Prediction of 3D RNA Structures from Sequence Using Energy Landscapes of RNA Dimers: Application to RNA Tetraloops

Ivan Isaac Riveros[1] and Ilyas Yildirim[1*]

[1] Department of Chemistry and Biochemistry, Florida Atlantic University, Jupiter, FL 33458 USA

* Authors to whom correspondence is addressed. Phone: (561) 799-8325. Email: iyildirim@fau.edu (I.Y.).

Abstract.

Access to the three-dimensional structure of RNA enables an ability to gain a more profound understanding of its biological mechanisms, as well as the ability to design RNA-targeting drugs, which can take advantage of the unique chemical environment imposed by a folded RNA structure. Due to the dynamic and structurally complex properties of RNA, both experimental and traditional computational methods have difficulty in determining RNA's 3D structure. Herein, we introduce TAPERSS (Theoretical Analyses, Prediction, and Evaluation of RNA Structures from Sequence), a physics-based fragment assembly method for predicting 3D RNA structures from sequence. Using a fragment library created using discrete path sampling calculations of RNA dinucleoside monophosphates, TAPERSS can sample the physics-based energy landscapes of any RNA sequence with relatively low computational complexity. We have benchmarked TAPERSS on 21 RNA tetraloops, using a combinatorial algorithm as a proof-of-concept. We show that TAPERSS was successfully able to predict the apo-state structures of all 21 RNA hairpins, with 16 of those structures also having low predicted energies as well. We demonstrate that TAPERSS performs most accurately on GNRA-like tetraloops with mostly stacked loop-nucleotides, while having limited success with more dynamic UNCG and CUYG tetraloops, most likely due to the influence of the RNA force field used to create the fragment library. Moreover, we show that TAPERSS can successfully predict the majority of the experimental non-apo states, highlighting its potential in anticipating biologically significant yet unobserved states. This holds great promise for future applications in drug design and related studies. With discussed improvements and implementation of more efficient sampling algorithms, we believe TAPERSS may serve as a useful tool for a physics-based conformational sampling of large RNA structures.

INTRODUCTION.

Ribonucleic acid (RNA) plays a crucial role in various cellular processes in part due to its complex structural dynamics. Its high conformational variability enables a wide range of functions beyond the traditional "central dogma" primarily through non-coding RNAs (ncRNAs). NcRNAs serve multiple roles in genetic regulation and expression, including riboswitches, miRNAs, siRNAs, tRNAs, as well as catalytic activities in rRNA and ribozymes.[1-2] Moreover, RNA's ability to fold into intricate structures contributes to its regulatory function in coding RNA as well. The folded structure of pre-mRNA can either inhibit or promote the binding of small nuclear ribonucleoproteins, modulating genetic splicing and thus gene expression.[3] Understanding the three-dimensional (3D) structure of RNA provides valuable insights into its biological functions and enables the design of small molecules that can selectively target RNA. The detailed structures of RNA motifs, such as stem loops, internal loops, and bulges, are particularly useful in RNA-targeted drug design studies, as these motifs create unique molecular environments that can be targeted with high specificity.[4-6] For instance, stabilization of an intronic adenine-bulge by a small molecule has been shown to regulate aberrant alternative splicing in frontotemporal dementia associated with chromosome 17.[7-9] Risdiplam, another small molecule targeting RNA for treatment of spinal muscular atrophy, as well as similar compounds, were also shown to bind to and stabilize an associated mutant pre-mRNA A-bulge, also likely regulating aberrant splicing.[10-11]

While access to RNA’s 3D structure can provide valuable insight, investigating RNA's 3D structure often becomes a complex and laborious task. The traditional "thermodynamic hypothesis"[12] suggests that the native (or biologically active) state of biomacromolecules corresponds to the global energy minimum, however, experimental findings have shown that RNA possesses a complex folding landscape with multiple stable conformations that it can transition between overextended timescales.[13-15] Often, several of these stable conformations are biologically relevant, and play a role in the RNA's cellular function, as is the case with riboswitches.[16] Obtaining a comprehensive view of an RNA’s structural dynamics may then require experimental elucidation of these multiple states. Furthermore, the experimental determination of RNA’s three-dimensional structure fundamentally presents challenges due to the difficulties in crystallization and the need for extreme non-biological conditions. This is exemplified by the significant disparity between the number of resolved RNA structures (1,691) and protein structures (175,121) in the Protein Data Bank (PDB).[17]

Computational studies of RNA structure are complicated by its intricate landscape as well. Even with the utilization of GPU technology and massive parallelization, conventional molecular dynamics (MD) simulations fall short in covering biologically relevant timescales for studying even relatively small RNA sequences. To gain a comprehensive understanding of molecular structure, it is necessary to consider the statistical ensemble of the molecule, which includes its metastable states and the kinetics of transitions between them. However, the often-lengthy simulation time of MD restricts the ability to explore the molecule's full configuration space. Additionally, MD simulations often get trapped in local energy minima further limiting conformational sampling. To address these limitations, several techniques have been developed to enhance the efficiency of energy landscapes sampling. These methods include umbrella sampling[18], replica-exchange MD (REMD)[19], Markov state models (MSMs)[20] and discrete path sampling (DPS)[21]. Unlike the other methods listed, DPS does not rely on MD for sampling conformational space. Instead, it employs transition state theory and a modified nudged elastic band method to explore a significant portion, if not the entire, conformational landscape. DPS generates multiple energetic minima connected by transition states, along with associated free energies for each minimum. By utilizing DPS and other landscape sampling methods, it becomes feasible to comprehensively sample larger biomolecular systems than what is achievable with conventional MD simulations. However, there are still inherent limitations in studying large systems with numerous degrees of freedom, particularly in the case of RNA. The limits to "traditional" computational methods have led to a growing emphasis on the development of novel computational methods to predict 3D RNA structures, garnering attention over the past two decades.

Non-traditional methods for predicting 3D RNA structures can be broadly categorized as knowledge-based or physics-based, with the latter often employing coarse-grained (CG) modeling and simulation. CG modeling simplifies RNA models to reduce computational complexity, enabling exploration of the folding landscape over longer timescales. However, simplified models can lead to inaccurate predictions due to cumulative errors and challenges in reconstructing atomistic details. To compensate, CG methods may incorporate knowledge-based elements such as empirically derived energy functions or input secondary structures. Methods in this category include iFoldRNA[22-24], SimRNA[25], HiRE-RNAv3[26], IsRNA2[27] and cgRNASP-CN[28]. Physics-based methods in principle have unrestricted conformational sampling, hence, the conformations accessible and available to study are not limited by the available experimental data. Knowledge-based methods, on the other hand, typically rely directly on experimental data potentially limiting the explorable conformational space, ultimately leading to biased sampling. Between the two categories however, knowledge-based methods have the computational advantage, and can theoretically study systems larger than may be possible with physics-based methods. Knowledge-based methods also typically rely on the input or prediction of an RNA secondary structure to guide tertiary structure construction. These methods can be classified as homology modeling, motif assembly, or fragment assembly. Homology modeling predicts structures based on sequence alignments with templates, though it may encounter difficulties when suitable templates are unavailable. Tools like ModeRNA[29], can address minor template issues through fragment insertion. Motif assembly methods, such as MC-Fold/MC-Sym[30], RNAComposer[31-32], VFold[33-35], 3dRNA[36] and F-RAG[37], utilize loop motif fragments from respective databases to construct 3D structures. F-RAG and RNAComposer employ graph representations of RNA to guide tertiary structure construction. Methods like 3dRNA, MC-Fold/MC-Sym, and Vfold can predict secondary structures to use as guides in building tertiary structures. Of particular relevance to this work are fragment assembly methods, like FARNA/FARFAR[38-39] and FARFAR2[40], which use small RNA fragments of one to three nucleotides (nt) in length to construct larger structures. FARNA/FARFAR uses a 3nt fragment library created using the large ribosomal subunit from *Haloarcula marismortui*, assembling structures with a Monte Carlo simulator. The Monte Carlo simulations are guided by an empirically derived, low resolution energy function, which promotes selection of folded states. The Rosetta all-atom energy function, with specific modifications for promotion of RNA folding, is then used to improve the quality of predictions by removing stereochemical violations and incorporating more realistic energies.[38-39, 41] The original FARFAR showed sampling issues for structures greater than 12nt, which is improved upon by FARFAR2. FARFAR2 uses an expanded fragment library, additional Monte Carlo moves, improved scoring functions and additional optimizations. The improved method now shows effective sampling for sequences up to 80nt long.

Herein, we introduce TAPERSS (Theoretical Analyses, Prediction, and Evaluation of RNA Structures from Sequence), a physics-based fragment assembly method for predicting 3D RNA structures. TAPERSS leverages dinucleotide monophosphate (DNMP) fragments generated by DPS (**Figure 1**), enabling a simplified physics-based sampling of conformational landscape while preserving the low computational complexity and atomistic nature of fragment assembly methods. We have used DPS to explore the conformational space of all 16 non-post-transcriptionally modified RNA DNMPs. The resulting structures were used to create our 16 fragment libraries containing 52-150 structures in each, with individual structures having an associated relative 'folding' energy. As an initial proof-of-concept, we developed a combinatorial algorithm that predicts RNA 3D structures solely based on the input sequence and a simple physics-based energy function. While the combinatorial algorithm's scalability is limited due to the exponential increase in landscape size with additional nucleotides, we implemented key optimizations in the assembly and energy calculation steps, resulting in an efficient codebase.

To validate our method, we applied the combinatorial algorithm to study 21 RNA tetraloops with known experimental structures. Specifically, we focused on 6 nucleotide-long sequences forming an RNA tetraloop with one closing base pair. Employing the combinatorial algorithm, we explored the entire accessible conformational landscape of each sequence, generating over 1 billion structures for each, with thousands of valid tetraloops identified within the landscape. For benchmarking purposes, we have primarily compared predictions to experimental structures in the apo (or native) state, meaning no interactions with proteins and/or other biological systems. We demonstrate that our method can predict structures and associated free energies which fall in line with expectations and experimental data. Our method was able to structurally predict all 21 RNA tetraloops with root mean square deviation (RMSD) values less than 2 Å with respect to the experimental apo states, with 16 out of the 21 tetraloops also having low relative folding energies for the predicted apo states. Furthermore, we demonstrate that our method can also predict the majority of experimental non-apo states, indicating TAPERSS' potential to predict biologically relevant but unobserved states, which holds promise for future applications in drug design and related studies. The version of TAPERSS described herein should serve as a valuable tool for detailed structural investigations of small RNA hairpins. With our suggested improvements for future iterations, it may eventually serve as a powerful physics-based large RNA structure prediction method.

METHODS.

RNA DNMPs observed in RNA hairpin loops. The CoSSMos database[42] displayed 1460 NMR and X-ray structures available in the literature containing 3-/4-/5-/6-/7-nucleotide-long RNA hairpins (**Table S1, and Figure 2A**). Analyses showed 698, 7655, 2574, 2207, and 2393 conformations for 3-, 4-, 5-, 6-, and 7-nucleotide-long hairpins, respectively, which were reduced to 676, 7405, 2512, 2134, and 2353 after discarding structures having non-Watson-Crick terminal basepairs. The final list included 15080 RNA hairpins with terminal basepairs in Watson-Crick AU, UA, GC, CG, GU, and UG forms. We then extracted all the RNA DNMP structures observed in these RNA hairpins, which yielded a range of conformations from 1913 (in RNA CC) to 10672 (in RNA AA) (**Figure 2A**). These RNA DNMPs were then compared to the structures observed in discrete path sampling (DPS) calculations to create the fragment library (*vide infra*).

Discrete Path Sampling (DPS) calculations. The DPS method[21, 43] is an approach based on geometry optimization, where the configuration space is efficiently sampled to extract global and local minima. "Rare event" dynamics typically inaccessible in molecular dynamics simulations can be investigated with DPS. We previously utilized the DPS approach in different RNA systems[44-47] including RNA DNMP,[21] where amber99 force field[48] with revised χ[49] and α/γ[45] torsional parameters were employed in the studies. The results of the RNA DNMP[21] were used in the creation of the RNA fragment library (*vide infra*). The harmonic superposition approximation[50-52] was employed to estimate the free energies for the database of RNA DNMPs, which are exercised to describe the free energy change between two conformational states. The OPTIM and PATHSAMPLE codes (https://www-wales.ch.cam.ac.uk/software.html) were used in the DPS calculations. Results of DPS display total number of minima in the ranges of 123K-161K, 49K-63K, 60K-80K, and 43K-56K for Purine-Purine, Purine-Pyrimidine, Pyrimidine-Purine, and Pyrimidine-Pyrimidine, respectively (**Figure 2B**).

Creation of RNA DNMP Fragment Library. An RNA DNMP has two RNA residues connected with a phosphate group (**Figure 1**). As described above, the structures predicted by DPS calculations were compared to the RNA DNMPs extracted from 3-/4-/5-/6-/7-nucleotide-long RNA hairpins. To do so, three RMSD values between the DPS and experimental structures are calculated; base and sugar heavy atoms of i) first (rms_1) and ii) second (rms_2) residues, and iii) all heavy atoms of RNA DNMP except phosphate group ($rms_{1,2}$). If $rms_1 \leq 0.3$ Å, $rms_2 \leq 0.3$ Å, and $rms_{1,2} \leq 0.8$ Å, we decided that the DPS structure is a good representation of the experimentally observed RNA DNMP (**Script S1**). During the comparisons, structures predicted by DPS with energies greater than 40 kcal/mol with respect to global minimum are discarded. Results yielded total number of structures in the ranges of 5450-8955, 2480-4422, 2705-7118, and 1704-4040 for Purine-Purine, Purine-Pyrimidine, Pyrimidine-Purine, and Pyrimidine-Pyrimidine, respectively (**Figure 2B**). We then performed cluster analyses on each RNA DNMP, where structures with $rms_1 \leq 0.3$ Å, $rms_2 \leq 0.3$ Å, and $rms_{1,2} \leq 0.5$ Å are grouped into same cluster (**Script S2**). From the clusters holding more than one configuration, the structures with the lowest free energies are extracted from their respective clusters and included in the RNA DNMP fragment library with predicted 'folding' free energies with respect to the global minima (**Script S3**). Final analyses display total number of fragments in the ranges of 116-146, 73-117, 96-143, and 52-150 for Purine-Purine, Purine-Pyrimidine, Pyrimidine-Purine, and Pyrimidine-Pyrimidine, respectively (**Figure 2B**).

Experimental RNA Tetraloop Structures. The global minimum predicted by TAPERSS should ideally represent the apo form of an RNA sequence. Thus, to have proper comparisons between predictions and experiments, experimental structures displaying the apo forms were required for benchmarking purposes. RNA tetraloop structures extracted from the CoSSMos database[42] were analyzed for this objective. We identified 21 RNA tetraloop sequences (**Table 1**) with experimental apo states available for our structural comparison studies (**Table S2**). Furthermore, all the PDB structures in the literature containing these RNA tetraloop sequences were extracted and analyzed (**Table S2**). For each sequence, we identified a single apo structure which preferentially came from NMR studies and was otherwise ensured to have no external contacts distorting the structure (**Table S2**). This apo structure was used as a reference in further analyses. Several non-apo conformations were discovered for several RNA tetraloop sequences such as C$\underline{\text{UUCG}}$G, and G$\underline{\text{GAAA}}$C (**Table 1 and Figures S1-S21**). The non-apo conformations arose from close contact with nearby residues in buried RNA tetraloops and/or crystal packing contacts, implying excited state conformations for these RNA tetraloop sequences. Thus, another benchmark we performed was to test if TAPERSS can predict these excited state conformations in addition to global minima (*vide infra*).

Combinatorial Fragment Assembly. The assembly of structures by TAPERSS occurs in a combinatorial, iterative fashion, assembling every possible structure with the created fragment library, as directed by our combinatorial algorithm (**Figure 3**). This method is briefly described as follows: To assemble any given sequence, DNMPs from the fragment library are attached together iteratively. Each attachment is validated through a two-phase steric clash check (SCC) to ensure there are no stereochemical violations, and a root-mean-square deviation (RMSD) check (*vide infra*). If at any point during construction either test fails, the DNMP responsible for the failure is removed and replaced with a new one. Once a full structure is constructed, we ensure that a closing base pair is present, verifying that the predicted structure is a tetraloop, and proceed to calculate the energy of the structure (see SI, section "Details of Combinatorial Fragment Assembly" and **Pseudocodes S1** and **S2**).

Overlapping Fragments. The most fundamental aspect of the fragment assembly method is the attachment or assembly of the DNMP fragments to create a larger structure. In our method, this is performed using the Kabsch algorithm[53], which calculates the optimal rotation matrix for the minimization of error between two matrices. At any step where a new fragment is being added to an existing model, the 5′ nucleoside of the new fragment is overlapped with the 3′ nucleoside of the previous fragment. Only the base and ribose heavy atoms are included in the calculation of the rotation matrix. After the overlap, root-mean-square deviation (RMSD) between the two coordinate sets is calculated. If the RMSD is below 0.5 Å, then the attachment is accepted and the algorithm moves on, otherwise it is rejected, and a new fragment is selected from the fragment library.

Steric Clash Check. At each successful attachment of a new DNMP fragment, we perform a steric clash check (SCC) to ensure that there are no stereochemical violations imposed by the new fragment. We perform the SCC at each attachment, rather than after complete assembly, so that structures with steric clashes are discarded as soon as possible. This way calculation time is not wasted on completing unphysical structures. Our SCC encompasses two phases: a coarse-grained SCC (CG-SCC) and an atomistic SCC. Our atomistic SCC calculates the distance between every atom in the new nucleotide and any non-neighboring, already assembled, nucleotides. This is, of course, a very time-consuming step and having it repeated at every iteration causes a significant performance bottleneck. Thus, we utilize an opportunistic CG-SCC, which uses the centroids of each nucleotide (ribose, nucleobase, and phosphate group) in the structure to define large van der Waals "beads" which encompass the full nucleotide. Rather than performing an atomistic check immediately, we check the distances between only the centroids initially. If there is overlap between any two beads, then only the nucleotides represented by the overlapping beads will have an atomistic SCC performed. The use of the CG-SCC typically allows for around 50% of the atomistic SCCs to be bypassed, even for the small 6 nucleotide sequences we studied in this work, significantly reducing calculation times.

Energy Calculation and Inclusion of Non-neighboring Hydrogen Bonds. Once a complete structure is generated, the energy for that structure is calculated. We utilize the following energy function to determine the folding energies.

$$E_i = \sum_{j=1}^{N-1} \Delta G_j + \sum_{j=1}^{N-1} \Delta E_{A,j} \qquad \text{Eq. 1}$$

In eq. 1, E_i represents the predicted folding energy of the structure i (in kcal/mol), ΔG_j is the 'relative' folding free energy of the jth DNMP fragment ranging from $j = 1$ to $j = N - 1$, and $\Delta E_{A,j}$ is the energy due to hydrogen bonds formed between the jth DNMP fragment and all the non-neighboring residues. ΔG_j are the energies stored in the fragment library for each DNMP fragment originating from the DPS calculations representing 'relative' folding energies with respect to the global minimum of the investigated DNMP. As a result, ΔG_j are based on the RNA force field utilized in the DPS calculations. We assume that summing the energies of each fragment and the non-neighboring interactions will provide an effective approximation to the true folding energy of the structure, with the quality of this approximation closely dependent on the quality of the RNA force field.

For calculation of $\Delta E_{A,j}$, we assume that the functional groups in the nucleobases will form a favorable hydrogen bond interaction with each other given they are within certain distances. We ensure that only unique interactions are included in this calculation. For the purposes of this study, we define a hydrogen bond as a favorable energy reduction of 1 kcal/mol. Because our fragment library contains no protons, we set the interaction distance cutoff to 3.6 Å as the typical distances between donor and acceptor atoms are around 2.7Å and the bond lengths in O-H and N-H groups are approximately 1Å.

In highly folded structures it is likely that many functional groups will be within proximity, forming a variety of possible hydrogen bonding networks. In our method, we ensure that the maximum number of hydrogen bonds are formed, where any donor or accepter shares a single bond with its counterpart. To identify this maximized bond network, we treat the set of potential networks as a problem in graph theory by mapping the set of positive and negative atoms to a bipartite graph. In this graph, one set of vertices are the positively charged groups, and the other set are the negatively charged groups, and the edges between the vertices represent the potential hydrogen bonding network. With the bipartite graph created, we use the Hopcroft-Karp[54] algorithm to identify the maximum number of matchings, or vertices which share a single edge, and thus the hydrogen bonding network with the greatest number of hydrogen bonds (**Figure 4**).

Tetraloop Filter. To avoid generation of billions of structures which are not of interest, we have implemented a simple filter to detect tetraloops as soon as they are created, writing only those structures to file while all others are discarded. This is done by using a set of model Watson-Crick paired nucleosides as a reference to overlap the 1st and 6th nucleotides of the final structure generated. The Watson-Crick models were created using AMBER LEaP module.[55] If the RMSD between the model Watson-Crick pair and the predicted structure's 1st and 6th nucleotides is less than 0.5 Å, then the structure is accepted as a tetraloop.

Implementation. All the methods discussed regarding our fragment assembly method were implemented in C++ using the GNU Scientific Library (GSL) BLAS libraries. All calculations were performed on FAU's KoKo High Performance Computing (HPC) cluster. The TAPERSS main code was compiled with GCC version 8.3 using the *-ffast-math* and *-march=native* flags. The TAPERSS code is freely available under the GPLv3.0 license. For access, please contact the corresponding author.

Clustering. We performed cluster analyses on both the predicted and experimental structures to determine the unique conformations. An in-house code was utilized for this purpose (**Script S4**). Our clustering method is comparable to the DBSCAN[56] algorithm. Starting with any structure, any other structure which satisfies our RMSD condition is included in an initial cluster. Then the centroid of the initial cluster is calculated, and the RMSD comparison is performed again, adding any new structures which satisfy the RMSD condition with the centroid to the cluster, and removing any which now do not. Then this process of calculating the centroid and updating the cluster is repeated until the cluster "converges", such that the set of elements of the cluster does not change upon recalculation, or more precisely, that the current set is a repeat of any previous iteration. Once a cluster is converged, a new, non-clustered structure is selected as an initial point and the "convergence clustering" process repeats until all points belong to some cluster. This method conceptually finds the densest region local to the starting point of the cluster, as the centroid will "move" towards the center of these dense regions. Additionally, all structures in a cluster satisfy the RMSD condition with the centroid, and all structures will belong to a cluster, although some clusters may contain only a single structure (**Pseudocode S3**). We have used two metrics for our distance criteria. First the structures must have an RMSD of at most 1.5Å with the cluster centroid, with the backbone excluded in the RMSD calculations. Second, every nucleotide must have an RMSD of at most 0.5Å *per nucleoside* with the centroid, meaning the RMSD between centroid and structure is calculated one by one for each residue. The per nucleoside RMSD does include fitting. This was done to limit the torsional variance between structures in the same cluster. This procedure was used for clustering both the predicted and experimental structures.

Calculating Folding Free Energies for Predicted Clusters. While TAPERSS can predict the entire conformational landscape of an RNA structure (given our modelling method), we have investigated only a subsection of this conformational landscape, which represents all the RNA tetraloops one can build using our RNA DNMP fragment library. This complete set can be utilized to calculate first the partition function, and then the probabilities for each structure. Using these probabilities, one can then calculate the folding free energies for each predicted cluster. To determine the energies corresponding to the predicted clusters for each RNA tetraloop system, the partition function is calculated as follows:

$$Z = \sum_{i=1}^{N} \exp(-\Delta E_i / RT) \qquad \text{Eq. 2}$$

In eq. 2, Z is the partition function, ΔE_i (in kcal/mol) is the 'relative' folding energy of predicted structure i with respect to the predicted global minimum as calculated using eq. 1, R is the gas constant, T is the temperature, and N is the total number of RNA tetraloops predicted for a given sequence. Using this partition function, a probability for each predicted structure was calculated using the following equation:

$$p_i = \frac{\exp(-\Delta E_i/RT)}{Z} \qquad \text{Eq. 3}$$

where p_i is the probability assigned to the predicted structure i. Once all the p_i values are determined, probabilities of individual clusters, $p_{C,i}$, where C, i stands for cluster i, can be calculated by summing up all the probabilities of structures observed in C_i as follows:

$$p_{C,i} = \sum_{j \in C_i} p_j = \sum_{j \in C_i} \frac{\exp\left(-\Delta E_j/RT\right)}{Z} \qquad \text{Eq. 4}$$

The cluster with highest probability, $p_{C,0}$, will represent the predicted global minimum cluster. Once all the probabilities are assigned to each cluster, folding free energies of each cluster with respect to global minimum, $\Delta G_{C,i}$, can be determined using the following equation:

$$\Delta G_{C,i} = -RTln\left(\frac{p_{C,i}}{p_{C,0}}\right) \qquad \text{Eq. 5}$$

Benchmarking Predictions to Experiments. The predicted clusters were compared to experimental structures using RMSD as the metric to decide how closely the predictions represent experimental structures. During the RMSD calculations, only the heavy atoms of nucleotides were utilized. The primary benchmark was between our predictions and what we determined to be the experimental structures representing the apo states. The experimental cluster which contained our apo reference structure, described above, is the cluster we use as our primary comparison. All other clusters are used as “excited” states. There are two criteria we use to decide if a prediction is successful. First, we compare the experimental apo states to the predictions; if the RMSD is less than 2Å, we say that the predicted structure represents the apo state. Second, we compare the folding free energies of the predicted clusters with respect to the global minimum. Ideally, the global minimum should represent the apo state, however, practically we classify a successful prediction as one where the free energy difference between the global minimum and predicted cluster is less than 5.1 kcal/mol.

RESULTS and DISCUSSION

TAPERSS generates between 700 million to 2.5 billion valid structures for each RNA hexamer. 21 RNA hexamer sequences with known experimental apo-states were used as benchmarks for TAPERSS. Depending on the sequence, TAPERSS generated 700 million to 2.5 billion valid structures for each RNA hexamer, with roughly 1 in every 100,000 structures being in tetraloop form. Run times for each calculation varied proportionally to the total number of structures predicted, but generally fell within 3 to 8 hours of total run time (**Table S3**).

The global minimum structures of RNA hexamers predicted by TAPERSS are linear. TAPERSS utilizes our RNA DNMP fragment library, which was created from DPS data predicted by the RNA amber99 force field with revised χ[49] and α/γ[45] parameters. The energy landscapes predicted by DPS calculations display global minima in A-form-like orientations. As a result, linear and A-form-like structures are hypothetically expected to be the global minimum for short RNA sequences. This is indeed the case for the 21 RNA hexamer systems we investigated (**Figures S22-S42)**. For the purposes of this study, the linear structures produced are not of interest but do indicate that our energy function, eq. 1, does generate structures with expected energies for the short 6nt sequences studied. We maintain that this is realistic as the 21 RNA hexamers studied here are not long enough to remain in hairpin-like states due to the torsional strains imposed on the RNA backbone. As a result, we do not expect TAPERSS to yield low energies to RNA hexamers in folded, hairpin-like states. This agrees with experimental investigations of short RNA structures, four to six nucleotides in length, which show that these sequences generally prefer A-form like conformations.[57-62] A recent investigation into the conformational ensembles of RNA UCAAUC oligonucleotide generated by FARFAR2 compared to MD simulations, NMR and SAXS show that FARFAR2 was unable to produce the experimentally confirmed A-form like structures, rather it generated highly folded structures with several intramolecular interactions.[60] For an RNA sequence to fold into an RNA hairpin motif, stabilization from other interactions is necessary to maintain the bent hairpin. Specifically, the stability gained due to formation of Watson-Crick base pairs in the stem regions of longer RNA systems will overcome the ‘unfavorable’ bending observed in RNA hairpins. As an example, to our knowledge the shortest RNA sequence which can fold and form an RNA hairpin motif is 5′-GGG<u>CGUG</u>CCC-3′, which is a 10 nucleotide long sequence forming three GC base pairs with the tetraloop (underlined) in a folded state.[63] For a rough approximation, if we assume a hydrogen bond counts for approximately 1-2 kcal/mol, the Watson-Crick pairs in the listed sequence will impose an energy reduction of 9-18 kcal/mol, which is hypothetically enough to compensate bending at the hairpin site. This simple approximation also falls in line with the predictions made by TAPERSS, as the difference in folding free energy between the true global minimum and the tetraloop minimum is between 8.5 and 11.4 kcal/mol (**Figures S22-S42**). Thus, if TAPERSS is able to correctly predict a tetraloop as a global minimum within a subset tetraloop landscape, we can assume that additional stabilization by a larger stem region would allow for the TAPERSS predicted tetraloop to be the true global minimum structure. This may also provide some indication of its future performance with probabilistic sampling methods, such as Monte Carlo algorithms.

Experimental structures of RNA tetraloops extracted from literature include apo and excited states. To identify the unique conformations for each sequence in our dataset of experimental structures, we clustered the experimental structures using the same procedure we performed on predicted structures. This analysis enabled an approximate separation of apo from non-apo states, by designating the cluster which contained our identified apo state structure as the apo cluster, and all others as “excited states”. These excited states typically arise from experimental structures where the tetraloop is interacting with non-neighboring nucleotides and/or proteins in a highly folded state, or from other crystal packing contacts distorting the RNA structure.[64] As described in the methods section, our identified apo state came preferentially from NMR structures, ensuring that there were no apparent interactions between non-tetraloop nucleotides. 15 out of 21 sequences we have studied had at least one experimental cluster classified as an excited state. Due to the restrictive clustering cutoffs, there are some clusters which we technically classify as “excited”, as they do not contain our identified apo structure, but can arguably be considered apo as their differences with the apo cluster are minor (**Figures S43-S57**). For example, sequences GGUGAC, UGAAAG and UGCAAG, have multiple clusters identified as unique, however they may not necessarily be considered distinct from the apo cluster as the RMSD between any two cluster averages (for the same sequence) is always less than 2, for these 3 systems specifically (**Figures S49, S51, and S52**). Other excited states which do not represent the apo state are largely states distorted by crystal packing contacts. Predictions of these excited states are further discussed below.

With the exception of CUACGG, TAPERSS predicts the experimental apo states among the 10 lowest energy structures. The focus of this study is to validate the effectiveness of TAPERSS in predicting apo states of RNA tetraloops. Structural comparisons demonstrate that TAPERSS was able to predict all the 21 RNA tetraloops with RMSD values less than 2Å (**Tables 1** and **2,** and **Figure 5)**. However, the predicted folding free energy of these structures shows that they are not all global minimum structures within the tetraloop landscape (**Table 1**). Although the predicted RNA tetraloops have significantly higher folding energies than the A-form states, which are true global minimum structures according to TAPERSS, we still place importance on the ability of TAPERSS to generate minima which represent the apo state for RNA tetraloops. Being able to predict experimental apo states with folding free energies either as global minima or energetically close to them indicates that our RNA DNMP model and its extension to 3D RNA structure prediction is indeed realistic. As was discussed, experimental RNA tetraloops are only possible due to stabilization by Watson-Crick base pairs in the stem region, which overcomes the unfavorable bending observed in hairpins. Nevertheless, the conformation of the hairpin loop, in isolation of the stem region, is still in the energetically most favorable orientation due to stabilizing interactions within the hairpin. Being able to predict these experimental RNA tetraloop motifs with relatively low folding free energies is a promising result and TAPERSS is in fact able to predict the experimental apo states within the 10 lowest energy clusters, excluding CUACGG (**Table 2**). Indeed, 13 of the 21 RNA tetraloop motifs were predicted as either global minimum or the first excited state (**Table 1**). Furthermore, in many of the sequences tested, several of the predicted lowest energy clusters display the apo state (**Table 2 and Figures S58 to S78**). For example, in AGAAAU seven out of the 10 lowest energy clusters display the apo structure. However, this does not hold true for all sequences, which we discuss in the subsequent sections.

The global minima of AGAAAU, CGAAAG, CGAGAG, CGCAAG, UGAAAG, and UGAGAG predicted by TAPERSS perfectly overlap with experimental apo structures. There were 6 sequences we tested which had what we will categorize as ideal predictions, in which the global minimum structure successfully predicted the apo state (**Table 1** and **Figure 5ACDEKL**). Additionally, at least 3 out of the 10 lowest energy clusters predicted for these sequences display structures similar to the experimental apo state (**Table 2**). All the sequences which fall into this category are GNRA tetraloops. GNRA tetraloops, the dominant tetraloop found in the RNA world, are well understood and known to be particularly stable due to having all loop nucleotides in stacked conformations, as well as a common G-A sheared base-pair.[65-67] These highly stacked conformations may be an explanation for why TAPERSS performs well in predicting the structures of these sequences, as our fragment library exhibits the properties of the RNA-IL force field, which generally favors stacked conformations and more accurately describes purine-purine DNMPs than others.[68]

The first excited states of AGUGAU, CGGAAG, CGUAAG, GGAAAC, UGAAAA, UGUGAA, and CUAACG predicted by TAPERSS represent the experimental apo states. The 3D structures of RNA AGUGAU, CGGAAG, CGUAAG, GGAAAC, UGAAAA, UGUGAA, and CUAACG hairpins are predicted by TAPERSS as the first excited states with relative folding free energies less than 3.6 kcal/mol with respect to the global minima (**Table 1**). Like the 'ideal' predictions, these sequences are mostly GNRA tetraloops, excluding CUAACG, a UMAC (M = A or C) tetraloop known to exhibit GNRA-like conformations and stabilities.[69] Although they are not the global minima, the predicted structures have very low folding free energies, and in the cases of AGUGAU, CGGAAG, CUAACG, and UGUGAA the energies are 0.5 kcal/mol higher than the global minima. As with the previous sections, all apo structures in this category have stacked nucleotides, likely contributing to the success of the predictions (**Figure 5BFGHJNO**).

The excited states of GGUGAC, CUCACG, CCUCGG, CCUUGG, and CUUCGG, which have relative folding free energies in the ranges of 3.6 and 5.1 kcal/mol higher than global minima represent the experimental apo states. Less than ideal, but still well performing, the experimental apo states for these sequences were predicted by TAPERSS as one of the 10 lowest minima, but had comparatively higher energies than those previously discussed (**Tables 1** and **2,** and **Figure 5IPQRT**). These sequences are composed of mostly pyrimidines, excluding one GNRA sequence, GGUGAC. In sequences CCUCGG, CCUUGG, (CUYG tetraloops) and CUUCGG (UNCG tetraloop), there are loop nucleotides interacting with solution, making these sequences more challenging for TAPERSS. CUYG is another set of tetraloops commonly found in the RNA world,[67] which are much more dynamic than GNRA tetraloops, especially when a second loop U is present, as is the case in CCUUGG.[70-71] UNCG tetraloops are also commonly found and, despite having dynamic loop nucleotides, they are known to be thermodynamically stable. This stability is attributed to hydrogen bonding between the U and G nucleotides within the tetraloop,[72] as well as solvent interactions.[66] TAPERSS incorrectly predicts the global minimum of these sequences to be in highly stacked states. For example, the global minimum structure TAPERSS predicted for CCUCGG (**Figure S74A**) has its 2nd loop nucleotide, U, stacked with its 3rd loop nucleotide, C. The experimental apo state shows that the 2nd loop nucleotide is instead unstacked interacting with solution (**Figure S17A** and **S74G**). Although the apo state predictions are not the global minimum, these results are still quite promising as the native state conformation is still observed in the lowest energy clusters with fairly close relative folding energies with respect to global minima (**Table 1**).

Poor predictions are observed in UGCAAG, CUACGG, and GUUCGC, where the predicted structures representing experimental apo states have folding free energies ranging from 6.5 to 8.8 kcal/mol higher than the predicted global minima. These are the worst predictions made by TAPERSS in this study, as the respective apo structures were predicted to have comparatively high energies. In the worst case, the apo state of CUACGG was predicted as the 35th excited state (**Table 1** and **Figure 5S**). For CUACGG and GUUCGC, which are UNCG tetraloops, the 2nd and 3rd loop nucleotides both interact with solution and have very similar conformations (**Figures Y5SU**). Despite the current fragment library's limited ability to accurately predict the energetics of these UNCG structures, there is potential for their recovery through enhancements in RNA force fields and advancements in our electrostatic energy calculations (*vide infra*). Perhaps the most unique sequence with regards to performance is UGCAAG, a GNRA tetraloop whose apo structure is standard in comparison to the other GNRA loops studied, however it could not be predicted by a low folding free energy (**Figure 5M**).

TAPERSS can predict the experimentally observed excited states. In addition to predicting the apo states for RNA tetraloops, which were generally successful, we also investigated the ability of TAPERSS to predict non-apo states available for the sequences we studied. We consider the experimental clusters not resembling the apo states as "excited states", which in some cases may be a misleading description as previously discussed (**Figure S43-S57**). Nonetheless, when available

TAPERSS was able to predict many of the excited states for all relevant sequences tested. Structures likely distorted by crystal packing forces and/or non-neighboring residues usually had several unstacked nucleotides and deviated significantly from the apo state. These distorted structures were often predicted with higher energies, which is in line with experimental data. Structures which TAPERSS failed to predict were often those with nucleotides in syn orientations being significantly distorted by contacts, for example in GGAAAC (**Figure S8A,B,I,L,O,P,Q,R,U,X,AA**). Occasionally, there are experimental structures that undergo distortion due to crystal packing forces or are buried loops, yet they contain several stacked nucleotides. This is exemplified by CGAGAG as observed in 3CPW[73] (**Figure S4D**) and CGCAAG as observed in 4M2Z[74] (**Figure S5F**) and 5WNQ[75] (**Figure S5H**). These non-apo structures are observed to have low folding free energies despite having external contacts, with the example in CGAGAG having a folding free energy of 0.5 kcal/mol (**Figure S4D**) and the examples in CGCAAG having relative folding free energies of 4 kcal/mol (**Figure S5H**) and 0.2 kcal/mol (**Figure S5F**). The 4.0 kcal/mol CGCAAG structure (**Figure S5H**) is, at least visually, resemblant of the apo state of typical GNRA loops. While TAPERSS correctly predicts the true apo state for CGCAAG (**Figure S4A**), it assigns relatively low folding energies to these distorted states, which do not resemble the apo structure. This suggests that these low-energy non-apo structures may represent real metastable states, which are rarely observed experimentally but are predicted by TAPERSS as local minima for the given sequence.

Limitations and potential improvements to TAPERSS. As was noted several times above, TAPERSS performs well with highly stacked GNRA-like RNA tetraloops but may struggle with more dynamic ones. The RNA force field which was used to create the RNA DNMP fragment library and the types of interactions considered in our energy calculations are the two main factors we suspect are the culprits. There is an association between the success of our predictions and the composition of the tetraloops studied. Specifically, purine-rich sequences containing over 4 purines typically had the apo states predicted by one of the three lowest energy clusters, while pyrimidine-rich sequences often had the apo states predicted with higher energies (**Table 1** and **Figure 5**). This clearly coincides with the fact that purine rich GNRA tetraloops prefer stacked states, while pyrimidine rich CUYG and UNCG tetraloops often take more dynamic, unstacked conformations. This result aligns with our earlier investigations, which also highlighted the deficiencies in the abilities of RNA-IL and ff99OL3 RNA force fields to accurately predict the stacking thermodynamics of various RNA DNMPs, especially those involving RNA UU DNMPs.[68] Hence, a clear weakness with the current version of TAPERSS is its propensity to favor stacked states. A common low folding free energy motif predicted by TAPERSS is a fully stacked loop section, in which all loop nucleotides are stacked on top of one another almost forming an A-form-like tetraloop (**Figures S58CI**, **S60HJ**, **S61B**, **S62BCG, S63GH, S64D, S66B,**

S67AEH, S68BI, S69CDF, S70ADF, S71CEJ, S73C, S74B, S75AD). While this motif is sometimes experimentally observed, as in CGCAAG (**Figure S4D**), it is seemingly over predicted by TAPERSS. It commonly appears as at least one of the ten lowest folding free energy clusters in many of the sequences tested (**Figure S58-S78**). The direct answer to this flaw is of course to improve the underlying RNA force field to correct the well-known stacking inaccuracies.

However, there are tetraloops studied here which may have benefited from a more robust method of detecting electrostatic interactions. For example, in GNRA tetraloops, which TAPERSS generally perform well, there is a sheared G-A pair known to occur. This interaction does not factor in TAPERSS, as the O2′ group involved in sheared base-paring[76] is currently not considered a charged group available for hydrogen bonding. We do not use any functional groups outside of those used in canonical Watson-Crick pairing for detection of hydrogen bonds. Improvements to the energy function, thus, can potentially improve poorly predicted structures despite force field weaknesses. To achieve further improvement, modifying the Hopcroft-Karp algorithm[54] to consider hydrogen bond distances instead of solely maximizing the total number of hydrogen bonds formed may be necessary. With our current model, the energy reduction imposed is equivalent, regardless of the distance. However, future iterations of TAPERSS may utilize a more nuanced energy function where a maximum weight matching algorithm will be more appropriate.[77] Beyond improvements to the energy function and RNA force field, a potential performance improvement may be possible by changing our overlap method. Although the overlap calculation is not the most significant performance bottleneck, future iterations of this method may utilize the quaternion-based characteristic polynomial (QCP) method[78] for minimizing the error between two matrices, which has a significantly lower computational cost compared to the Kabsch method.[53]

CONCLUSION.

We have developed a physics-based fragment assembly method, TAPERSS, an acronym for Theoretical Analyses, Prediction, and Evaluation of RNA Structures from Sequence, which aims to predict the three-dimensional structures of RNA from sequence with energies derived from physical principles. The physics-based nature of this method arises from the use of an RNA DNMP library created using DPS. We have benchmarked our RNA model using a combinatorial method of fragment assembly, which, while highly limiting the maximum length sequence we can feasibly study, has enabled a complete scan of the accessible conformational landscape given our library. We have discussed the combinatorial method and provided two major algorithms, the course-grain steric clash check and hydrogen-bond maximization algorithms, which should prove useful for future iterations of this method. We have benchmarked TAPERSS on 21 RNA tetraloop sequences, each with a single closing base pair. TAPERSS predicted A-form-like linear structures to be the ‘true’ global minimum for these sequences, which falls within our expectations for the small systems we studied. TAPERSS has also shown the ability to correctly predict the experimental RNA tetraloop structures with realistic folding free energies. We have found that TAPERSS generally performs well energetically when predicting the apo structures of highly stacked tetraloops, as observed in GNRA tetraloops, but can have difficulty in predicting structures of tetraloops which contain dynamic nucleotides, such as CUYG tetraloops. We discuss that the likely reason for this discrepancy arises from the RNA force field, RNA-IL, which was used to generate the RNA DNMP fragment library. RNA-IL is known to have weaknesses in predicting populations of stacked state accurately particularly in purine-pyrimidine, pyrimidine-purine, and UU DNMPs. Although the effect of the RNA force field on our predictions is apparent, we also note that a more nuanced and robust energy function could improve predictions by capturing interactions, such as sheared base-pairing, which stabilize the dynamic tetraloops TAPERSS struggles with. The presented version of TAPERSS may act as a useful tool for any detailed investigation of RNA tetraloops, and with the suggested improvements as well as implementation of probability-based sampling, we believe TAPERSS can be become an effective method for physics-based sampling of RNA's conformational landscape.

ASSOCIATED CONTENT

Supporting Information. Details of combinatorial fragment assembly; pseudocodes used in combinatorial algorithm; scripts utilized to compare structures, perform cluster analyses, and extract fragments; PDB ids of RNA hairpins; RNA tetraloops benchmarked; details of the predictions for each sequence; extended version of **Table 2**; comparison between all experimental clusters and predictions; comparison of true global minima to predicted tetraloops; 2D RMSD matrix for experimental clusters; 10 lowest energy clusters predicted.

Acknowledgments. We thank Dr. Kye Won Wang for valuable discussions regarding the computational approach we developed, and Dr. Andy W. C. Lau from Florida Atlantic University for helpful suggestions. Computations were performed using the High-Performance Computing (HPC) cluster, KoKo, at the Florida Atlantic University. This work was supported by the Florida Atlantic University startup grant (IY), by the David and Lynn Center for Degenerative Disease Research program (IY), and by the NIH grant R15GM146199 (IY).

Table 1. Results comparing performance of TAPERSS to experimental data for 21 RNA tetraloop sequences.

Tetraloop Type	Sequence[a]	Lowest Energy Apo Prediction (LEAP) Cluster			Experimental Clusters	
		ΔE_C Ranking[b]	ΔE_C[c] (kcal/mol)	RMSD[d] (Å)	Count[e]	% Predicted[f]
GNRA	AGAAAU	A	0	1.42	1	100.00%
	AGUGAU	B	0.45	1.18	1	100.00%
	CGAAAG	A	0	0.74	20	60.00%
	CGAGAG	A	0	0.93	6	83.33%
	CGCAAG	A	0	1.09	15	93.33%
	CGGAAG	B	0.31	1.06	3	100.00%
	CGUAAG	B	2.5	1.05	7	100.00%
	GGAAAC	B	3.63	0.89	27	59.26%
	GGUGAC	C	4.47	0.99	3	100.00%
	UGAAAA	B	2.69	1.55	11	45.45%
	UGAAAG	A	0	1.01	2	100.00%
	UGAGAG	A	0	0.86	1	100.00%
	UGCAAG	H	6.49~~5~~	1.78	2	100.00%
	UGUGAA	B	0.25	1.46	1	100.00%
UMAC	CUAACG	B	0.28	1.12	1	100.00%
	CUCACG	E	4.97	1.45	1	100.00%
CUYG	CCUCGG	G	3.72	0.94	7	85.71%
	CCUUGG	E	5.09	1.13	5	100.00%
UNCG	CUACGG	X[g]	8.77	0.66	5	100.00%
	CUUCGG	F	3.63	1.84	55	90.91%
	GUUCGC	H	7.39	0.79	8	62.50%

[a] The sequence of the RNA tetraloop. [b] The ranking of the predicted apo state within the predicted cluster sets. For example, A means the predicted apo state is the global minimum while B means the predicted apo state is observed as first excited state. [c] The energy difference between the predicted apo state and the global minima found. For example, if ranking is 0, then ΔE_C=0 by definition as the predicted apo state is already the global minimum. [d] The RMSD of the predicted apo structure with respect to the experimental apo state (**Figure 5**). [e] The number of experimental clusters observed for a sequence. If the count is 1, it implies that experimental data does not have any excited but apo state for that studied sequence. [f] Percentage of experimental clusters predicted by TAPERSS. 100% means that all the experimentally observed structures are predicted. [g] X represents the 35th excited state.

Table 2. RMSD (in Å) comparison of experimental apo states to predicted 10 lowest energy clusters for each sequence.

Sequence[a]	ΔE_C Ranking[b]									
	A	B	C	D	E	F	G	H	I	J
AGAAAU[d]	**1.42**	**0.76**	3.04	3.46	**1.76**	**1.36**	**0.89**	**1.12**	3.44	**1.57**
AGUGAU[e]	2.78	**1.18**	2.64	2.98	3.60	2.47	3.13	3.22	2.11	2.90
CGAAAG[f]	**0.74**	3.47	**1.35**	**1.00**	3.82	**1.17**	3.77	3.55	2.00[c]	3.57
CGAGAG[g]	**0.93**	3.29[c]	**1.37**	**1.19**	3.19	**1.16**	**1.50**	4.26	**1.60**	2.41
CGCAAG[h]	**1.09**	3.86[c]	3.96[c]	4.07	2.68[c]	**1.26**	4.37[c]	2.85[c]	4.35	**1.73**
CGGAAG[i]	2.64	**1.06**	3.89	3.51	3.55	3.27	3.80	3.62	**1.09**	3.32
CGUAAG[j]	2.40	**1.05**	2.79	3.81	**1.23**	3.35	3.42	2.41	2.74	3.26
GGAAAC[k]	3.68	**0.89**	3.06	**1.51**	3.13	3.47[c]	**1.12**	2.86	3.74	2.71
GGUGAC[l]	2.82	2.86	**0.99**	2.98	2.38	2.05[c]	2.39	**1.58**	3.79	3.77
UGAAAA[m]	3.24	**1.55**	3.29	**1.17**	3.45	3.78	3.05	3.96	**1.13**	**1.25**
UGAAAG[n]	**1.01**	3.14	**1.48**	4.81	2.07[c]	4.73	**1.30**	**0.90**	3.21	**1.83**
UGAGAG[o]	**0.86**	**1.47**	3.33	3.44	**1.30**	3.38	3.34	4.38	4.23	4.13
UGCAAG[p]	3.15	2.55	3.40	3.61	2.63	3.25	2.90	**1.78**	**1.04**	3.74
UGUGAA[q]	2.91	**1.46**	3.54	3.60	3.40	**1.61**	2.60	3.88	2.57	4.19
CUAACG[r]	2.73	**1.12**	**0.91**	**1.39**	2.54	4.15	2.77	3.52	3.34	3.15
CUCACG[s]	2.57	3.26	3.48	3.61	**1.45**	**0.97**	3.76	3.68	2.85	**1.83**
CCUCGG[t]	3.14	3.27	2.90	2.88	4.28	2.62	**0.94**	4.57	3.06	3.89
CCUUGG[u]	2.67	2.41	2.09	2.94	**1.13**	2.75	2.18	2.84	2.57	2.71
CUACGG[v]	4.01	3.68	4.59	3.36	2.80	4.78	3.73	4.72	3.71	3.73
CUUCGG[w]	4.55[c]	2.67[c]	3.52[c]	3.53[c]	3.01[c]	**1.84**	2.89	4.04	2.84	**0.77**
GUUCGC[x]	3.52	2.75	3.51	2.60	3.90	4.03	3.75	**0.79**	4.72	3.52

[a] Sequence of the RNA tetraloop. [b] RMSD (in Å) with respect to the experimental apo state for each predicted cluster. Note that ranking A represents the predicted global minima, while the rest represents excited states arranged in ascending order. Values highlighted bold displays predicted clusters representing the experimental apo states. Compare to **Table 1**. [c] Represents structures comparable to experimentally observed excited states (see **Table S4** for details). [d] See **Figure S44**. [e] See **Figure S45**. [f] See **Figure S46**. [g] See **Figure S47**. [h] See **Figure S48**. [i] See **Figure S49**. [j] See **Figure S50**. [k] See **Figure S51**. [l] See **Figure S52**. [m] See **Figure S53**. [n] See **Figure S54**. [o] See **Figure S55**. [p] See **Figure S56**. [q] See **Figure S57**. [r] See **Figure S58**. [s] See **Figure S59**. [t] See **Figure S60**. [u] See **Figure S61**. [v] See **Figure S62**. [w] See **Figure S63**. [x] See **Figure S64**.

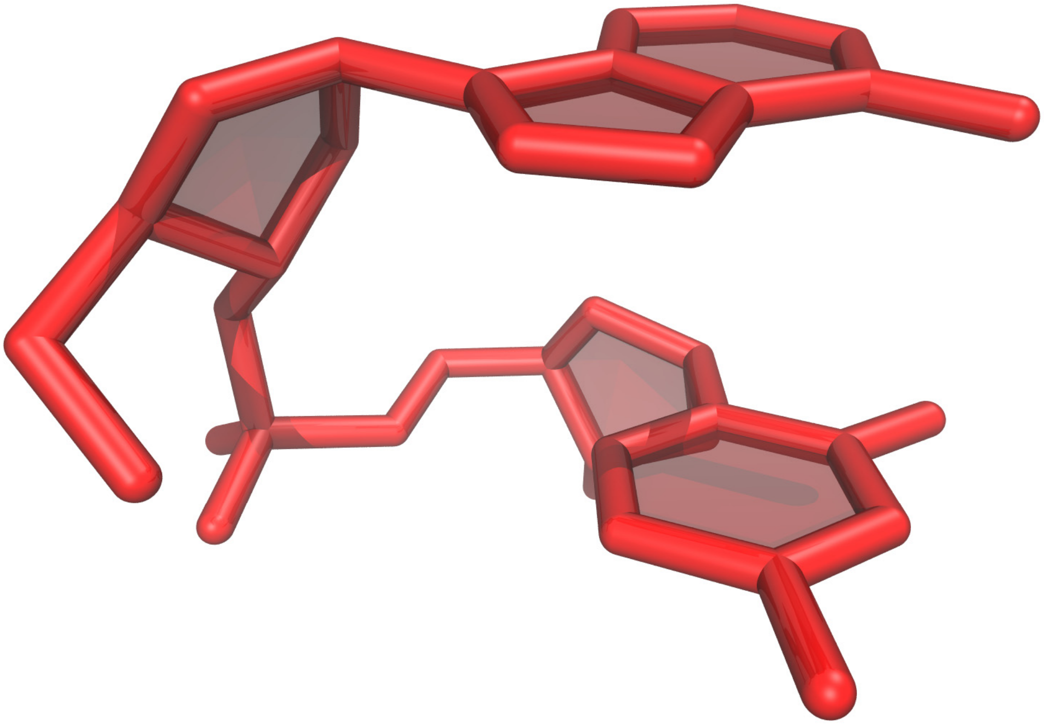

Figure 1. A model RNA dinucleoside (AC) monophosphate (DNMP). The fragment library created and utilized in TAPERSS includes 16 such RNA DNMPs.

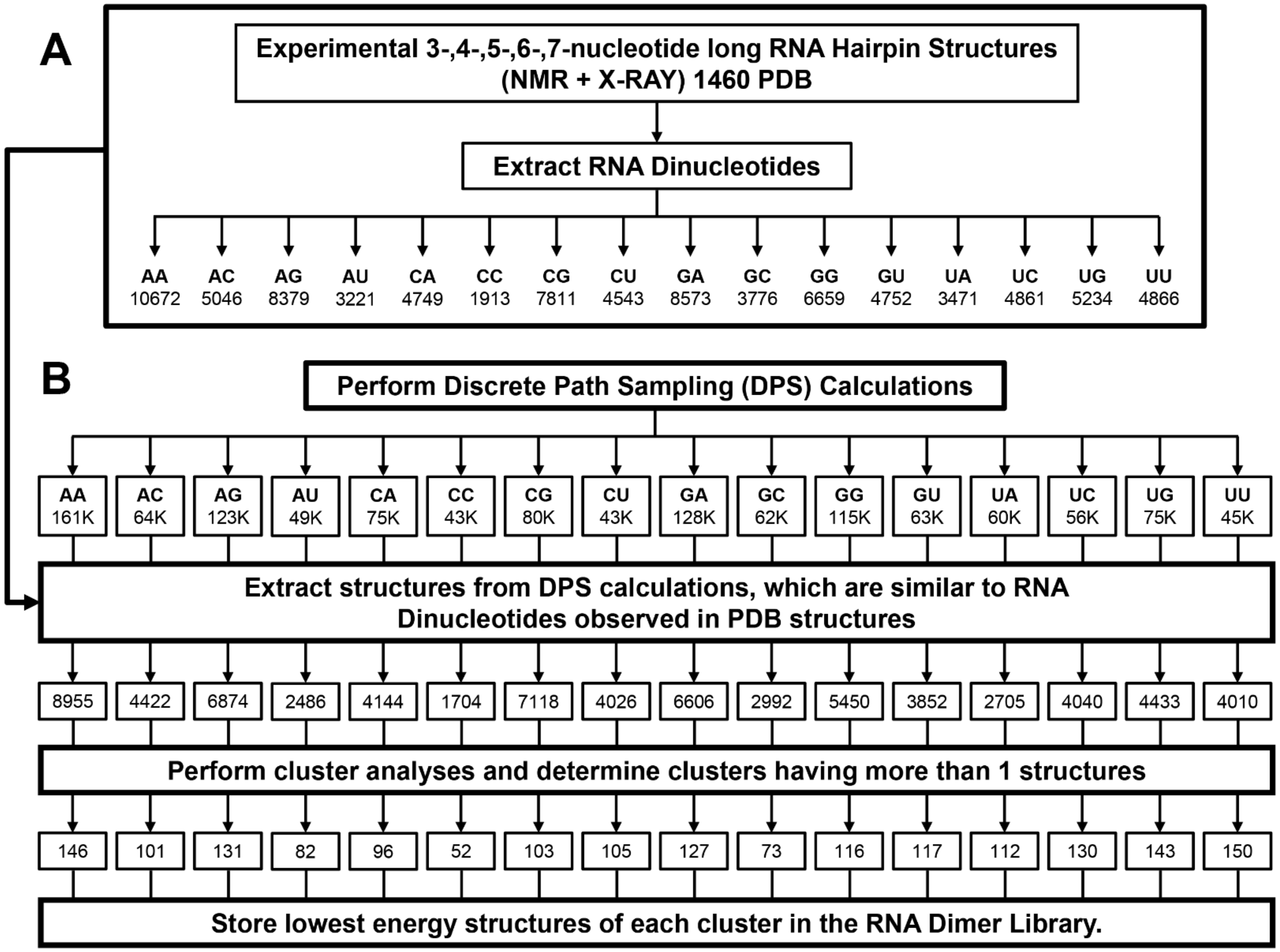
A
Experimental 3-,4-,5-,6-,7-nucleotide long RNA Hairpin Structures (NMR + X-RAY) 1460 PDB
Extract RNA Dinucleotides
AA 10672
AC 5046
AG 8379
AU 3221
CA 4749
CC 1913
CG 7811
CU 4543
GA 8573
GC 3776
GG 6659
GU 4752
UA 3471
UC 4861
UG 5234
UU 4866
B
Perform Discrete Path Sampling (DPS) Calculations
AA 161K
AC 64K
AG 123K
AU 49K
CA 75K
CC 43K
CG 80K
CU 43K
GA 128K
GC 62K
GG 115K
GU 63K
UA 60K
UC 56K
UG 75K
UU 45K
Extract structures from DPS calculations, which are similar to RNA Dinucleotides observed in PDB structures
8955
4422
6874
2486
4144
1704
7118
4026
6606
2992
5450
3852
2705
4040
4433
4010
Perform cluster analyses and determine clusters having more than 1 structures
146
101
131
82
96
52
103
105
127
73
116
117
112
130
143
150
Store lowest energy structures of each cluster in the RNA Dimer Library.

Figure 2. Steps used to create the RNA DNMP Fragment Library. (**A**) Experimental RNA hairpin structures are used to extract the RNA dinucleotides observed in experimental database. (**B**) DPS calculations are utilized to determine the conformational landscapes of all 16 RNA DNMPs. Results of **A** are then compared to **B** to determine structures in DPS calculations representing the experimentally observed conformations for all the RNA DNMPs (for details see Methods).

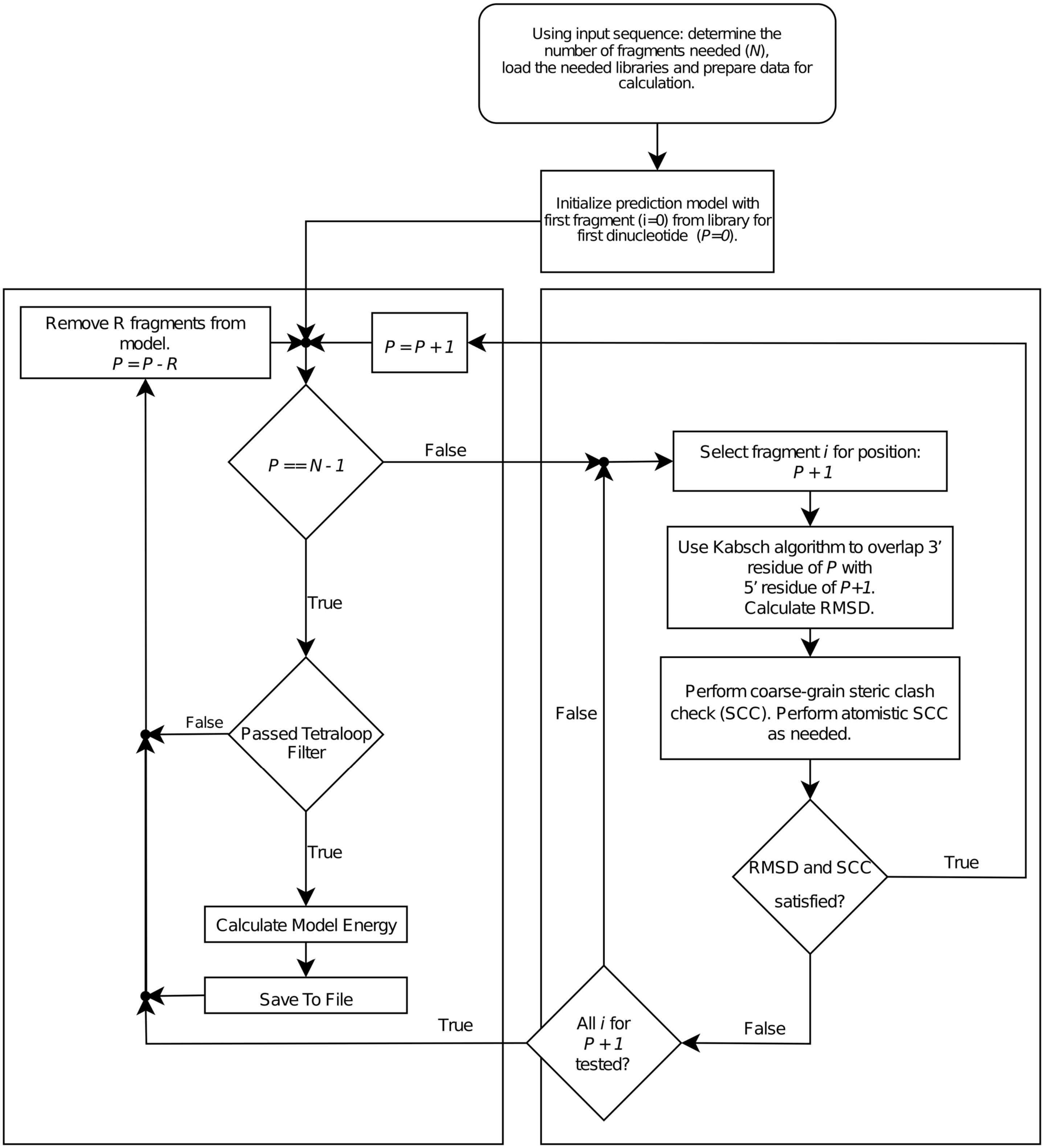
Using input sequence: determine the number of fragments needed (*N*), load the needed libraries and prepare data for calculation.
Initialize prediction model with first fragment (i=0) from library for first dinucleotide (*P=0*).
Remove R fragments from model. *P = P - R*
P = P + 1
P == N - 1
False
True
Select fragment *i* for position: *P + 1*
Use Kabsch algorithm to overlap 3' residue of *P* with 5' residue of *P+1*. Calculate RMSD.
Perform coarse-grain steric clash check (SCC). Perform atomistic SCC as needed.
RMSD and SCC satisfied?
True
False
False
Passed Tetraloop Filter
True
Calculate Model Energy
Save To File
True
All *i* for *P + 1* tested?
False

Figure 3. Workflow diagram of combinatorial algorithm used in TAPERSS. Determination of R is described in SI.

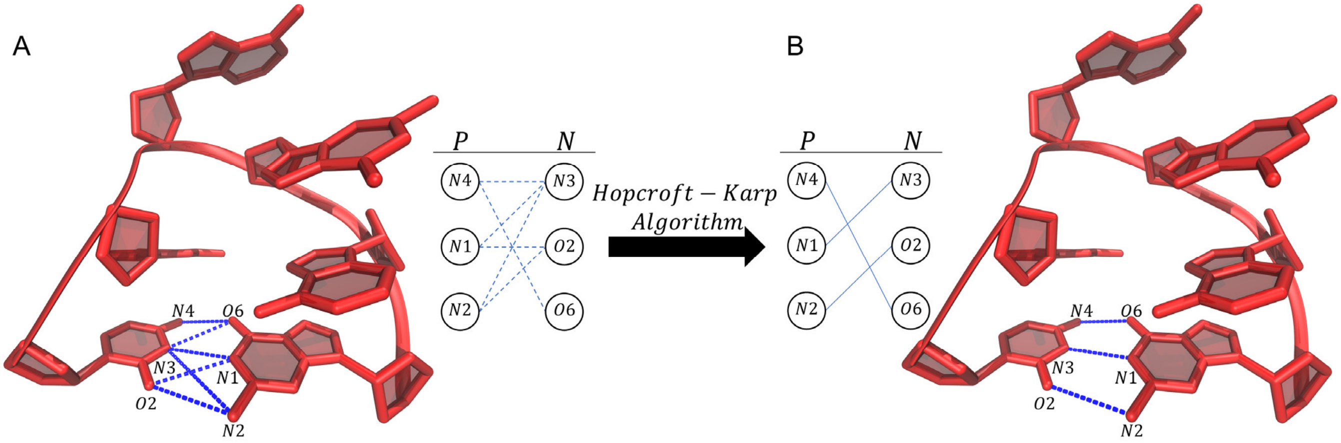

Figure 4. Visualization of the Hopcroft-Karp algorithm used in TAPERSS to maximize total number of non-neighboring interactions in a predicted structure, shown for sequence CGAGAG. **A** and **B** represent non-neighboring interactions before and after the maximization process, respectively. Dashed blue lines in **A** indicate potential non-neighboring interactions. Adjacent to the shown structure is a visualization of the bipartite graph. The **P** column vertices represent the positively charged or acceptor groups, while the **N** column vertices represent the negatively charged or donor groups. Note that each pair of functional groups (donor/acceptor) can exhibit only one non-neighboring interaction.

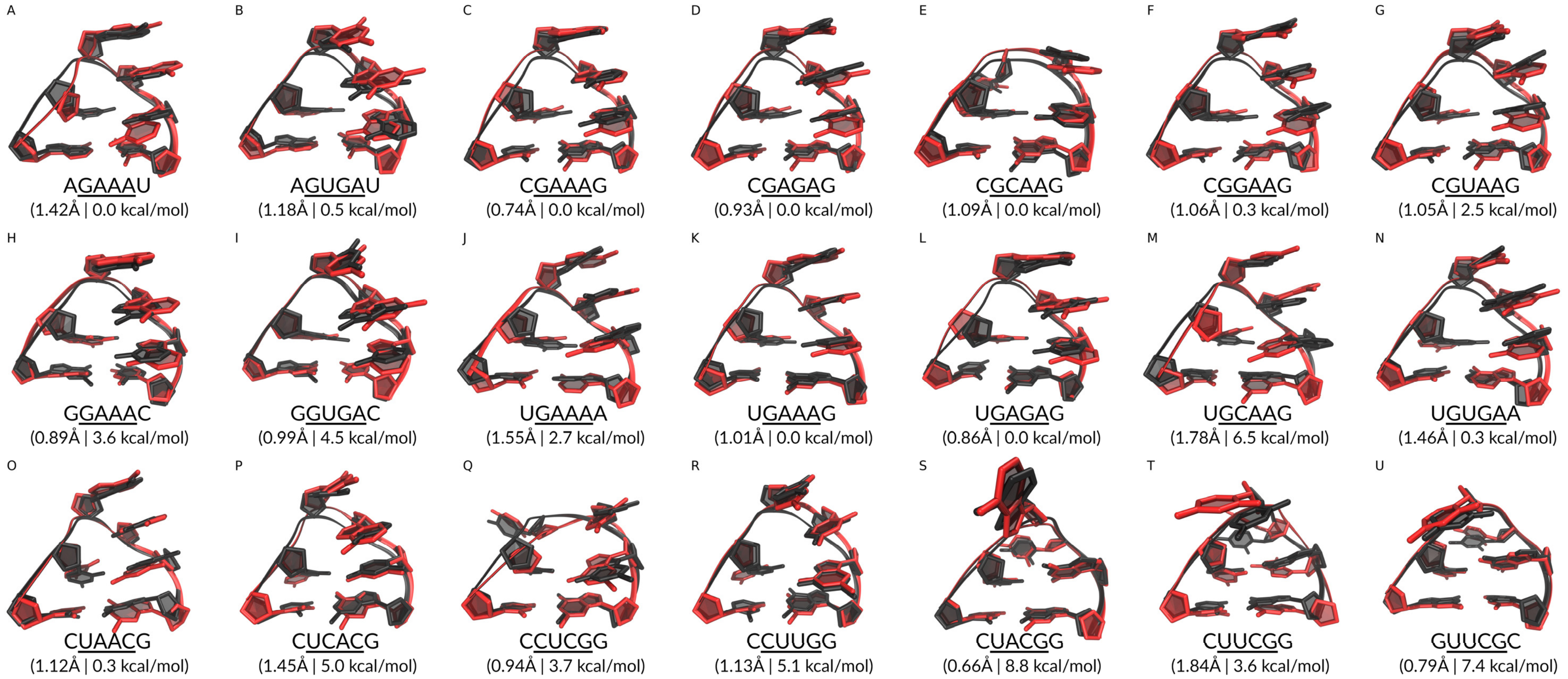

Figure 5. Overlap of Experimental Apo State with Predicted Clusters for 21 Tested Sequences (**A** to **U**). The figure illustrates the overlap of the experimental apo states (black) with the predicted clusters (red), which closely resembles the apo state, for all 21 tested sequences (**Table 1**). Structures represent cluster averages. Predicted structures display the lowest energy apo prediction (LEAP) cluster. Average structures are displayed after minimization. Each structure is accompanied by the sequence name, the root-mean-square deviation (RMSD) between the black and red structures, and the relative 'folding' free energy of the predicted cluster with respect to global minima presented as (RMSD | ENERGY) (**Table 1**).

References:

1. Ganser, L. R.; Kelly, M. L.; Herschlag, D.; Al-Hashimi, H. M., The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.* **2019,** *20* (8), 474-489.
2. Hombach, S.; Kretz, M., Non-coding RNAs: Classification, Biology and Functioning. *Adv. Exp. Med. Biol.* **2016,** *937*, 3-17.
3. Schärfen, L.; Neugebauer, K. M., Transcription Regulation Through Nascent RNA Folding. *J. Mol. Biol.* **2021,** *433* (14), 166975.
4. Childs-Disney, J. L.; Yang, X.; Gibaut, Q. M. R.; Tong, Y.; Batey, R. T.; Disney, M. D., Targeting RNA structures with small molecules. *Nat. Rev. Drug Discovery* **2022,** *21* (10), 736-762.
5. Warner, K. D.; Hajdin, C. E.; Weeks, K. M., Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discovery* **2018,** *17* (8), 547-558.
6. Meyer, S. M.; Williams, C. C.; Akahori, Y.; Tanaka, T.; Aikawa, H.; Tong, Y.; Childs-Disney, J. L.; Disney, M. D., Small molecule recognition of disease-relevant RNA structures. *Chem. Soc. Rev.* **2020,** *49* (19), 7167-7199.
7. Chen, J. L.; Zhang, P. Y.; Abe, M.; Aikawa, H.; Zhang, L. Y.; Frank, A. J.; Zembryski, T.; Hubbs, C.; Park, H.; Withka, J.; Steppan, C.; Rogers, L.; Cabral, S.; Pettersson, M.; Wager, T. T.; Fountain, M. A.; Rumbaugh, G.; Childs-Disney, J. L.; Disney, M. D., Design, Optimization, and Study of Small Molecules That Target Tau Pre-mRNA and Affect Splicing. *J. Am. Chem. Soc.* **2020,** *142* (19), 8706-8727.
8. Luo, Y.; Disney, M. D., Bottom-up design of small molecules that stimulate exon 10 skipping in mutant MAPT pre-mRNA. *ChemBioChem* **2014,** *15* (14), 2041-4.
9. Chen, J. L.; Moss, W. N.; Spencer, A.; Zhang, P.; Childs-Disney, J. L.; Disney, M. D., The RNA encoding the microtubule-associated protein tau has extensive structure that affects its biology. *PLoS One* **2019,** *14* (7), e0219210.
10. Campagne, S.; Boigner, S.; Rüdisser, S.; Moursy, A.; Gillioz, L.; Knörlein, A.; Hall, J.; Ratni, H.; Cléry, A.; Allain, F. H. T., Structural basis of a small molecule targeting RNA for a specific splicing correction. *Nat. Chem. Biol.* **2019,** *15* (12), 1191-1198.
11. Singh, R. N.; Ottesen, E. W.; Singh, N. N., The First Orally Deliverable Small Molecule for the Treatment of Spinal Muscular Atrophy. *Neurosci. Insights* **2020,** *15*, 2633105520973985.
12. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G., Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995,** *21* (3), 167-95.
13. Solomatin, S. V.; Greenfeld, M.; Chu, S.; Herschlag, D., Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature* **2010,** *463* (7281), 681-684.
14. Al-Hashimi, H. M.; Walter, N. G., RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.* **2008,** *18* (3), 321-9.

15. Mustoe, A. M.; Brooks, C. L.; Al-Hashimi, H. M., Hierarchy of RNA functional dynamics. *Annu. Rev. Biochem.* **2014,** *83*, 441-66.

16. Nudler, E., Flipping Riboswitches. *Cell* **2006,** *126* (1), 19-22.

17. Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ganesan, S.; Goodsell, D. S.; Ghosh, S.; Green, R. K.; Guranović, V.; Guzenko, D.; Hudson, B. P.; Lawson, C. L.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Persikova, I.; Randle, C.; Rose, A.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Tao, Y. P.; Voigt, M.; Westbrook, J. D.; Young, J. Y.; Zardecki, C.; Zhuravleva, M., RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2021,** *49* (D1), D437-d451.

18. Kastner, J., Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011,** *1* (6), 932-942.

19. Earl, D. J.; Deem, M. W., Parallel tempering: Theory, applications, and new perspectives. *PCCP* **2005,** *7* (23), 3910-3916.

20. Husic, B. E.; Pande, V. S., Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018,** *140* (7), 2386-2396.

21. Wales, D. J., Discrete path sampling. *Mol. Phys.* **2002,** *100* (20), 3285-3305.

22. Sharma, S.; Ding, F.; Dokholyan, N. V., iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* **2008,** *24* (17), 1951-2.

23. Krokhotin, A.; Houlihan, K.; Dokholyan, N. V., iFoldRNA v2: folding RNA with constraints. *Bioinformatics* **2015,** *31* (17), 2891-2893.

24. Williams, B., II; Zhao, B.; Tandon, A.; Ding, F.; Weeks, K. M.; Zhang, Q.; Dokholyan, N. V., Structure modeling of RNA using sparse NMR constraints. *Nucleic Acids Res.* **2017,** *45* (22), 12638-12647.

25. Boniecki, M. J.; Lach, G.; Dawson, W. K.; Tomala, K.; Lukasz, P.; Soltysinski, T.; Rother, K. M.; Bujnicki, J. M., SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **2016,** *44* (7), e63.

26. Cragnolini, T.; Laurin, Y.; Derreumaux, P.; Pasquali, S., Coarse-Grained HiRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings. *J. Chem. Theory Comput.* **2015,** *11* (7), 3510-3522.

27. Zhang, D.; Chen, S.-J.; Zhou, R., Modeling Noncanonical RNA Base Pairs by a Coarse-Grained IsRNA2 Model. *J. Phys. Chem. B* **2021,** *125* (43), 11907-11915.

28. Tan, Y.-L.; Wang, X.; Yu, S.; Zhang, B.; Tan, Z.-J., cgRNASP: coarse-grained statistical potentials with residue separation for RNA structure evaluation. *NAR Genom. Bioinform.* **2023,** *5* (1).

29. Rother, M.; Rother, K.; Puton, T.; Bujnicki, J. M., ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.* **2011,** *39* (10), 4007-22.

30. Parisien, M.; Major, F., The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **2008,** *452* (7183), 51-5.

31. Popenda, M.; Szachniuk, M.; Antczak, M.; Purzycka, K. J.; Lukasiak, P.; Bartol, N.; Blazewicz, J.; Adamiak, R. W., Automated 3D structure composition for large RNAs. *Nucleic Acids Res.* **2012,** *40* (14), e112.

32. Antczak, M.; Popenda, M.; Zok, T.; Sarzynska, J.; Ratajczak, T.; Tomczyk, K.; Adamiak, R. W.; Szachniuk, M., New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure. *Acta Biochim. Pol.* **2016,** *63* (4), 737-744.

33. Xu, X.; Zhao, C.; Chen, S. J., VfoldLA: A web server for loop assembly-based prediction of putative 3D RNA structures. *J. Struct. Biol.* **2019,** *207* (3), 235-240.

34. Xu, X.; Zhao, P.; Chen, S. J., Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One* **2014,** *9* (9), e107504.

35. Zhao, C.; Xu, X.; Chen, S. J., Predicting RNA Structure with Vfold. *Methods Mol Biol* **2017,** *1654*, 3-15.

36. Zhang, Y.; Wang, J.; Xiao, Y., 3dRNA: Building RNA 3D structure with improved template library. *Comput. Struct. Biotechnol. J.* **2020,** *18*, 2416-2423.

37. Jain, S.; Schlick, T., F-RAG: Generating Atomic Coordinates from RNA Graphs by Fragment Assembly. *J. Mol. Biol.* **2017,** *429* (23), 3587-3605.

38. Das, R.; Baker, D., Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (37), 14664-9.

39. Das, R.; Karanicolas, J.; Baker, D., Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **2010,** *7* (4), 291-294.

40. Watkins, A. M.; Rangan, R.; Das, R., FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure* **2020,** *28* (8), 963-+.

41. Chou, F. C.; Kladwang, W.; Kappel, K.; Das, R., Blind tests of RNA nearest-neighbor energy prediction. *Proc. Natl. Acad. Sci. U. S. A.* **2016,** *113* (30), 8430-5.

42. Vanegas, P. L.; Hudson, G. A.; Davis, A. R.; Kelly, S. C.; Kirkpatrick, C. C.; Znosko, B. M., RNA CoSSMos: Characterization of Secondary Structure Motifs-a searchable database of secondary structure motifs in RNA three-dimensional structures. *Nucleic Acids Res.* **2012,** *40* (D1), D439-D444.

43. Wales, D. J., Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **2004,** *102* (9-10), 891-908.

44. Wales, D. J.; Disney, M. D.; Yildirim, I., Computational Investigation of RNA A-Bulges Related to the Microtubule-Associated Protein Tau Causing Frontotemporal Dementia and Parkinsonism. *J. Phys. Chem. B* **2019,** *123* (1), 57-65.

45. Wales, D. J.; Yildirim, I., Improving Computational Predictions of Single-Stranded RNA Tetramers with Revised α/γ Torsional Parameters for the Amber Force Field. *J. Phys. Chem. B* **2017,** *121* (14), 2989–2999.

46. Chen, J. L.; VanEtten, D. M.; Fountain, M. A.; Yildirim, I.; Disney, M. D., Structure and Dynamics of RNA Repeat Expansions That Cause Huntington's Disease and Myotonic Dystrophy Type 1. *Biochemistry* **2017,** *56* (27), 3463-3474.

47. Yildirim, I.; Chakraborty, D.; Disney, M. D.; Wales, D. J.; Schatz, G. C., Computational Investigation of RNA CUG Repeats Responsible for Myotonic Dystrophy 1. *J. Chem. Theory Comput.* **2015,** *11* (10), 4943-4958.

48. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995,** *117* (19), 5179-5197.

49. Yildirim, I.; Stern, H. A.; Kennedy, S. D.; Tubbs, J. D.; Turner, D. H., Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine. *J. Chem. Theory Comput.* **2010,** *6* (5), 1520-1531.

50. Wales, D. J., *Energy Landscapes*. Cambridge University Press, UK: 2003.

51. Strodel, B.; Wales, D. J., Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide. *Chem. Phys. Lett.* **2008,** *466* (4-6), 105-115.

52. Wales, D. J., Coexistence in small inert gas clusters. *Mol. Phys.* **1993,** *78* (1), 151-171.

53. Kabsch, W., A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **1978,** *34* (5), 827-828.

54. Hopcroft, J. E.; Karp, R. M., An $n^{5/2} $ Algorithm for Maximum Matchings in Bipartite Graphs. *SIAM J. Comput.* **1973,** *2* (4), 225-231.

55. Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Greene, D.; Harris, R.; Homeyer, N.; Huang, Y.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Salomon-Ferrer, R.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. *AMBER 18*, University of California: San Francisco, CA, 2018.

56. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X., A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press: Portland, Oregon, 1996; pp 226–231.

57. Condon, D. E.; Kennedy, S. D.; Mort, B. C.; Kierzek, R.; Yildirim, I.; Turner, D. H., Stacking in RNA: NMR of Four Tetramers Benchmark Molecular Dynamics. *J. Chem. Theory Comput.* **2015,** *11* (6), 2729-2742.

58. Yildirim, I.; Stern, H. A.; Tubbs, J. D.; Kennedy, S. D.; Turner, D. H., Benchmarking AMBER force fields for RNA: Comparisons to NMR spectra for single-stranded r(GACC) are improved by revised χ torsions. *J. Phys. Chem. B* **2011,** *115* (29), 9261–9270.

59. Zhao, J.; Kennedy, S. D.; Turner, D. H., Nuclear Magnetic Resonance Spectra and AMBER OL3 and ROC-RNA Simulations of UCUCGU Reveal Force Field Strengths and Weaknesses for Single-Stranded RNA. *J. Chem. Theory Comput.* **2022,** *18* (2), 1241-1254.

60. Bergonzo, C.; Grishaev, A.; Bottaro, S., Conformational heterogeneity of UCAAUC RNA oligonucleotide from molecular dynamics simulations, SAXS, and NMR experiments. *RNA* **2022,** *28* (7), 937-946.

61. Doornbos, J.; Wreesmann, C. T. J.; Boom, J. H.; Altona, C., Conformational analysis of the single-stranded ribonucleic acid AACC. A one-dimensional and two-dimensional proton NMR study at 500 MHz. *Eur. J. Biochem.* **1983,** *131* (3), 571-579.

62. Isaksson, J.; Acharya, S.; Barman, J.; Cheruku, P.; Chattopadhyaya, J., Single-stranded adenine-rich DNA and RNA retain structural characteristics of their respective double-stranded conformations and show directional differences in stacking pattern. *Biochemistry* **2004,** *43* (51), 15996-16010.

63. Rijnbrand, R.; Thiviyanathan, V.; Kaluarachchi, K.; Lemon, S. M.; Gorenstein, D. G., Mutational and Structural Analysis of Stem-loop IIIc of the Hepatitis C Virus and GB Virus B Internal Ribosome Entry Sites. *J. Mol. Biol.* **2004,** *343* (4), 805-817.

64. Pujari, N.; Saundh, S. L.; Acquah, F. A.; Mooers, B. H. M.; Ferré-D'Amaré, A. R.; Leung, A. K.-W., Engineering Crystal Packing in RNA Structures I: Past and Future Strategies for Engineering RNA Packing in Crystals. *Crystals* **2021,** *11* (8), 952.

65. Heus, H. A.; Pardi, A., Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **1991,** *253* (5016), 191-194.

66. Antao, V. P.; Lai, S. Y.; Tinoco, I., Jr., A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* **1991,** *19* (21), 5901-5.

67. Bottaro, S.; Lindorff-Larsen, K., Mapping the Universe of RNA Tetraloop Folds. *Biophys. J.* **2017,** *113* (2), 257-267.

68. Taghavi, A.; Riveros, I.; Wales, D. J.; Yildirim, I., Evaluating Geometric Definitions of Stacking for RNA Dinucleoside Monophosphates Using Molecular Mechanics Calculations. *J. Chem. Theory Comput.* **2022,** *18* (6), 3637-3653.

69. Zhao, Q.; Huang, H. C.; Nagaswamy, U.; Xia, Y.; Gao, X.; Fox, G. E., UNAC tetraloops: to what extent do they mimic GNRA tetraloops? *Biopolymers* **2012,** *97* (8), 617-28.

70. Baumruk, V.; Gouyette, C.; Huynh-Dinh, T.; Sun, J. S.; Ghomi, M., Comparison between CUUG and UUCG tetraloops: thermodynamic stability and structural features analyzed by UV absorption and vibrational spectroscopy. *Nucleic Acids Res.* **2001,** *29* (19), 4089-96.

71. Jucker, F. M.; Pardi, A., Solution Structure of the CUUG Hairpin Loop: A Novel RNA Tetraloop Motif. *Biochemistry* **1995,** *34* (44), 14416-14427.

72. Comolli, L. R.; Ulyanov, N. B.; Soto, A. M.; Marky, L. A.; James, T. L.; Gmeiner, W. H., NMR structure of the 3' stem-loop from human U4 snRNA. *Nucleic Acids Res.* **2002,** *30* (20), 4371-9.

73. Ippolito, J. A.; Kanyo, Z. F.; Wang, D.; Franceschi, F. J.; Moore, P. B.; Steitz, T. A.; Duffy, E. M., Crystal Structure of the Oxazolidinone Antibiotic Linezolid Bound to the 50S Ribosomal Subunit. *J. Med. Chem.* **2008,** *51* (12), 3353-3356.

74. Court, D. L.; Gan, J.; Liang, Y.-H.; Shaw, G. X.; Tropea, J. E.; Costantino, N.; Waugh, D. S.; Ji, X., RNase III: Genetics and Function; Structure and Mechanism. *Annu. Rev. Genet.* **2013,** *47* (1), 405-431.

75. Choi, J.; Indrisiunaite, G.; DeMirci, H.; Ieong, K.-W.; Wang, J.; Petrov, A.; Prabhakar, A.; Rechavi, G.; Dominissini, D.; He, C.; Ehrenberg, M.; Puglisi, J. D., 2′-O-methylation in mRNA disrupts tRNA decoding during translation elongation. *Nat. Struct. Mol. Biol.* **2018,** *25* (3), 208-216.

76. Cheong, C.; Cheong, H.-K., RNA Structure: Tetraloops. In *Encyclopedia of Life Sciences*.

77. Mulmuley, K.; Vazirani, U. V.; Vazirani, V. V., Matching is as easy as matrix inversion. *Combinatorica* **1987,** *7* (1), 105-113.

78. Liu, P.; Agrafiotis, D. K.; Theobald, D. L., Fast Determination of the Optimal Rotational Matrix for Macromolecular Superpositions. *J. Comput. Chem.* **2010,** *31* (7), 1561-1563.