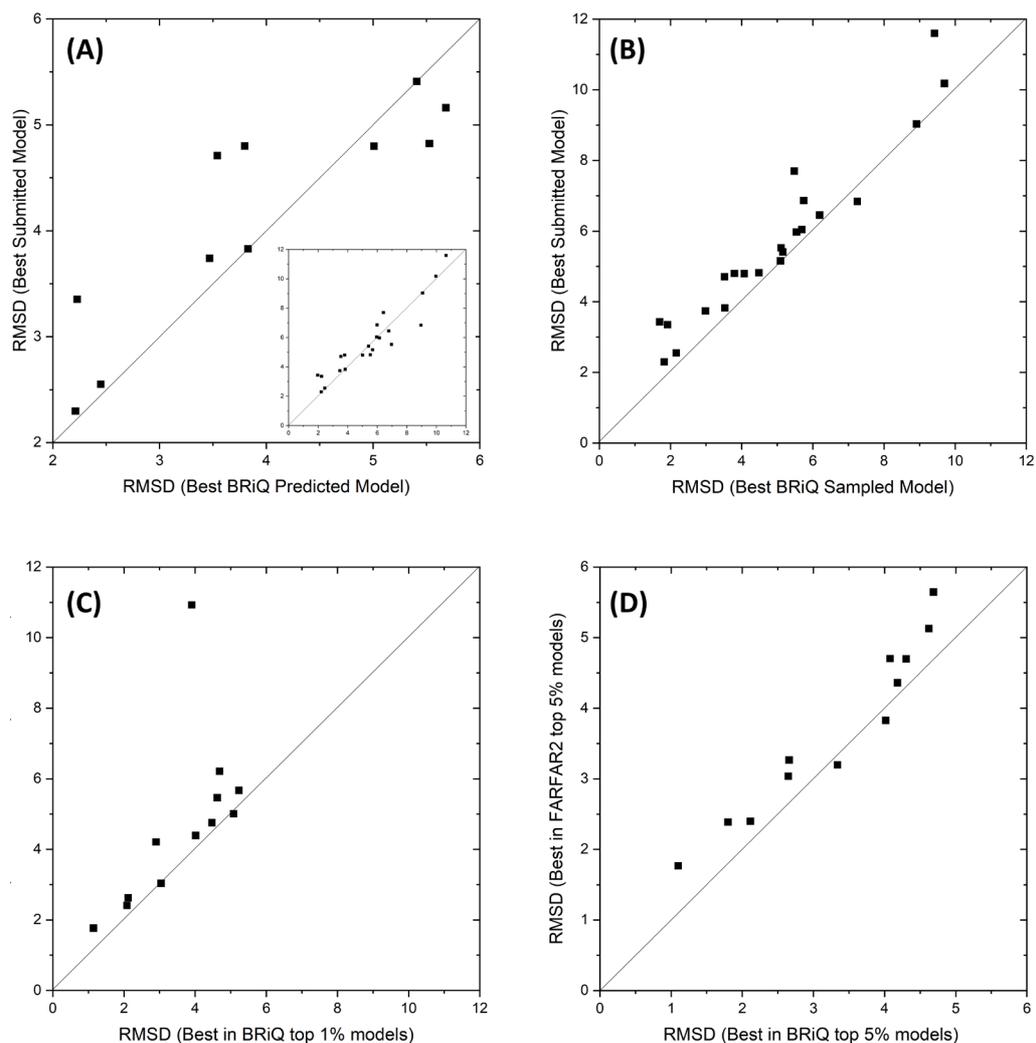


Figure 4. BRiQ refinement achieves consistent improvement over the best-submitted models for RNA puzzles containing model structures with $\text{RMSD} < 4\text{\AA}$ (A), over all RNA puzzles in conformational sampling (B), over FARFAR2 models for 10 out of 12 RNA puzzles in top 1% (C) and top 5% (D) predicted models.

Comparison to FARFAR2. We also test our refinement method starting from the recently developed FARFAR2 predicted models²⁴ but only for those 12 RNA puzzles with RMSD of best FARFAR2-predicted models less than 6\AA as shown in Figure 4A. This is because it is not possible to refine those models that are far from native. For each puzzle, we select 2000 lowest energy models according to FARFAR2 (1547 models for PZ20) to run BRiQ refinement, and compare the best RMSD values of top 1% and 5% predicted models by FARFAR2 and BRiQ energy scores, respectively. As shown in Supplementary Table S3 and Figures 4C and 4D, almost all (10/12) have reduced RMSD values after refinement with a median RMSD reduction of 0.48\AA for top 1% prediction and 0.45\AA for top 5% prediction. The only two RNA puzzles with increased RMSD values after refinement have very small changes ($<0.08\text{\AA}$ for top 1% and $<0.19\text{\AA}$ for top 5% predictions). This further indicates the consistency of BRiQ refinement.

Figure 5 compare the conformations sampled by FARFAR2 and those further refined by BRiQ for Puzzles 4 and 18, respectively. Both show significant moves toward native structures of RNAs. Similar to the motif cases, BRiQ refinement yields improved backbone conformations over FARFAR2 models.

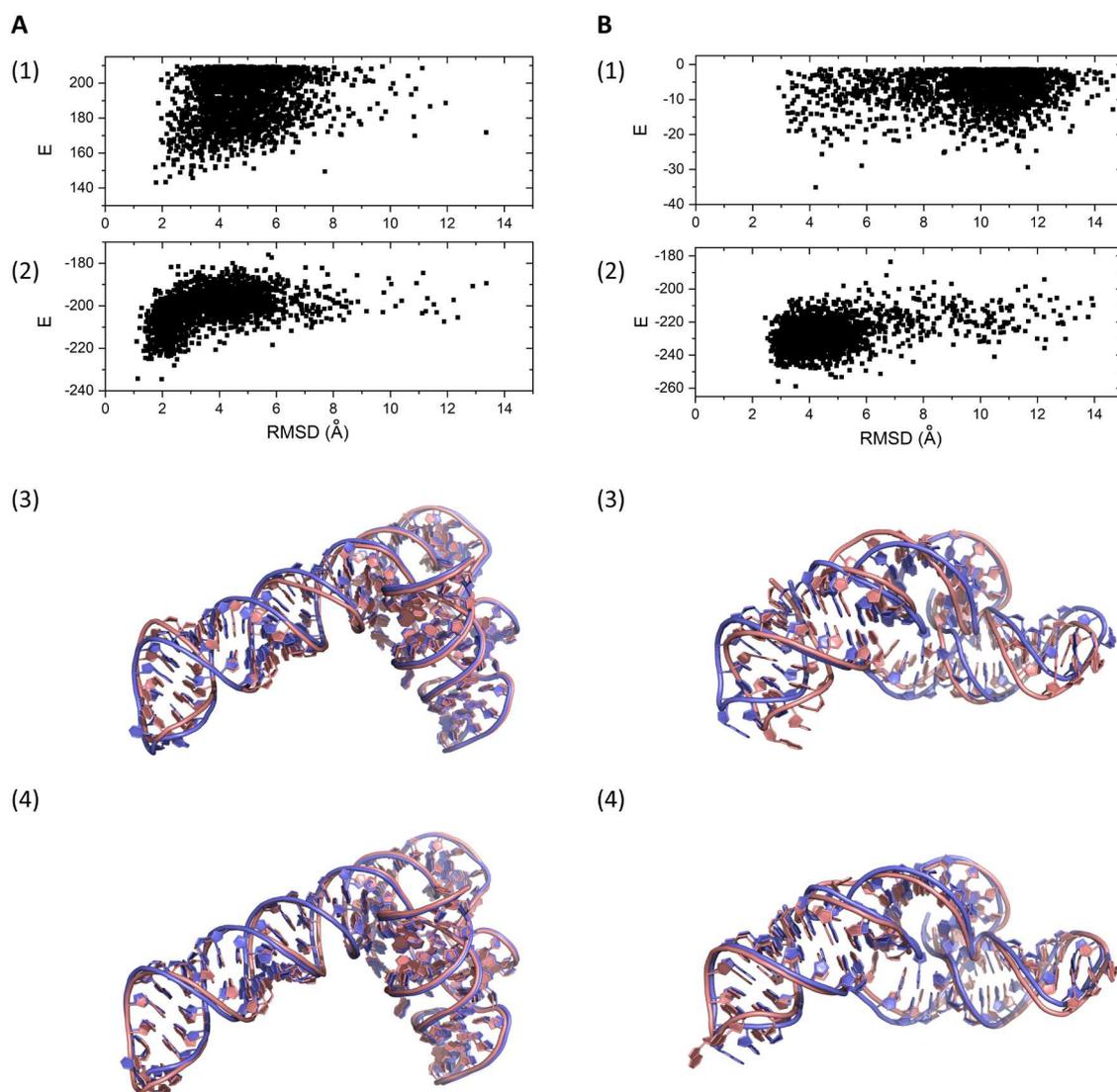


Figure 5. Energy versus RMSD values of the conformations sampled by (1) FARFAR2, (2) refinement of FARFAR2 conformations by BRiQ for (A) RNA Puzzle 4 and (B) RNA Puzzle 18. Structure alignment of native (blue) to the best in top 1% predicted models by FARFAR2 (3) and BRiQ refinement (4) in the bottom panel.

Discussion

This work presents a protocol for refining RNA model structures produced by other modelling tools. The protocol is nucleobase-centric in its knowledge-based energy score as well as the sampling algorithm. The fully knowledge-based score (BRiQ) captures the orientation dependence of base-base and base-oxygen, and oxygen-oxygen interactions by using a local coordinate system for each nucleobase, whereas the structures of ribose and phosphate backbones was governed by rotameric statistical potentials and empirical internal energies. Moreover, the dominant base-base interaction was scaled by quantum mechanical calculations for removing possible indirect interactions contained in structural statistics. This statistical potential was coupled with the NuTree algorithm that samples the conformations around pre-determined base pairs. This refinement protocol improves 81% Rosetta-SWM models with less than 2Å RMSD, 100% RNA puzzle models with RMSD<4Å and 83% FARFAR2 models with RMSD<6Å. The CPU time cost of refining a single RNA puzzle model is around 40 minutes for a model of 50 nt, and 2 hours for a model of 100 nt on Inter Xeon Gold 6242 CPU 2.80GHz.

In order to secure as many RNA structures as possible, we included all high-resolution RNA structures to collect base and backbone statistics for our BRiQ knowledge-based score. This is because a few redundancies will have an insignificant effect on all statistics collected at the atom or base level. Indeed, we found that the difference in the energy scores of a single base pair before and after excluding all homologous structures of RNA puzzles based on CD-HIT with 70% sequence identity (the sequence similarity for the RNA "twilight zone" is 80%)³⁶ is mostly negligible (the difference is <0.001 for most cases). Indeed, the difference before and after removing homologs for the refined FARFAR2 structures is only 0.02Å for rp04 and 0.2Å for rp18 for the best in top 1%, respectively.

The NuTree algorithm samples the RNA conformations around predetermined base pairs from a given model. These preformed base pairs serve as the nucleus to speed up the folding of the rest of an RNA chain. However, if a base pair was incorrectly modelled, it will be energetically difficult to correct the mistake. This is the main reason why the current refinement protocol works best for those near-native structures, in which the majority of base pairs were correctly identified. Unlike Rosetta, the NuTree algorithm can fix both nested and non-nested (pseudoknot) base pairs. In other words, this algorithm will become more useful as secondary and tertiary base pairs are increasingly more accurately predicted by employing co-variational analysis of RNA homologous sequences³⁷⁻⁴⁰ and deep learning techniques⁴¹.

The BRiQ statistical energy score is built on the knowledge that backbone conformations are determined by more stable stacked base pairs. This was done by focusing on base-base, base-oxygen, and oxygen-oxygen interactions while treating the backbone as rotameric states around the bases. The robustness of the scoring function is illustrated by the consistent improvement in refining a diverse set of RNA models produced from different methods participated in RNA puzzles including ModeRNA⁶, FARNNA⁷, MC-Fold/MC-Sym⁸, ifoldRNA^{10,11}, Vfold^{12,13}, 3dRNA^{14,15}, and SimRNA¹⁶.

This work is limited to RNA structural refinement. Another question is whether or not the RBiQ statistical energy function coupling with the NuTree algorithm can serve as an effective method for *ab initio* structure prediction. Initial studies suggest that a new sampling algorithm is likely required because *ab initio* folding requires more frequent breaking and

forming of base pairs than what are typically involved in the NuTree algorithm. The work in this area is in progress.

Methods

BRIQ Energy Score

(1) Base-base interaction energy (E_{bb})

Relative orientation between two bases. The planar shaped bases make the orientation dependence a must for base-base interactions. Here, we consider each base i as a rigid body with a local coordinate system CS_i whose origin is located at atom C1'. The x-axis is in the direction of C1'N9 and the z-axis is in the direction of C1'N9×N9C4 for A and G whereas the x-axis is in the direction of C1'N1 and the z-axis is in the direction of C1'N1×N1C2 for U and C. As shown in Figure 1A, the orientation and distance between bases i and j can be described by 1) the distance r_{ij} between the origins of CS_i and CS_j , 2) the rotational angle ω_{ij} around r_{ij} , 3) the directional vector r_{ij} in CS_i and 4) the directional vector r_{ji} in CS_j (a total of 6 dimensions). The distance r_{ij} has a range of 0 to 15 Å with a uniform grid space of 0.3 Å. The rotational angle ω_{ij} varied from -180° to 180° with a uniform grid of 8° . We represent the orientation of the directional vector r_{ji} by 2000 uniformly distributed points on a sphere from Monte Carlo simulated annealing of points with repulsive interactions proportional to the inverse of squared distances between the points. Thus, the space between the two bases is separated into $50 \times 45 \times 2000 \times 2000$ discrete regions. Once the energy values for all these regions are known, the energy value at a given distance and orientation can be linearly interpolated. The above coordinate system for representing the base-base relative orientation was similar to but is more sophisticated than what was proposed for a coarse-grained two-particle representation of RNA chain⁴². The 6-dimensional base-base interactions were pre-calculated and stored in a table and only negative values were loaded into computer memory for quick access.

Base-orientation representation. To define an orientation-dependent density, it is necessary to measure the structural similarity first. One common measure is the root-mean-squared distance (RMSD) between two base pairs. However, it is too time-consuming to calculate RMSD. To speed up the calculation, we assumed that four fixed points (T1, T2, T3, T4) in the local coordinate system of a base (Supplementary Figure S1a) are sufficient to represent the orientation of each base and the mean squared distance between these points in two bases

A and B ($DDM = \sqrt{\sum_{i=1}^4 \sum_{j=1}^4 d(T_{Ai}, T_{Bj})^2 / 16}$) can approximate RMSD. We have

optimized T1, T2, T3, T4 so that DDM has the highest correlation coefficient to RMSD. The final four points in each base coordinate system are: T1 = (2.158, 3.826, 1.427), T2 = (-0.789, -0.329, -1.273), T3 = (4.520, -3.006, 1.586) and T4 = (6.018, 1.903, -1.638). The resulting correlation coefficient is 0.974 (Supplementary Figure S1b).

Orientation distribution density. We employed a modified radial basis function kernel $h(d)$ to calculate the orientation distribution of one base around another base $f(x)$ by the following equations:

$$h(d) = \begin{cases} 1, & \text{if } d \leq 0.15 \\ e^{-0.5((d-0.15)/0.1)^2}, & \text{if } (d > 0.15) \end{cases} \quad (2)$$

and

$$f(x) = \sum_{i=1}^N h(DDM(x, x_i)) \quad (3)$$

where N is the number of base pairs in the database of RNA structures and $DDM(x, x_i)$ is the distance between a target base pair and a base pair in the database. If $DDM(x, x_i)$ between two bases is less than 0.15\AA , we consider two bases shared the same orientation. Here and below, we have employed an empirical value (0.1) for defining the spread of the kernel based on statistics from the PDB structures and a few trials. An example is shown in Supplementary Figure S2.

The orientation distribution densities of base pairs with different sequence-separation distances (1, 2 or >2) were calculated separately. For non-local base pairs (separation >2), the density was re-weighted by quantum mechanical energy.

Quantum-mechanical-energy-weighted orientation distribution density. One issue associated with a statistical energy function is that a high population of specific orientation between two bases may not be due to a strong direct interaction at this specific orientation but due to the orientation preference of the interaction with another base. Here, we minimize the possibility of this type of indirect interactions by a weighting factor $w(x_i)$ according to quantum calculations. The assumption is that the strength of the directional interaction is related to the strength of the quantum interaction energy. Here,

$$w(x_i) = e^{-\frac{E_{QM}(\tilde{x}_i)}{4.32\text{kcal/mol}}} / f(\tilde{x}_i), \quad (4)$$

where \tilde{x}_i is the nearest representative configuration to x_i , $E_{QM}(\tilde{x}_i)$ is the quantum energy. The modified orientation distribution $f'(x)$ satisfies

$$f'(x) = \sum_{i=1}^N w(x_i) h(DDM(x, x_i)) \quad (5)$$

Here, quantum mechanical (QM) calculations were performed by using Gaussian 09⁴³. For each base pair, we selected 80 representative configurations based on the conformations obtained from the PDB structures. For each representative configuration, we employed 5 nearest conformations in the PDB structures as the initial configurations for QM calculations. The steps for calculating QM base-base interactions are as follows: 1) remove all backbone atoms, 2) replace the C1' atom with a H atom, 3) optimize the position of the H atom by *ab initio* Hartree-Fock calculations with the basis set 6-31G* and 4) calculate the base-base interaction energy by $E(AB) = H(AB) - H(A) - H(B)$ with the density-functional theory method M06-2X⁴⁴ and the basis set 6-31+G(d,p). The average value of five QM calculations is considered as the QM energy for the representative configuration. A total of 4,000 QM calculations for 10 possible base pairs ($80 \times 10 \times 5$) were performed. Figure 1C shows some examples of the calculation results. The scaling parameter (4.32kcal/mol) in Eq. (4) between QM calculations and statistical energy functions is obtained from the slope of the regression analysis of hydrogen-bonded base pairs (Figure 1C).

Base-base interaction energy. The final energy function for base-base interactions is calculated by the following equations:

$$E_0(x, sep) = -\ln f'(x) / f_{ref}(sep) \quad (6)$$

and,

$$E_{bb}(x, sep) = \begin{cases} E_0(x, sep), & \text{if } E_0(x, sep) < 0, \\ 0, & \text{if } E_0(x, sep) \geq 0. \end{cases} \quad (7)$$

Where sep is the sequence separation distance, its value could be 1, 2 or 2+. $f_{ref}(2+)$ was employed to scale the lowest energy value of non-local base pair to -8.0 based on the lowest energy from QM calculations and the scaling between statistical and QM calculations (Figure 1C). $f_{ref}(1)$ and $f_{ref}(2)$ were employed to scale the lowest energy value of local base pair to -4.0. Here, we also set the maximum base-base interaction energy to zero because repulsive interactions from the above six-dimensional statistical analysis are not reliable. Figure 1B shows a schematic example of before and after QM scaling.

(2) Interaction energy between a base and a main-chain oxygen atom $E_{bo}(x)$

The interaction energy between a base and a main-chain oxygen atom is the interaction between a polar group and a polar atom. The orientation of a base is still represented by 4 points obtained previously. The relative orientation and position between an oxygen atom and a base can be described by the distance between the oxygen atom and the four points representing the base. The kernel function $h(d)$ and the density $f(x)$ for $E_{bo}(x)$ are given by

$$h(d) = e^{-0.5(d/0.16)^2} \quad (8)$$

and

$$f(x) = \sum_{i=1}^N h(d(x, x_i)) \quad (9)$$

where x and x_i represent the target base-oxygen and a base-oxygen pair in the database, respectively, and $d(x, x_i)$ are the RMSD between the two base-oxygen pairs. Unlike E_{bb} , the repulsive (positive) interaction was not set to zero but reduced by a decaying weighting function $w_{bo}(d_{bo}^{min}) = 1 - 1/(1 + e^{(3.7 - d_{bo}^{min})/0.08})$, where d_{bo}^{min} is the minimum distance between an oxygen atom and any atoms in a base, 3.7Å is the approximate interaction distance between a mainchain oxygen atom and a base and 0.08 is an empirical parameter to control the rate of decay. We also impose an orientation dependence for the hydrogen bonding between an oxygen atom and the corresponding atoms in a base according to the θ angle between the base atom N or O, the main chain atom O (O2', OP1, OP2) and the main chain atom C or P bonding to O. This was done by locating the angle range in RNA structures after removing top 3% angles from each side and defining $s(\theta)=0$ for outside the angle range, $s(\theta)=1$ when $\theta=\theta_M$, the median value of θ , and

$$s(\theta) = \begin{cases} 1 - \left(\frac{\theta - \theta_M}{\theta_L - \theta_M} \right)^2, & \text{if } \theta_L < \theta < \theta_M \\ 1 - \left(\frac{\theta - \theta_M}{\theta_U - \theta_M} \right)^2, & \text{if } \theta_M < \theta < \theta_U \end{cases}$$

where θ_u and θ_L are the upper and lower bounds for θ , respectively. The final expression for $E_{bo}(x)$ is

$$E_{bo}(x) = -w_{bo}(d_{bo}^{min})s(\theta) \ln f(x)/f_{ref} \quad (10)$$

where f_{ref} is employed to scale the lowest energy value empirically to -3.0, which we set to be about one-third strength of GC pairs. As an example, Figure 1D shows the distribution of a nucleobase C and an oxygen atom in a phosphate group and that of a nucleobase A and O4 in ribose.

(3) Hydrogen-bonding energy between main-chain oxygen atoms $E_{oo}(x)$

We consider hydrogen bonding interactions between O2'-O2', O2'-OP and OP-OP. The configuration of hydrogen bonds is determined by four atoms (two oxygen atoms and their connecting C or P atoms). The distance d between two hydrogen-bonding configurations is calculated by RMSD between the four atoms. Similar to $E_{bo}(x)$, we have

$$E_{oo}(x) = -w_{oo}(d_{oo}) \ln \frac{f(x)}{f_{ref}} \quad (11)$$

with $f(x) = \sum_{i=1}^N e^{-0.5(d(hb, hb_i)/0.1)^2}$, f_{ref} is employed to scale the lowest energy value empirically to -3.0 for O2'-O2', -2.0 for O2'-OP, -1.5 for OP-OP and the decay function $w_{oo}(d_{oo}) = 1 - 1/(1 + e^{(3.3-d)/0.07})$ with 3.3Å is the typical hydrogen bond length between two oxygen atoms and 0.07 is the empirical parameter to control the decay rate. Here the angle dependence is implicitly accounted for by using 4 atomic coordinates (e.g. atoms C2'-O2'-OP-P for the hydrogen bond between O2' and OP) to define RMSD.

(4) Energy for Atomic clashes (E_{clash})

Due to lack of adequate statistics for hard-core exclusion in all statistical energy functions, we have set repulsive terms to 0 in $E_{bb}(x)$ or quickly decay to 0 by weighting functions $w_{bo}(d_{bo}^{min})$ and $w_{oo}(d_{oo})$ in $E_{bo}(x)$ and $E_{oo}(x)$, respectively, when the distance is less than a preset value. To avoid direct atomic clashes, we introduce an empirical energy function as below.

$$E_{clash}(d) = \begin{cases} (k_{clash} * 0.4)^4 - 4(k_{clash} * 0.4)^3(d - r_0 + 0.4), & \text{if } d \leq r_0 - 0.4, \\ (k_{clash} * 0.4)^4, & \text{if } r_0 - 0.4 < d < r_0, \\ 0, & \text{if } d > r_0, \end{cases} \quad (12)$$

where d is the atomic distance between two atoms, r_0 is the shortest statistical distance from RNA structures between the two atoms and k_{clash} is an empirical parameter to control the increasing rate of repulsion when two atoms approach to each other. We set $k_{clash} = 3$ after the energy minimized structure does not change much for k_{clash} between 2 and 5. r_0 is orientation independent for atoms in SP³ hybridization but is orientation-dependent for atoms in SP² hybridization. That is, r_0 is orientation-dependent between any atom with an atom in a base but is orientation-independent between two main-chain atoms. For orientation-independent r_0 , it is set as the top 5% shortest distance between two atoms found in RNA structures. For orientation-dependent r_0 between a given atom a and an atom b in base B, the atomic position of a is transformed to the local coordinate system of the base B and clustered around the atom b . The top 5% shortest distance in different orientations is set as r_0 in different orientations. The functional form of the clash energy is shown in Supplementary Figure S3.

(5) Main-chain rotameric energy (E_{rot})

For protein structure prediction, the sidechain conformational space is discretized into rotamers. Here, we introduce main-chain rotameric states because stacked bases are more rigid than the main chains. The conformational space of a ribose rotamer is mainly determined by two dihedral angles (Figure 1E). The first is the rotational angle χ rotating around the chemical bond connecting the base and ribose (N9-C1' for bases A and G and N1-C1' for bases U and C). The second is the improper angle ν between the C2'-C4'-O4' plane and the C2'-C4'-C3' plane within the connected ribose. Each dihedral angle can be separated into two regions, and each region into 300 to 500 rotamers, and a total of 1500 rotamers is employed to describe a ribose rotameric state.

The ribose rotamers are clustered according to RMSD. This is done by transforming atomic coordinates of eight ribose atoms to the local coordinate system for the base. RMSD is the average distance of ribose atoms in the local coordinate system. Once RMSD is known, we can use the kernel density estimation method to calculate the density of each rotamer, and then take the negative logarithm and convert it to the rotamer energy $E_{rot}(x)$ as below.

$$E_{rot}(x) = -\ln \frac{f(x)}{f_{max}} \quad (13)$$

where $f(x) = \sum_{i=1}^N e^{-0.5(d(x,x_i)/0.15)^2}$, x_i denotes a rotamer in the database and the summation is over all the rotamers in the database.

(6) The internal energy $E_{internal}$

We introduce the following internal energies ($E_{internal}$) between a phosphate group and two connecting riboses to account for energetics associated to changes in bond lengths E_{bond} , bond angles E_{angle} , and dihedral angles $E_{torsion}$

$$E_{bond}(u) = \begin{cases} -2u - 1, & \text{if } u < -1, \\ u * u, & \text{if } -1 \leq u \leq 1, \\ 2u - 1, & \text{if } u > 1 \end{cases} \quad (14)$$

where $u(bl) = k_{bond}(bl - 1.422)$ and bl denotes the bond length between O5'_{i+1} and C5'_{i+1}. We made this bond slightly flexible so that we can add a phosphate group (atom P, OP1, OP2, O5') between two fixed riboses at fixed dihedral angles, other bond lengths and bond angles. The average bond length found in the RNA structures (1.422Å) is employed here. Parameter k_{bond} is set to 5.0, which is optimized by backbone modelling (rebuilt main-chain atoms after fixing base positions).

$$E_{angle}(u) = \begin{cases} -2u - 1, & \text{if } u < -1, \\ u * u, & \text{if } -1 \leq u \leq 1, \\ 2u - 1, & \text{if } u > 1 \end{cases} \quad (15)$$

where $u(\theta_{POC}) = k_{angle}(\theta_{POC} - 120.7)$ and $u(\theta_{OCC}) = k_{angle}(\theta_{OCC} - 111.1)$. θ_{POC} is between P_{i+1}-O5'_{i+1}-C5'_{i+1} and θ_{OCC} is between O5'_{i+1}-C5'_{i+1}-C4'_{i+1}. Similar to the bond length, we made these two angles flexible so that we can add a phosphate group between two fixed riboses at fixed dihedral angles, other bond lengths and bond angles. The average values for these two angles in RNA structures are 120.7° and 111.1°, respectively. Parameter k_{angle} is set to 0.1, after a few trials of backbone modelling. For both bond-angle

and bond-length energies, a harmonic function with linear extrapolations at both ends was employed to improve conformational sampling efficiency.

$E_{torsion}$ is calculated from three statistical energy functions (see Figure 1E for angle definitions).

$$E_{torsion} = E(\nu_i, \varepsilon_i, \zeta_i) + E(\zeta_i, \alpha_{i+1}, \beta_{i+1}) + E(\beta_{i+1}, \gamma_{i+1}, \nu_{i+1}) \quad (13)$$

with $E(\nu_i, \varepsilon_i, \zeta_i) = -\ln P(\varepsilon_i, \zeta_i | \nu_i)$, $E(\zeta_i, \alpha_{i+1}, \beta_{i+1}) = -\ln P(\alpha_{i+1} | \zeta_i, \beta_{i+1})$, and $E(\beta_{i+1}, \gamma_{i+1}, \nu_{i+1}) = -\ln P(\beta_{i+1}, \gamma_{i+1} | \nu_{i+1})$. Here, two improper angles (ν_i and ν_{i+1}) belong to the two neighboring riboses connected by a phosphate group. Two dihedral (angles ε_i involving C2'-C3'-O3'-P_{i+1} and ζ_i involving C3'-O3'-P_{i+1}-O5'_{i+1}) determines the position of the phosphate group. In addition, α_{i+1} , β_{i+1} , and γ_{i+1} dihedral angles involve O3'_{i-1}-P_{i+1}-O5'_{i+1}-C5'_{i+1}, P_{i+1}-O5'_{i+1}-C5'_{i+1}-C4'_{i+1} and O5'_{i+1}-C5'_{i+1}-C4'_{i+1}-O4'_{i+1}, respectively⁴⁵. Here, the coupling between neighboring three dihedral angles, rather than the statistics of a single torsion angle was considered to improve the accuracy of main-chain modeling.

Structure Database

All statistical energy terms in Eq. (1) were derived from 2247 RNA structures obtained by X-ray crystallography and cryogenic electron microscopy with a resolution higher than 3.0Å downloaded on January 23, 2020. We did not remove any redundant sequences because we need a database as large as possible. Moreover, RNAs can have different structures in different complexes and we want to capture all possible conformations. We can do this because statistics are collected at the atomic or base level. See the discussion in more details.

Conformational Sampling Algorithm

1) Confirmation representation by the NuTree

For conformational sampling, each RNA structure is represented by the NuTree. Each node in the NuTree denotes an RNA residue including its base position (the local coordinate system), the ribose rotameric state and the phosphate position attached to the 3' position of the ribose (determined by dihedral angles ε and ζ , Figure 1E). The edge of the NuTree represents relative positions between the local coordinate systems of two bases, which are described by a 3x4 coordinate transformation matrix. There are ten types of coordinate transformation matrices to describe nine types of edges: 1) sequence-neighboring base pairs along 5' to 3' and 3' to 5' directions in the loop regions, 2) sequence-neighboring base pairs along 5' to 3' and 3' to 5' directions in the Watson-Crick pairing regions, 3) sequence-neighboring base pairs along 5' to 3' and 3' to 5' directions in the non-Watson-Crick pairing regions or connection between Watson-Crick pairing and non-Watson-Crick pairing regions, 4) Watson-Crick pairs, 5) Non-Watson-Crick pairs, 6) chain connection without a specific positional relation (jump). Except for jump, each edge has a corresponding move set. The move set was collected from all possible conformations of the same type in the structural database that were discretized into 60 representative conformations and each representative conformation was further discretized into 60 sub-representative conformations. These 3600 representative conformations form the move set for each edge type. Figure 1F shows the NuTree for the GCAA tetraloop with a sequence of G₀C₁G₂C₃A₄A₅G₆C₇ where the edge 0-1

is sequence-neighboring base pairs along the 5' to 3' direction in the Watson-Crick pairing region, edges 1-2 is sequence-neighboring base pairs along the 5' to 3' direction in non-Watson-Crick pairing regions, 2-3, and 3-4 are sequence-neighboring base pairs along the 5' to 3' direction in the loop regions, edges 0-7, 1-6 are Watson-Crick pairs and 2-5 is non-Watson-Crick pair. Although there is no edge between 4 and 5, we need to build the phosphate connection and calculate the internal energy between them.

2) Conformational Sampling Moves.

Conformational sampling involves node and edge sampling. Node sampling allows either a random selection of ribose rotamers or a small adjustment of the base local coordinate system ($<0.2\text{\AA}$ translational and $<2^\circ$ rotational motions). Both moves are local. They do not change the positions of other nodes and riboses. Edge sampling makes both local and global moves. The local move makes minor adjustment ($<0.2\text{\AA}$ translational and $<2^\circ$ rotational motions). The global move randomly selects a move from the set according to the edge type. Edge sampling moves will change the positions of all downstream nodes but not the relative positions between the nodes.

The above sampling may change the relative positions of some neighbouring riboses and make it necessary to rebuild the phosphate connection. Such connection can be rapidly built by the grid search from representative points for the lowest $E_{internal}$ value in the ϵ - ζ torsional angle space. According to the distribution of conformations (Supplementary Figure S4), we have divided ϵ - ζ space into 50 representative points, each representative point into 50 sub-representative points and each sub-representative point with 25 local points in 1° extension. That is, a total of 125 calculations is conducted for searching the lowest $E_{internal}$ value.

3) Monte Carlo Simulated Annealing.

All nodes in a NuTree are divided into three types: those with positions fixed (A), those with relative internal positions fixed but not absolute positions (B), and those with both relative and absolute positions changed (C). The energy change after each move is calculated by

$$\Delta E = dE(AB) + dE(AC) + dE(BC) + dE(CC) + dE_{rot} + dE_{internal} \quad (16)$$

where $dE(AB)$, $dE(AC)$, $dE(BC)$ and $dE(CC)$ denote the changes in interaction energies with the AB, AC, BC, and CC regions, respectively, and dE_{rot} and $dE_{internal}$ are the changes in rotameric and internal energies, respectively. Each move is accepted or rejected according to the Metropolis criterion. We employed simulated annealing for energy optimisation with an initial temperature set at 2.5, which is decreased by a factor of 0.95 at each round until temperature reached 0.01. In the refinement protocol, the initial temperature is set at 0.5 which is decreased to 0.01 by a factor of 0.9. The sampling steps at each temperature N_{step} is proportional to the edge number in the NuTree.

$$N_{step} = 400 * n(wc \text{ edge}) + 2000 * (nwc \text{ edge}) + 4000 * n(\text{other edge})$$

where $n(wc \text{ edge})$ is the edge number of Watson-Crick pair or helix neighbor, $n(nwc \text{ edge})$ is the edge number of Non-Watson-Crick pair or NWC neighbor, $n(\text{other edge})$ is the number of other edges.

To increase the efficiency of conformational sampling, we give *E_{clash}* and *E_{internal}* a low weight of 0.05 at the initial temperature and gradually increase the weight to 1.

Motif Test set

The test set is obtained from the benchmark of Rosetta-SWM³⁴, downloaded from <https://purl.stanford.edu/fq893cm4516>. After removing redundant RNAs and those with more than 50 bases and non-standard RNA bases, we have 48 motifs listed in Supplementary Table S1.

RNA Puzzle and FARFAR2 Test sets

RNA Puzzle data set is downloaded from https://github.com/RNA-Puzzles/raw_dataset_and_for_assessment. The prediction results of FARFAR2²⁴ are downloaded from <https://purl.stanford.edu/wn364wz7925>.

Data availability

The structural data and test sets used by BRiQ is publicly available at <http://servers.sparks-lab.org/downloads/BRiQ-dataset.tar.gz>

Code availability

The source code of BRiQ refinement is available at <http://servers.sparks-lab.org/downloads/RNA-BRiQ.tar.gz>

References

1. Rose, P.W. *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* **43**, D345-56 (2015).
2. Petrov, A.I. *et al.* RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Research* **43**, D123-D129 (2015).
3. Sun, L.Z., Zhang, D. & Chen, S.J. Theory and Modeling of RNA Structure and Interactions with Metal Ions and Small Molecules. *Annu Rev Biophys* **46**, 227-246 (2017).
4. Antunes, D., Jorge, N.A.N., Caffarena, E.R. & Passetti, F. Using RNA Sequence and Structure for the Prediction of Riboswitch Aptamer: A Comprehensive Review of Available Software and Tools. *Frontiers in Genetics* **8**(2018).
5. Flores, S.C., Wan, Y., Russell, R. & Altman, R.B. Predicting RNA structure by multiple template homology modeling. *Pac Symp Biocomput*, 216-27 (2010).
6. Rother, M., Rother, K., Puton, T. & Bujnicki, J.M. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res* **39**, 4007-22 (2011).
7. Das, R. & Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A* **104**, 14664-9 (2007).
8. Parisien, M. & Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**, 51-5 (2008).
9. Das, R. Atomic-Accuracy Prediction of Protein Loop Structures through an RNA-Inspired Ansatz. *Plos One* **8**(2013).
10. Krokhotin, A., Houlihan, K. & Dokholyan, N.V. iFoldRNA v2: folding RNA with constraints. *Bioinformatics* **31**, 2891-3 (2015).
11. Sharma, S., Ding, F. & Dokholyan, N.V. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* **24**, 1951-2 (2008).
12. Cao, S. & Chen, S.J. Physics-based de novo prediction of RNA 3D structures. *J Phys Chem B* **115**, 4216-26 (2011).

13. Xu, X. & Chen, S.J. Hierarchical Assembly of RNA Three-Dimensional Structures Based on Loop Templates. *J Phys Chem B* **122**, 5327-5335 (2018).
14. Zhao, Y. *et al.* Automated and fast building of three-dimensional RNA structures. *Sci Rep* **2**, 734 (2012).
15. Wang, J., Wang, J., Huang, Y. & Xiao, Y. 3dRNA v2.0: An Updated Web Server for RNA 3D Structure Prediction. *Int J Mol Sci* **20**(2019).
16. Boniecki, M.J. *et al.* SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res* **44**, e63 (2016).
17. Miao, Z. *et al.* RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **23**, 655-672 (2017).
18. Miao, Z. *et al.* RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **21**, 1066-84 (2015).
19. Miao, Z. *et al.* RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA* **26**, 982-995 (2020).
20. Feig, M. Computational protein structure refinement: Almost there, yet still so far to go. *Wiley Interdiscip Rev Comput Mol Sci* **7**(2017).
21. Adiyaman, R. & McGuffin, L.J. Methods for the Refinement of Protein Structure 3D Models. *Int J Mol Sci* **20**(2019).
22. Stasiewicz, J., Mukherjee, S., Nithin, C. & Bujnicki, J.M. QRNAS: software tool for refinement of nucleic acid structures. *BMC Struct Biol* **19**, 5 (2019).
23. Das, R., Karanicolas, J. & Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* **7**, 291-4 (2010).
24. Watkins, A.M., Rangan, R. & Das, R. FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure* **28**, 963-976 e6 (2020).
25. Alford, R.F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031-3048 (2017).
26. Zhou, Y., Zhou, H., Zhang, C. & Liu, S. What is a desirable statistical energy function for proteins and how can it be obtained? *Cell Biochemistry and Biophysics* **46**, 165-174 (2006).
27. Wang, J., Zhao, Y., Zhu, C. & Xiao, Y. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Res* **43**, e63 (2015).
28. Zhang, T.C., Hu, G.D., Yang, Y.D., Wang, J.H. & Zhou, Y.Q. All-Atom Knowledge-Based Potential for RNA Structure Discrimination Based on the Distance-Scaled Finite Ideal-Gas Reference State. *Journal of Computational Biology* (2019).
29. Tan, Y.L. *et al.* What is the best reference state for building statistical potentials in RNA 3D structure evaluation? *RNA* **25**, 793-812 (2019).
30. Tinoco, I. & Bustamante, C. How RNA folds. *Journal of Molecular Biology* **293**, 271-281 (1999).
31. Murray, L.J., Arendall, W.B., 3rd, Richardson, D.C. & Richardson, J.S. RNA backbone is rotameric. *Proc Natl Acad Sci U S A* **100**, 13904-9 (2003).
32. Dill, K.A. & MacCallum, J.L. The protein-folding problem, 50 years on. *Science* **338**, 1042-6 (2012).
33. Dunbrack, R.L., Jr. & Karplus, M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol* **1**, 334-40 (1994).
34. Watkins, A.M. *et al.* Blind prediction of noncanonical RNA structure at atomic accuracy. *Sci Adv* **4**, eaar5316 (2018).
35. Tan, D., Piana, S., Dirks, R.M. & Shaw, D.E. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc Natl Acad Sci U S A* **115**, E1346-E1355 (2018).
36. Gardner, P.P., Wilm, A. & Washietl, S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research* **33**, 2433-2439 (2005).
37. De Leonardis, E. *et al.* Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Research* **43**, 10444-10455 (2015).

38. Weinreb, C. *et al.* 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell* **165**, 963-75 (2016).
39. Wang, J. *et al.* Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. *Nucleic Acids Research* **45**, 6299-6309 (2017).
40. Cutarello, F., Tiana, G. & Bussi, G. Assessing the accuracy of direct-coupling analysis for RNA contact prediction. *RNA* **26**, 637-647 (2020).
41. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y.Q. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications* **10**(2019).
42. Pobleto, S., Bottaro, S. & Bussi, G. A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs. *Nucleic Acids Res* **46**, 1674-1683 (2018).
43. Frisch, M.J.T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, (Gaussian Inc., Wallingford, CT, 2009).
44. Zhao, Y. & Truhlar, D.G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **120**, 215-241 (2008).
45. Richardson, J.S. *et al.* RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **14**, 465-81 (2008).

Acknowledgement

This work was supported by Australian Research Council DP180102060 and DP210101875 to Y.Z. We also gratefully acknowledge the use of the High Performance Computing Cluster ‘Gowonda’ to complete this research, and the aid of the research cloud resources provided by the Queensland CyberInfrastructure Foundation (QCIF). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. PX would like to thank the hospitality of Shenzhen Bay Laboratory while he is a visiting scholar to the laboratory.

Author contribution

PX designed and performed the method and wrote the manuscript, RW assisted QM calculations, TZ assisted in dataset building, JZ and YZ conceived of the study, participated in the initial design, and assisted in data analysis. YZ drafted the whole manuscript. All authors read, contributed to the discussion, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Table S1. The test set of 48 RNA motifs with lower case letters indicating fixed regions predefined in Rosetta-SWM.

| motif name | Total length | Model length | sequence |
|----------------------------|--------------|--------------|---|
| 23s_rrna_44_49 | 45 | 6 | cuGAUGAAggcggauuuccgaaugauacaaucaaugaggagagga |
| 3P_j55a_group_I_intron | 13 | 4 | caacgcgACAUgg |
| 5P_j12_leadzyme | 15 | 4 | ggaggcCGAGcgagg |
| 5P_j55a_group_I_intron | 16 | 5 | ucAACAGcgacagguc |
| 5P_p1_m_box_riboswitch | 12 | 4 | gaaaacGAUUga |
| cg_helix | 8 | 6 | cGCGCGCg |
| gaaa_minor_dock | 18 | 4 | ccGAAAgggaccuagguu |
| gagua_pentaloop | 9 | 5 | cgGAGUAcg |
| gagu_forcesyn_blockstackU | 16 | 8 | acGAGUguacGAGUgu |
| gcaa_tetraloop | 8 | 4 | gcGCAAgc |
| gg_mismatch_fixed | 10 | 2 | caGgcgcGug |
| gir1_p2.lp5_kiss_fixed | 21 | 13 | ggCUAAUCAccggGACGAUcc |
| gir1_p2p9_gaaa_minor_fixed | 20 | 8 | ugGGuuccGAAAggaaGaca |
| g_quadruplex_fixed | 20 | 16 | gGGGuGGGgUGGgGUGGGGU |
| hammerhead_3WJ_cat_fixed | 23 | 11 | agCUGAUGAguacGAAacguCcu |
| hammerhead_tert_fixed | 20 | 10 | gUACUaccCAAUAagguauc |
| hepatitis_C_virus_ires_IIa | 14 | 5 | ggAACUAcuggcaa |
| hiv_rre_fixed | 13 | 5 | cgGGcgcgGUAcg |
| j23_group_II_intron | 29 | 7 | gcAGCAGAAgggaacgaaccgcacguggu |
| j24_tpp_riboswitch | 23 | 5 | gggagcCUUUGaaccugaggag |
| j31_glycine_riboswitch | 25 | 7 | aggagaaaauagaggcAAAAGGAcaca |
| j44a_p4p6 | 15 | 5 | ggAAAaggccuAAcc |
| j55a_P4P6_fixed | 21 | 9 | gGUCAACAGCcggaucagguc |
| just_tr_P4P6_fixed | 15 | 5 | ccUAAgucgauAUgg |

| | | | |
|-------------------------|----|----|---------------------------|
| kink_turn_fixed | 17 | 9 | guGGAggccGAUGAAac |
| kiss_add_fixed | 24 | 16 | aaUGAUUUGGUuuagCCUUAACu |
| kiss_add_L2_fixed | 24 | 9 | aaUGAUUUGGUuuagccuuaaacu |
| kiss_add_L3_fixed | 24 | 7 | aaugauauguuuagCCUUAACu |
| kiss_add_PK_dock | 24 | 12 | aaUGAUUUGGUuuagccUUAACu |
| l1_sam_II_riboswitch | 25 | 7 | gcGAUUUAAccguugcaugcuaaaa |
| l2_viral_rna_pseudoknot | 20 | 7 | gcggcgccgGAAACAAcG |
| loopE_fixed | 22 | 14 | ccGAUGGUAgugcGAGAGUAgg |
| P5b_connect | 24 | 18 | aguCUCAGGGGAAACUUUGAGaug |
| parallel_AA | 8 | 6 | aAAAAAAa |
| puzzle1 | 15 | 7 | cgCGCcaugCCUGug |
| puzzle18_zika_3WJ | 27 | 3 | gUcaggccgCcacaugugcAgccug |
| puzzle1_alt_fixed | 21 | 7 | ugCCUGugcgCGCcagccg |
| r2_4x4_fixed | 16 | 8 | ugAAGCccggAGGAca |
| srl_fixed | 19 | 9 | ccUAGUAcgagagGACCgg |
| srp_domainIV_fixed | 16 | 8 | guCAGGucgaAGCAgc |
| tandem_ga_imino_fixed | 12 | 4 | cgGAcgGAcg |
| tandem_ga_sheared_fixed | 12 | 4 | gcGAgcgGAcg |
| t_loop_fixed | 12 | 7 | gggUUCGAUUcc |
| tl_tr_P4P6 | 21 | 9 | gGAAAccUAAGucgauAUgg |
| tl_tr_P4P6_dock | 23 | 9 | ggGAAAcuccUAAGucgauAUgg |
| uucg_tetraloop | 8 | 4 | acUUCGgu |
| VS_rbzm_P2P3P6_fixed | 19 | 7 | ggAAcuaggcgcACAAGcc |
| VS_rbzm_P3P4P5_fixed | 22 | 10 | ccUCcaugUCguauGUGAUUgg |

Supplementary Table 2. The RNA puzzle test set along with RMSD values for the best pre-refinement model, the best model in BRiQ-predicted models and the best model in BRiQ-sampled models for each RNA puzzle experiment.

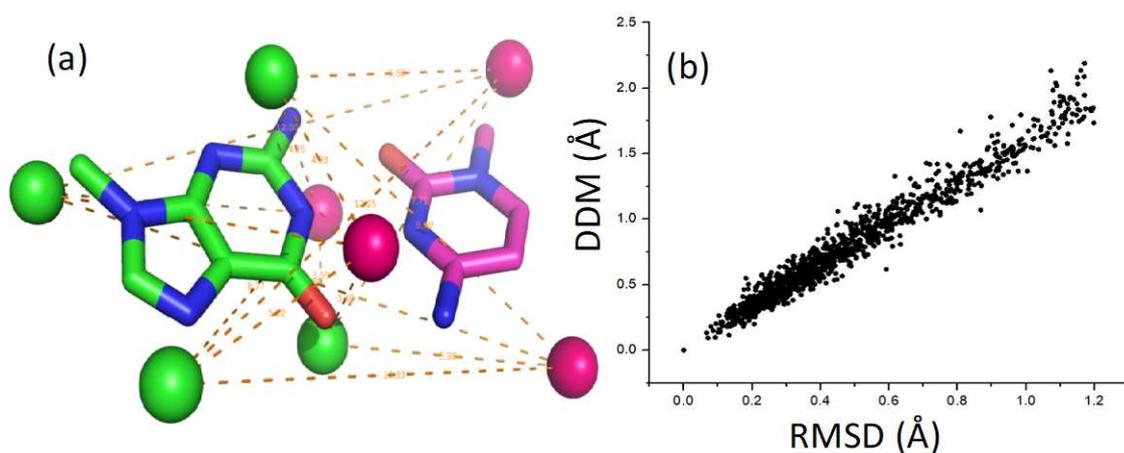
| Puzzle # | Length | RMSD (Best Model) | | |
|----------|--------|-------------------|----------------|--------------|
| | | pre-refinement | BRiQ-predicted | BRiQ-sampled |
| | | | | |

| | | | | |
|-------------|------------|--------------|--------------|--------------|
| PZ01 | 46 | 3.433 | 1.968 | 1.694 |
| PZ02 | 98 | 2.297 | 2.062 | 1.821 |
| PZ03 | 84 | 7.699 | 6.531 | 5.9 |
| PZ04 | 126 | 3.352 | 2.229 | 1.918 |
| PZ05 | 188 | 9.031 | 9.069 | 8.915 |
| PZ06 | 158 | 11.598 | 10.653 | 9.421 |
| PZ07 | 185 | 20.581 | 20.701 | 20.155 |
| PZ08 | 96 | 4.801 | 3.797 | 3.797 |
| PZ09 | 71 | 6.043 | 5.954 | 5.675 |
| PZ10 | 157 | 6.84 | 8.956 | 7.251 |
| PZ11 | 57 | 4.799 | 5.009 | 4.074 |
| PZ12 | 117 | 10.18 | 9.964 | 9.699 |
| PZ13 | 60 | 5.41 | 5.52 | 5.157 |
| PZ14_bound | 61 | 5.978 | 6.144 | 5.541 |
| PZ14_free | 58 | 6.864 | 6.258 | 5.302 |
| PZ15 | 68 | 6.451 | 6.779 | 6.192 |
| PZ17 | 58 | 5.161 | 5.683 | 5.093 |
| PZ18 | 71 | 3.74 | 3.469 | 2.984 |
| PZ19 | 62 | 5.529 | 6.966 | 4.17 |
| PZ20 | 68 | 4.71 | 3.542 | 3.522 |
| PZ21 | 41 | 3.83 | 3.829 | 3.526 |
| PZ23 | 37 | 8.268 | 8.657 | 7.991 |
| PZ24 | 111 | 4.824 | 5.528 | 4.486 |
| PZ25 | 69 | 2.551 | 2.448 | 2.159 |

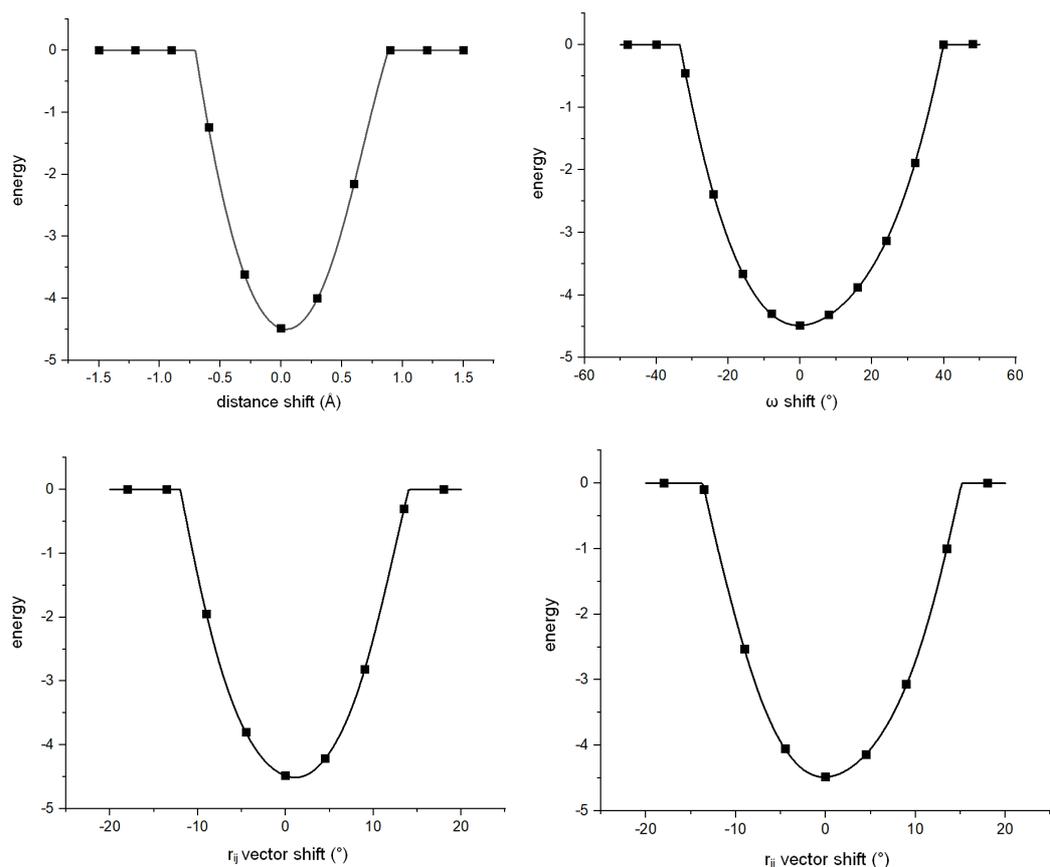
Supplementary Table S3. Refinement of 2000 FARFAR2-predicted models for 12 RNA Puzzles by BRiQ. Performance is compared by using the best model (smallest RMSD) within top 1% and top 5% predicted by FARFAR2 and BRiQ energy scores.

| | Top 1% | | Top 5% | |
|--|----------------------|-------------------|----------------------|-------------------|
| | FARFAR2 Best RMSD | BRiQ Best RMSD | FARFAR2 Best RMSD | BRiQ Best RMSD |
| | | | | |

| | (Å) | (Å) | (Å) | (Å) |
|-------|--------|-------|-------|-------|
| PZ1 | 2.626 | 2.115 | 2.398 | 2.115 |
| PZ2 | 2.409 | 2.083 | 2.387 | 1.801 |
| PZ4 | 1.766 | 1.139 | 1.766 | 1.099 |
| PZ8 | 5.464 | 4.624 | 5.129 | 4.624 |
| PZ9 | 4.394 | 4.016 | 3.828 | 4.016 |
| PZ11 | 5.009 | 5.081 | 4.703 | 4.077 |
| PZ13 | 6.211 | 4.687 | 5.646 | 4.687 |
| PZ14F | 10.929 | 3.904 | 3.197 | 3.338 |
| PZ18 | 4.21 | 2.9 | 3.267 | 2.659 |
| PZ19 | 4.753 | 4.472 | 4.699 | 4.304 |
| PZ20 | 3.036 | 3.044 | 3.036 | 2.647 |
| PZ21 | 5.674 | 5.23 | 4.361 | 4.18 |

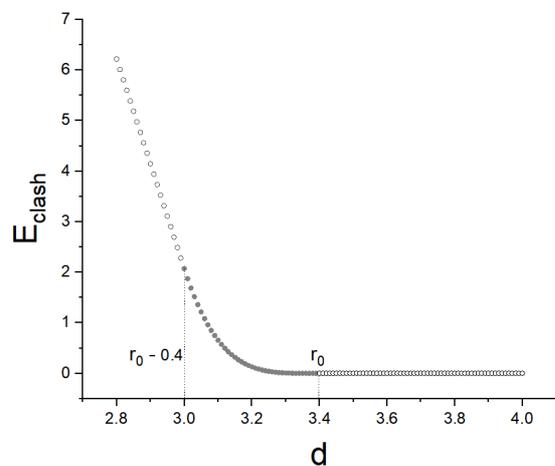


Supplementary Figure S1. a) Four points in the local base coordination system are employed to measure the relative orientation between two bases. b) DDM, the mean squared distance between the four points of two separate bases has a high correlation coefficient with the RMSD after optimizing the locations of four points.

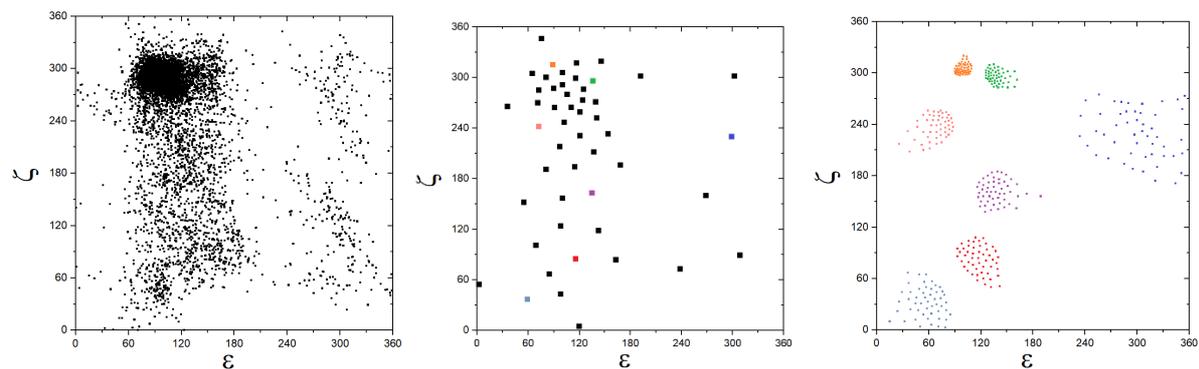


Supplementary Figure S2: An example of the statistical energy score for the AU pair with a single degree of freedom variation around the local minimum (distance, ω , r_{ij} and r_{ji} vector angle shifts). The black squares are the grid points. The energy values on grid points were pre-calculated and stored in an energy table. The final energy value of a target base pair was

calculated by the interpolation from nearest 36 grid points (nearest two points in distance and dihedral angle dimensions, nearest three points in r_{ij} and r_{ji} dimension, $2*2*3*3=36$).



Supplementary Figure S3. $E_{clash}(d)$ as a function of the atomic distance. Here $k_{clash} = 3$.



Supplementary Figure S4 (a) The conformational distribution in the ϵ - ζ torsional space in the database (b) the 50 representative points (c) 50 sub-representative points for several selected representative points.

Figures

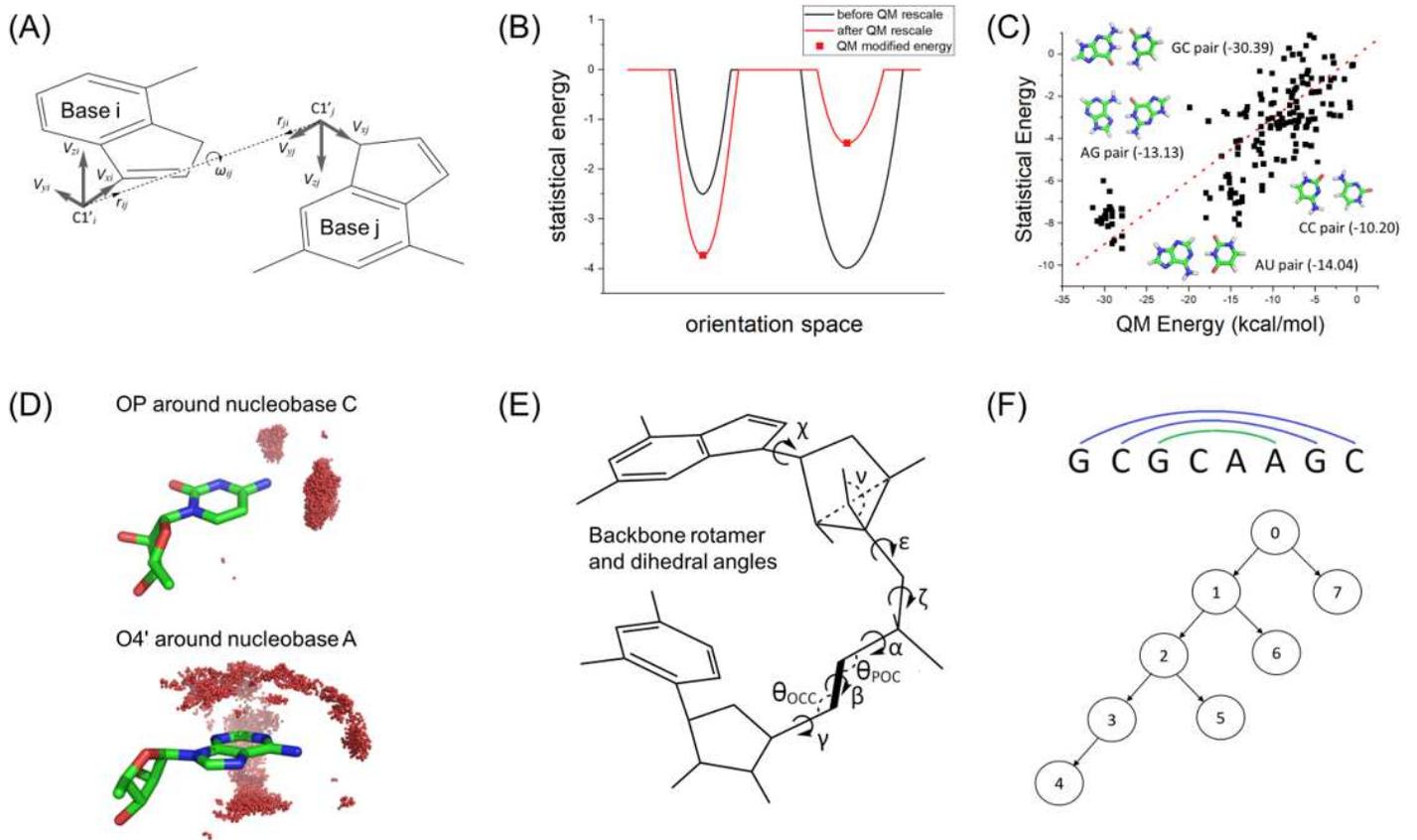


Figure 1

(A) Six-dimensional base-base statistical potentials with relative positions defined by r_{ij} and r_{ji} vectors along with the rotational angle ω_{ij} . (B) A schematic illustration of the orientation dependence before and after Quantum Mechanical (QM) energy scaling. (C) This QM scaling is based on the correlation between the statistical energy scores of hydrogen-bonded base pairs ($-\ln \text{Phb}$) and QM calculations. Phb is the probability of hydrogen-bonded base pairs. (D) The distribution of OP around nucleobase C and the distribution of O4' around the nucleobase A as labelled. (E) Backbone rotamers defined according to various torsion and improper angles that control the ribose (χ , ν) and phosphate (ϵ , ζ) backbone. Dihedral angles α , β , γ , bond angles θ_{POC} , θ_{OCC} and bond length of C5'O5' were required to calculate the internal energy. (F). Nucleobase-centric fold-tree (NuTree) algorithm for refinement by defining bases as the nodes and locally or globally connected bases as the edges, illustrated by the GCAA tetraloop.

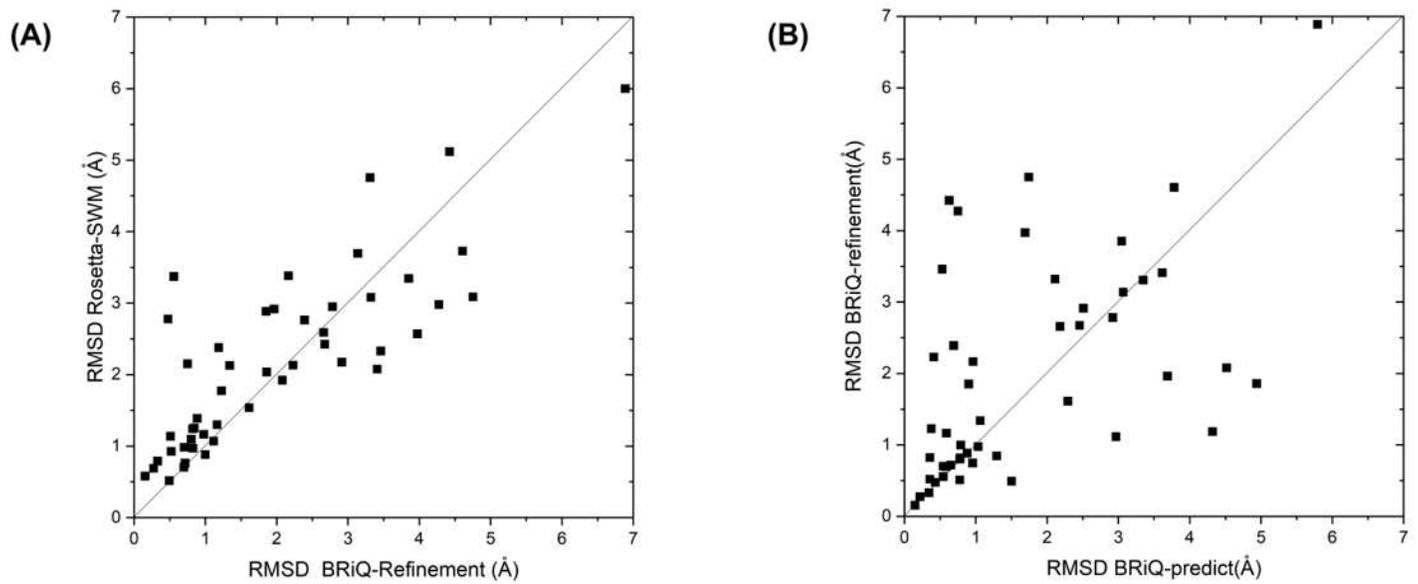


Figure 2

(A) Refining Rosetta-SWM motif models by BRiQ improves the majority of those model structures with $\text{RMSD} < 2\text{\AA}$ as demonstrated by RMSD comparison (Lowest RMSD of top 1% before refinement on Y-axis versus after refinement on X-axis). (B) Fixing all native base pairs with random initial conformations for all other regions and then folding the remaining structure leads to more accurate motif models for the majority of the motifs than refining Rosetta-SWM models that have pre-assigned, partially fixed base pairs (Lowest RMSD of top 1% BRiQ-refined models on Y-axis versus lowest RMSD of top 1% BRiQ models with fixed native base pairs on X-axis).

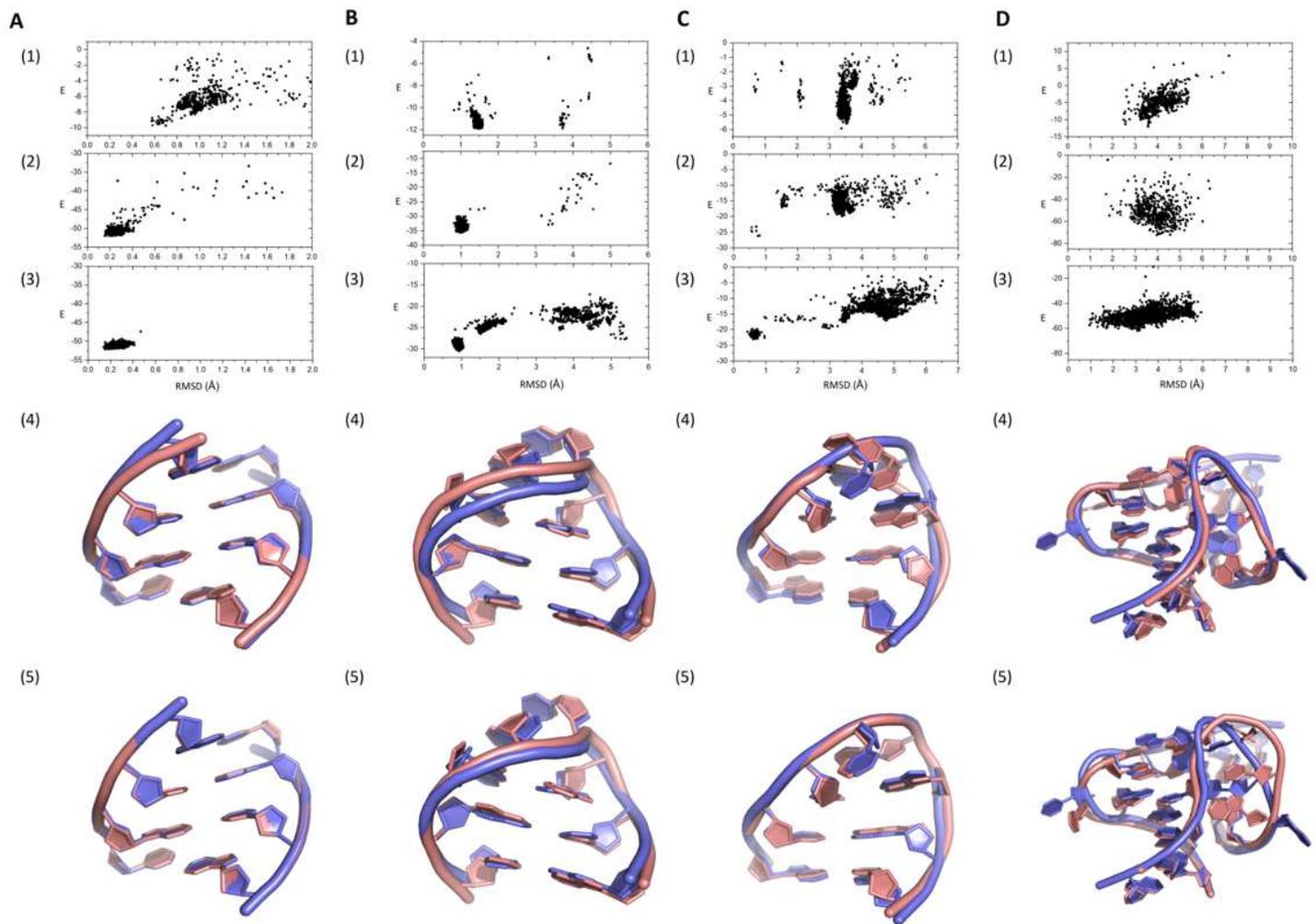


Figure 3

Energy versus RMSD values of the conformations sampled by (1) Rosetta-SWM, (2) refinement of Rosetta-SWM conformations by BRIQ, (3) structure prediction by BRIQ with all native base pairs fixed for (A) CG-helix, (B) GCAA_tetra loop, (C) UUCG_tetra loop, and (D) j55a_P4P6_fixed. Structure alignment of native (blue) to the best in top 1% predicted models by Rosetta-SWM (panel 4) and BRIQ refinement (panel 5) in the bottom panel.

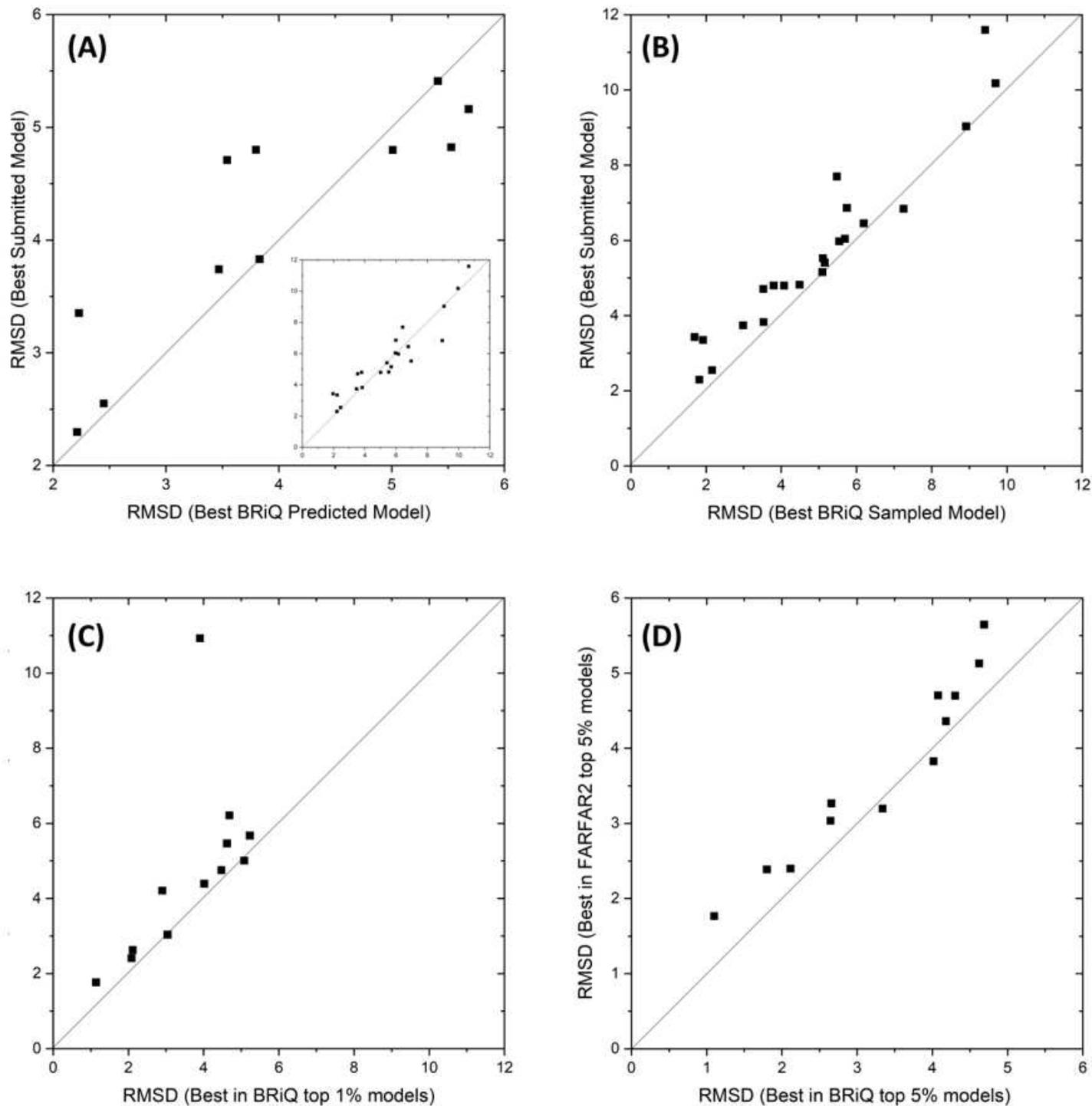


Figure 4

BRiQ refinement achieves consistent improvement over the best-submitted models for RNA puzzles containing model structures with RMSD $< 4\text{\AA}$ (A), over all RNA puzzles in conformational sampling (B), over FARFAR2 models for 10 out of 12 RNA puzzles in top 1% (C) and top 5% (D) predicted models.

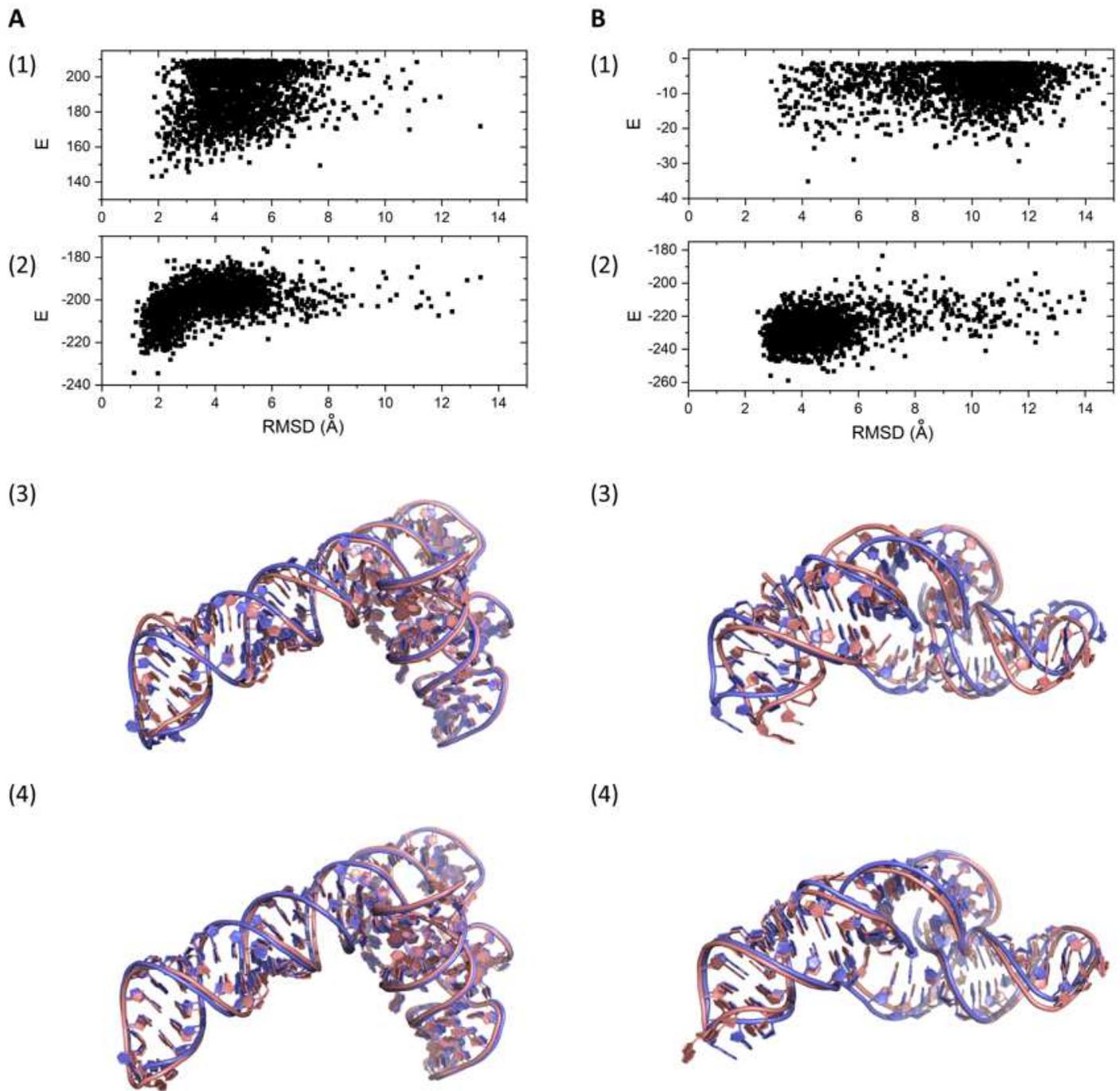


Figure 5

Energy versus RMSD values of the conformations sampled by (1) FARFAR2, (2) refinement of FARFAR2 conformations by BRiQ for (A) RNA Puzzle 4 and (B) RNA Puzzle 18. Structure alignment of native (blue) to the best in top 1% predicted models by FARFAR2 (3) and BRiQ refinement (4) in the bottom panel.