mathematics

MDPI

Article

A Mathematical Model for RNA 3D Structures

Sixiang Zhang and Liming Cai *

School of Computing, University of Georgia, Athens, GE 30602, USA; sixiang.zhang@uga.edu
* Correspondence: liming@uga.edu

Abstract: The computational prediction of RNA three-dimensional (3D) structures remains a significant challenge, largely due to the limited understanding of RNA folding pathways. Although the scarcity of resolved native RNA structures has hindered the effectiveness of machine learning-based prediction methods, small, local structural motifs are both recurring and abundant in the available data. Precisely modeling these geometric motifs presents a promising approach to improving 3D structure prediction. In this paper, we introduce a novel mathematical model that represents RNA 3D structures as collections of interacting helices with concise geometric descriptions. By using a small set of parameters for each modeled helix, our method maps RNA strand segments onto helices within a 3D space, facilitating the effective assembly of large RNA structures. Preliminary tests on RNA sequences from the Protein Data Bank demonstrated the model's potential in predicting key structural elements, including double helices, hairpin loops, and bulges.

Keywords: RNA 3D structure; nucleotide; base pair; RNA secondary structure; stem–loop; helix; double helix; bulge; rotation; translation; backbone conformation; machine learning

MSC: 92B20; 92B05

1. Introduction

RNA (ribonucleic acid) is a fundamental molecule in biology and a cornerstone of biomedical research. It has played a critical role in recent high-profile advancements in disease diagnosis and treatment [1]. The functional roles of RNA in cells—including protein synthesis, gene expression and regulation, and catalytic and structural contributions to various cellular processes [2]—are mediated through its spatial interactions with other biomolecules. Consequently, understanding the three-dimensional (3D) structures of RNAs is essential for unraveling their biological roles. Methods for RNA structure elucidation based on biological experiments [3], while accurate, are often slow and costly. To complement these efforts, computational algorithms for RNA structure prediction have emerged as a more efficient and cost-effective alternative. These algorithms can rapidly narrow down potential structure candidates, facilitating subsequent experimental validation [4,5].

Research in computational RNA 3D structure prediction, however, has lagged behind its protein counterpart by a big margin, largely due to the instability of RNA structures and much less understanding of their folding pathways. Despite decades of effort and the development of various computational methods [6], significant challenges persist. In particular, homology-based methods, the most accurate, rely on the known 3D structures of similar sequences [7–10], but these are often unavailable for novel structures. Folding algorithms transform and refine secondary structures [9,11,12], yet their accuracy depends on the reliability of the given (predicted) secondary structure. Fragment-based methods assemble small 3D motifs [7,8,13–15], but they are computationally intensive and struggle

check for updates

Academic Editor: Takashi Suzuki

Received: 4 February 2025
Revised: 8 April 2025
Accepted: 10 April 2025
Published: 21 April 2025

Citation: Zhang, S.; Cai, L. A Mathematical Model for RNA 3D Structures. *Mathematics* **2025**, *13*, 1352. https://doi.org/10.3390/math13081352

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

with large or novel RNAs. Physics-based methods, such as molecular dynamics simulations [16–18], offer either reduced or detailed modeling [19,20] but require significant computational resources.

Emerging AI and machine learning advancements are transforming biological data analysis, including RNA 3D structure prediction [6,21–24]. While deep learning approaches show promise, their performance is often limited by small-scale RNA sequence data and insufficient structural information. Even secondary structure predictions struggle without evolutionary or homology data, which are challenging to obtain [25]. More recently, large language models (LLMs), trained on vast datasets, have demonstrated significant potential in biological sequence analysis [26] as well as natural language processing [27]. Fine-tuning general-purpose LLMs with domain-specific datasets through data augmentation techniques can adapt them for specialized tasks. RNA-specific LLMs, in particular, have shown reasonable performance in predicting splicing sites, functional elements, and RNA secondary structures by modeling nucleotide arrangements and motifs [28]. However, their application to RNA 3D structure prediction remains hindered by the scarcity of experimentally validated native structures for training [28–30]. Specifically, only 1% of the approximately 200,000 resolved 3D structures available in the PDB are RNA 3D structures [29].

To advance machine learning-based RNA 3D structure prediction, in particular, by leveraging RNA LLMs, it becomes extremely desirable to increase the volume and diversity of plausible 3D models available for training purposes. Despite the limited number of resolved 3D RNA structures, numerous recurring local 3D structural motifs can be identified within the available native structures. These motifs can be leveraged to assemble plausible, full 3D models, including of previously unseen structures, for given RNA sequences. Toward this end, we propose a novel mathematical model for generating and predicting RNA 3D structures. In our model, every 3D structure consists of a collection of single (and possibly short) strands of nucleotides. Such strands are mathematically definable as non-uniform helical functions, and they are brought into spatial proximity by chemical interactions between nucleotides on these strands, yielding the 3D structure. With machine learning, the parameters of these helical functions can be learned from the local structural motifs that recur in the available native structures. Our model enables the effective assembly of plausible 3D RNA structures, representing each structure as a collection of interacting helices.

2. Background

2.1. Architecture of RNA 3D Structure

A ribonucleic acid (RNA) is a macro-molecule made up of nucleotides, each consisting of a phosphate group, a sugar, and an aromatic nucleobase, which is usually either adenine (`A`), cytosine (`C`), guanine (`G`), or uracil (`U`), with some exceptions. A phosphate group is attached to the 3′ position of one ribose and the 5′ position of the next, forming the phosphodiester bonds that connect the nucleotides of an RNA molecule in a linear sequence. Nucleotides can interact through chemical bonds, bringing them into spatial proximity of each other. The pairs with the strongest base–base interactions, called *canonical base pairs*, are the Watson–Crick pairs `A-U` between adenine and uracil, with two hydrogen bonds, and `C-G` between cytosine and guanine, with three hydrogen bonds, as well as the wobble pair `G-U` between guanine and uracil with one hydrogen bond (Figure 1).

The most fundamental structural elements of RNA are the canonical base pairs. They do not work alone but instead form a group with a stacked, ladder shape, bringing two individual, contiguous strands together into a base-paired region, called the *stem*. These, together with unpaired strands, called *loops*, constitute the *secondary structure* of RNA

molecules. Figure 2A shows the secondary structure of tRNA, including the base pairs, stem, and various loops.

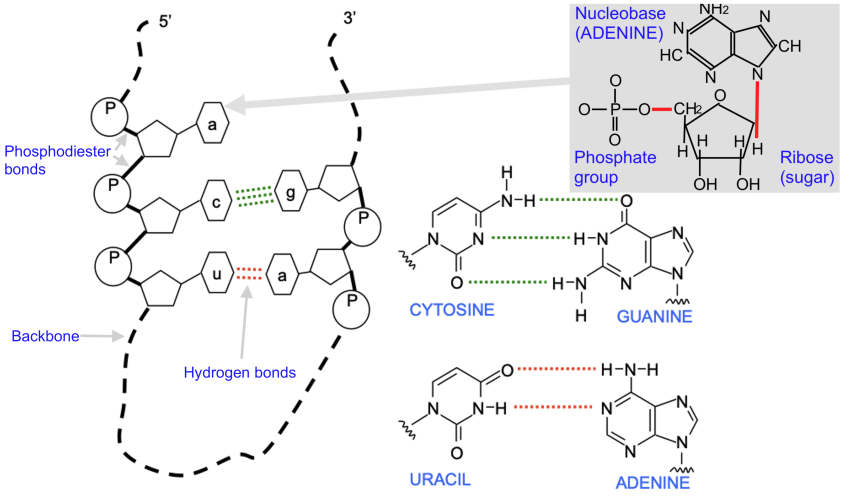


Figure 1. Each nucleotide consists of a phosphate group, a ribose, and one of the four nucleobases (`A, C, G, U`). Nucleotides are connected in a linear sequence by a phosphodiester bond, which connects adjacent nucleotides by linking the phosphate group of one nucleotide to the hydroxyl group on the ribose of the next nucleotide, forming the backbone. Hydrogen bonds that pair nucleobases make it possible for the sequence to fold spatially back onto itself. The figure shows the Watson–Crick pairs `C-G`, with three hydrogen bonds, and `A-U`, with two bonds.

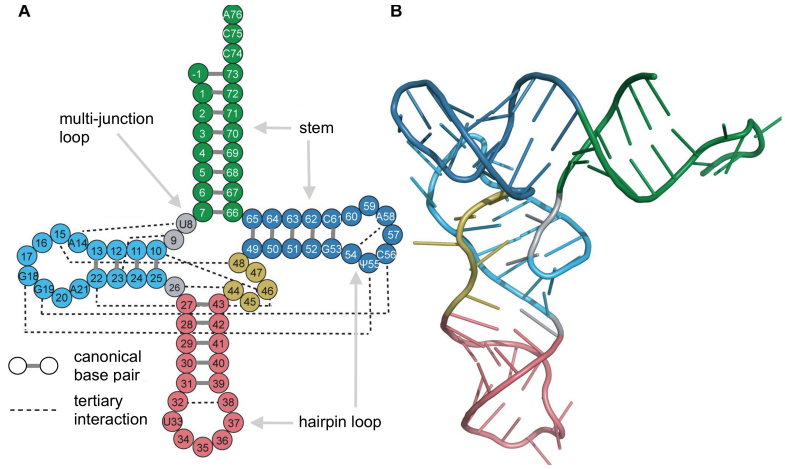


Figure 2. (**A**) The "clover leaf" shape of the tRNA secondary structure, where secondary structure elements include canonical base pairs, the stem, hairpin loops, and multi-junction loops. The figure in (**A**) also shows non-canonical, tertiary interactions between nucleotides (dotted lines). (**B**) The 3D structure of tRNA, where the 3D color-coded helices correspond to the 2D stems in (**A**). Both figures are taken (with permission) from Figure 1 of a publication by Krahn et al. [31], with some added annotations.

The secondary structure scaffolds the 3D structure, with stems corresponding to double helices in the 3D space. Figure 2 shows such a correspondence. While double helices are the main elements scaffolding the 3D structure of an RNA molecule, unpaired loops connect these double helices. Together with other types of nucleotide–nucleotide interactions, they form higher-order structures and ultimately the 3D conformation. This process not only involves the canonical base pairs but also non-canonical *tertiary* nucleotide–nucleotide interactions between two bases, a base and a phosphate, and a base and a ribose [32–36]. Figure 2A shows tertiary interactions in tRNA, which, together with canonical base pairs, are the factors behind the folding of a tRNA molecule into an L-shaped 3D structure.

2.2. Coarse-Grain Modeling

This research used RNA sequences from the Protein Data Bank (PDB) dataset [37], where each data entry includes atomic coordinates and other critical RNA nucleotide and structure information. As each nucleotide may contain about 27 to 30 atoms, our preliminary investigation focused on a *coarse-grain* model where every nucleotide was represented as a single pseudo-atom and the 3D structure as the backbone conformation. The C1′ atom of the ribose is the most common choice for the coordinate representation of RNA nucleotides, and it worked reasonably well for the stem region. However, our tests showed it was not a good fit for the unpaired hairpin loop segment since the C1′ atoms were closer to the base side-chain and far away from the sugar–phosphate backbone, making it difficult to measure the nucleotide–nucleotide distance in a situation where side-chains may flip outward (Figure 3).

The RNA 3D structures obtained from the PDB have been validated through experimental methods (X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy). These structures are accurate within 3 Å. Each entry contains atomic coordinates for the corresponding RNA molecule (chain). However, since not all biological experiments are perfect, not all atomic entries include coordinates in the molecule structure determined using these methods. For instance, hydrogen atoms may not be observed using X-ray crystallography methods [38]. Therefore, sometimes atoms may be missing from PDB coordinate files. When this was the case for O5′, we used the P or O3′ atom's position as a replacement for that of O5′ because these two atoms are also on the sugar–phosphate backbone, like O5′ atoms.

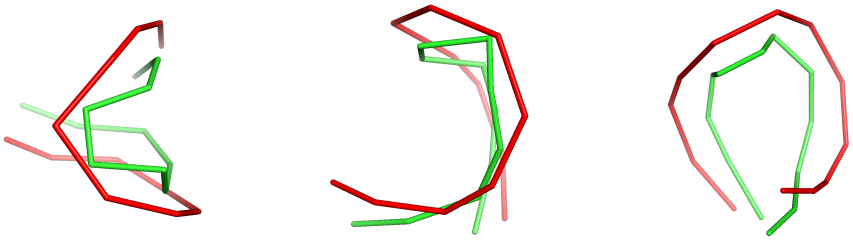


Figure 3. View from three different angles of difference between backbone conformations of two different pseudo-atom models for RNA molecule (PDB ID 1RNG). Red curve is formed from nucleotides represented by O5′ atoms and green from nucleotides represented by C1′ atoms. Figures were drawn with software PyMOL Version 3.1 [39].

3. The Model

In RNA 3D structures, we identify two distinct types of helices: base-paired double helices and unpaired single helices. From this perspective, every RNA 3D structure can be modeled as a collection of helices assembled into a 3D conformation. These helices interact through nucleotide interactions, which position them in close proximity and contribute to the overall structure. Each helix in the model has a set of parameters whose values are determined by the given RNA sequence and its secondary structure.

3.1. Parameters for Double Helices

A double helix can be viewed as two interacting single helices joined by base pairs. The distance between two base-paired nucleotides is the *diameter* of the hypothetical cylinder circled by the double helix (Figure 4a). It has been established that the nucleotide–nucleotide distance of a Watson–Crick pair is nearly 10.5 Å; the *backbone distance* between two neighboring nucleotides on the same strand is $b = 3.4$ Å when C1′ atoms represent nucleotides [40] in a coarse-grain model. Since our model uses O5′ atoms to represent nucleotides, we computed the diameter d as the distance between two O5′ atoms in two

base-paired nucleotides and the length b as the distance between two consecutive O5′ atoms for the available RNA 3D models extracted from the PDB. From the two histograms and O5′-O5′ plot below, it is clear that d and b nearly matched the normal distribution (Figure 5). Based on the principle that the sample mean $\overline{X}$ from a group of observations is an estimate of the population mean μ, we computed the means of d and b to be 16.5 Å and 5.8 Å, respectively. Furthermore, according to the literature [41], a double helix making a complete turn around its axis contains 11 base pairs.

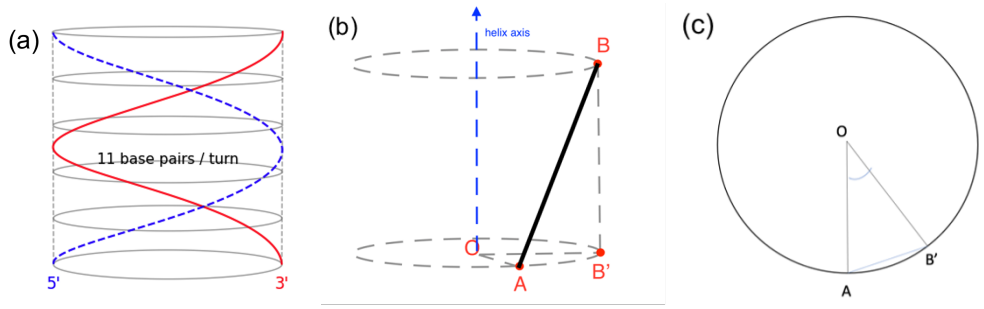


Figure 4. (**a**) A double helix consists of two interacting single helices (blue and red) circling around a hypothetical cylinder. An RNA molecule's growth is always from 5′ to 3′. (**b**) Relative positions of nucleotide B and nucleotide A on the same strand, with B being projected onto the same plane as nucleotide A and resulting in B'. (**c**) The rotation angle obtained by projecting nucleotide B onto plane AOB'.

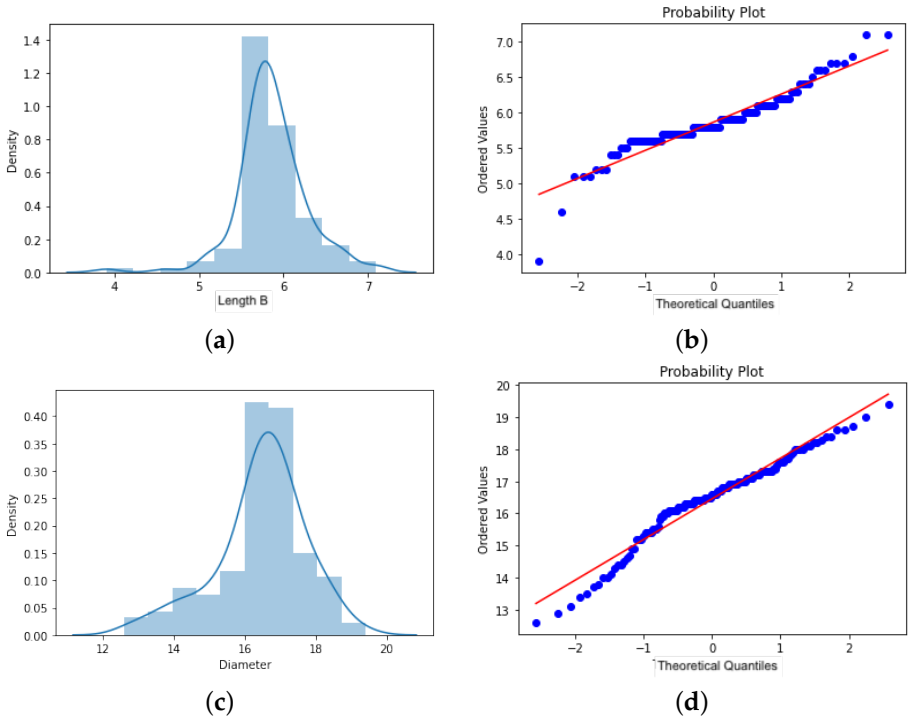


Figure 5. (**a**,**c**) The histograms of lengths b and d nearly follow a Gaussian distribution. (**b**,**d**) Length b's and d's plots indicate that the data approximately follow a normal distribution, lying close to a diagonal line through the main cluster of the points.

The plots are based on the 30 RNA crystal structures given in Table 1. A total of 138 Watson–Crick base pairs (for d) and 134 consecutive O5′–O5′ backbone steps (for b) were extracted. Whenever the leading O5′ atom was missing, the nearest P atom was used instead. Table 2 gives strong evidence that these two parameters followed near-Gaussian distributions. The plots in Figure 5 show that deviations from normality occurred almost exclusively in the extreme tails; the central 90% of the data followed a near-Gaussian

pattern. Because our geometric model used the sample means (16.5 Å and 5.8 Å), these tail deviations did not have an impact on subsequent calculations. In addition, as the Pearson, $r = 0.097$ ($p = 0.265$), and Spearman, $\rho = 0.090$ ($p = 0.31$), correlation coefficients revealed no significant correlations between parameters d and b, they were treated as independent variables.

Table 1. Performance of 3D structure prediction with the proposed model for 30 small RNA molecules. For each RNA, the RMSD_mean is the RMSD averaged over adjustments of the top loop helix segment, while the minimum RMSD is the lowest of them all.

RNA_num	PDB_id	Length (*nt*)	RMSD (Å)	RMSD_mean (Å)
1	1Q75	15	1.86	1.89
2	2GVO	18	3.09	4.33
3	1ATO	19	2.23	2.54
4	1UUU	19	3.09	3.15
5	2KOC	14	2.07	2.28
6	1RNG	12	1.9	2.07
7	2RLU	19	2.45	2.59
8	6PK9	20	3.05	3.1
9	2Y95	14	2.09	2.43
10	1ZIG	12	1.7	1.72
11	1BZ3	17	2.22	2.36
12	1HS3	13	2.16	2.16
13	1HS1	13	2.21	2.21
14	1HS8	13	2.08	2.08
15	1HS4	13	2.2	2.21
16	1HS2	13	2.18	2.18
17	1LK1	14	2.28	2.52
18	1WKS	17	2.71	2.81
19	1MT4	24	2.68	2.78
20	1E4P	24	2.73	3.12
21	2MXJ	11	2.17	2.23
22	1BN0	20	1.54	2.26
23	1AFX	12	1.46	1.62
24	3PHP	23	2.59	3.41
25	1F9L	22	2.35	2.48
26	2M5U	22	1.64	1.94
27	1JTJ	23	4.44	4.96
28	1ZIF	12	1.83	1.97
29	1ZIH	12	1.12	1.36
30	1K5I	23	2.09	2.34

Table 2. Statistics of d and b values extracted from 30 RNA crystal structures given in Table 1.

Parameter	Count	Mean $\pm$ Std, Å	Shapiro–W	p-Value
d	138	16.5 ± 1.29	0.964	0.0012
b	134	5.8 ± 0.42	0.919	6×10^{-7}

We now give more detailed descriptions of the parameters for double helices. Figure 4b shows two consecutive nucleotides, A and B, on the same helix strand. Since it takes 11 bases to make a full turn for a single helix, $\angle AOB' = \frac{360^\circ}{11} = 32.7272^\circ$. Be-

cause we have established the diameter $d = 16.5$ Å, the radius r of the cylinder is $\frac{16.5}{2} = 8.25$ Å, and the length AB = 5.8 Å. Then, we calculate the length AB′ as $\sqrt{2 * (1 - \cos \angle AOB')} = 4.648$ Å, and the increasing growth of each nucleotide on the helix axis is equal to length BB′ = $\sqrt{b^2 - AB'^2} = 3.469$ Å. With these parameters, we can draw the projected double helix (Figure 4).

3.2. Parameters for Unpaired Helices

We treat the unpaired strand as a single helix. In particular, for the apex loop of a double-helical structure (Figure 6a), we model this unpaired segment containing n nucleotides, including the enclosing base pairs (X_0 and X_n), as a single unpaired helix. Moreover, this single helix has an axis perpendicular to the bottom double helix axis, and the former has $n - 1$ nucleotides, connecting the two pairing strands of the double helix. As shown in Figure 6b, $\angle X_0 O X_n = \frac{360}{n-1}$, and in the case that $n = 6$, $\angle X_0 O X_n = 72°$.

The distance between the base-paired nucleotides, i.e., the diameter d, becomes the overall height of the apex loop structure. Moreover, the distance from each nucleotide to its neighbor on the same strand is $\frac{d}{n-1}$ Å, e.g., in the case that $n = 6$, $X_1 X_1' = 3.3$ Å. Furthermore, the direct distance $X_0 X_1$ between two successive nucleotides is still b. Therefore, the radius of the cylinder can be calculated as follows: $X_0 X_1' = \sqrt{b^2 - X_1 X_1'^2}$, where $b = 5.8$ Å. Finally, we place the projected top loop with these parameters on the top of the projected double helix top loop structure.

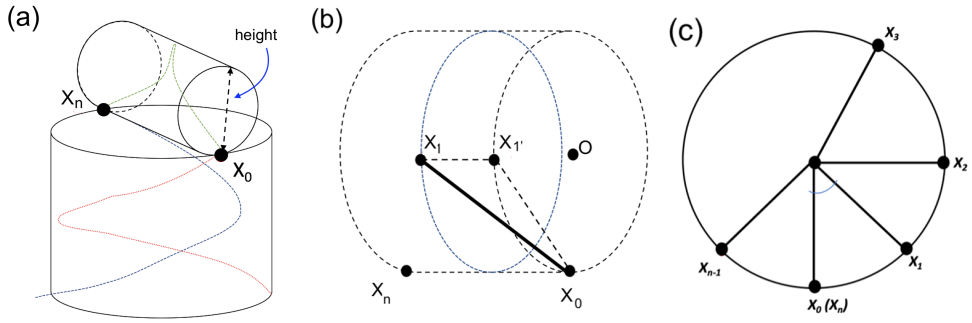


Figure 6. (**a**) The hairpin loop structure of RNA consists of a single-strand (unpaired) loop and two paired loops (red and dark-blue dotted curves) that form a double helix. The single-strand loop (green curve) sits on the top of the double helix. (**b**) Relative positions of nucleotide X_1 with respect to the position of nucleotide X_0 on the cylinder. Nucleotide X_1 is projected onto the same plane as nucleotide X_0, resulting in point X_1'. (**c**) The nucleotides on the top loop strand are projected onto plane $X_0 O X_1'$. This figure also shows the rotation angle for each nucleotide from the top loop segment.

3.3. Top Loop Position Adjustment

Sampled stem–loop structures viewed with tools (such as Pymol and UCSF Chimera [42]) have shown that the top loop is often not right on top of the bottom double helix. So, the parameters for the hairpin segment need to be adjusted.

We propose that the top loop structure can be made more specific based on three rotation angles, θ, γ, and δ, around the X, Y, and Z axes, respectively. Furthermore, the same distance d as between the two base-paired nucleotides A and B is assumed for the enclosing pairs. Thus, nucleotide B should be on the surface of the cylinder, and angles θ and δ should bear some relationship to each other (Figure 7), regarding which we believe that δ varies directly with θ, and vice versa.

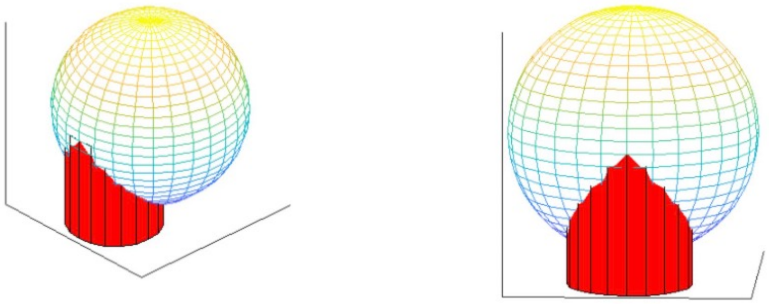


Figure 7. These two figures show the relationship between the two angles θ and δ (created using MATLAB 9.10.0 [43]). The sphere is centered at X_0, and the radius is $|X_0X_n|$. This sphere simulates all possible positions of X_n. Furthermore, X_n must be on the cylinder again because at the same time X_n is the nucleotide belonging to the bottom double helix.

We can calculate the following values (Figure 8):

$$\begin{cases} \text{Length } AB = d; \\ \text{Angle } BAB' = \theta; \\ \text{Angle } \angle AB'A' = \angle BAB'; \\ \text{Thus, length } AA' = d * \sin(\angle AB'A'); \\ \text{Length } A'B' = d * \cos(\angle AB'A'); \\ \text{Therefore, length } A'B'' = A'B'; \\ \text{Then, we know that } A'B'' \text{ also equals } d * \cos(\angle B'A'B''), \\ \text{where } \angle B'A'B'' = \delta; \\ \text{As a result , } |\theta| = |\delta|. \end{cases}$$

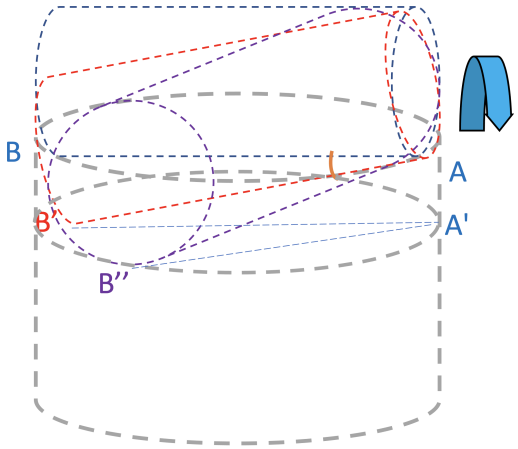


Figure 8. The top cylinder rotates at three angles around nucleotide A for position adjustment. The final position of nucleotide B falls on nucleotide B''.

Moreover, due to the base pairings between nucleotides, the angle around the Z-axis should be slightly smaller than 0; that is, $\delta < 0$. In this case, there are two directions to rotate. In this work, we assume that $\theta > 0$ (i.e., $\theta = -\delta$).

To compute γ and θ, we tested RNA sequences to gain some insights into these two parameters. θ took angle values from 0° to 90°, while γ had a value ranging from -180° to 180°. We substituted the angles into our coded program and then compared them with the actual RNA molecule structure. We were left with the result that was closest to

the accurate data. Interestingly, the results for the two angle parameters were close to a normal distribution (Figure 9). Once again, we took the mean of these two variables, setting $\theta = 21.0^\circ$ and $\gamma = -12.3^\circ$.

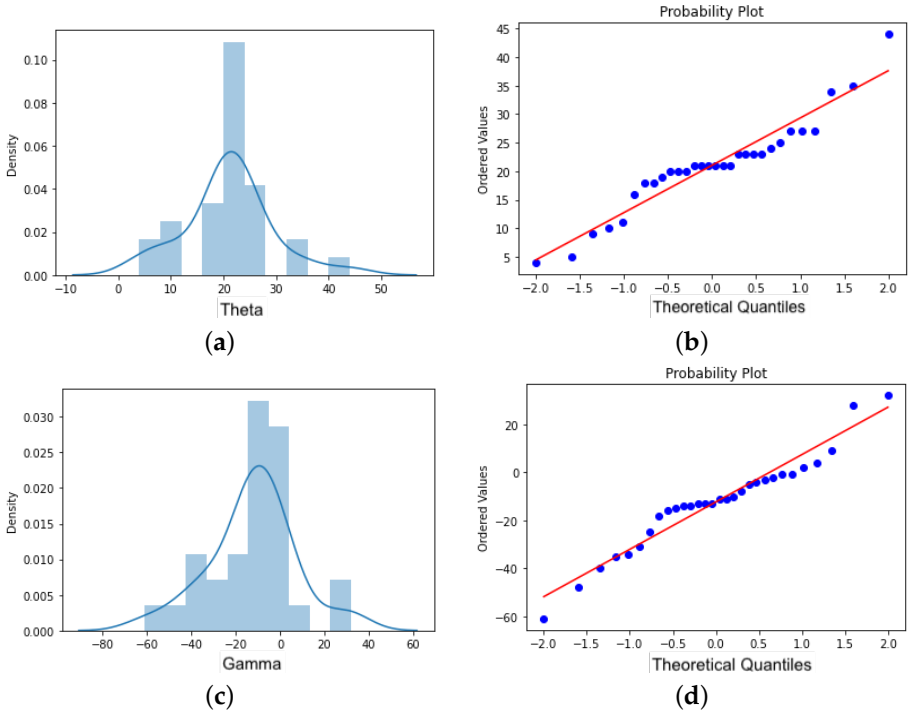


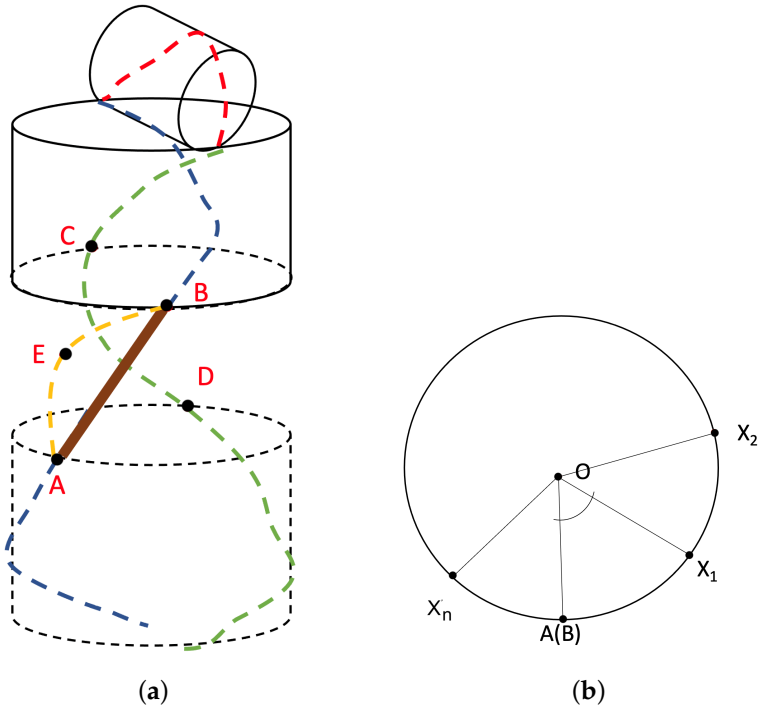
Figure 9. (**a**,**c**) The histograms of θ and γ nearly follow a Gaussian distribution. (**b**,**d**) θ's and γ's plots indicate that the data approximately follow a normal distribution, lying close to a diagonal line through the main cluster of the points.

3.4. Parameters for Bulges

A bulge is a small unpaired segment within one strand of a double helix. We model it with an approach modified from the one used for top hairpin loop modeling. The size of a bulge, the number of unpaired nucleotides, may vary from a single nucleotide to several nucleotides. Since bulges form intricate structures located within double helices, they may play important structural and functional roles [44].

First, we consider a bulge to be modeled as a single helix including k unpaired nucleotides, with two enclosing nucleotides, the same as for the top loop structure. Figure 10a illustrates this case. The rotated angle for the bulge segment is $\angle AOX_1 = \frac{360}{k-1}$; in most cases, $k = 3$, and $\angle AOX_1 = 180^\circ$ (Figure 10b).

For the cylinder enclosing the single helix in the bulge model, we need to calculate its height and radius. Our hypothesis is that the bulge occurs in between the original two neighboring nucleotides. The transformation process means that the bottom nucleotide A remains unchanged and the top nucleotide B rotates around its base-paired partner nucleotide, C. As a result, the distance between the two nucleotides A and B, after the rotation, becomes longer than their previously assumed distance, b (the distance between two consecutive O5′ atoms), which is also the height of the bulge column (Figure 10). To be more specific, let us set the three angles for the rotation. Similarly to the notations used for unpaired helices, we assume that angles θ, γ, and δ are around the X, Y, and Z axes, respectively.



(**a**) (**b**)

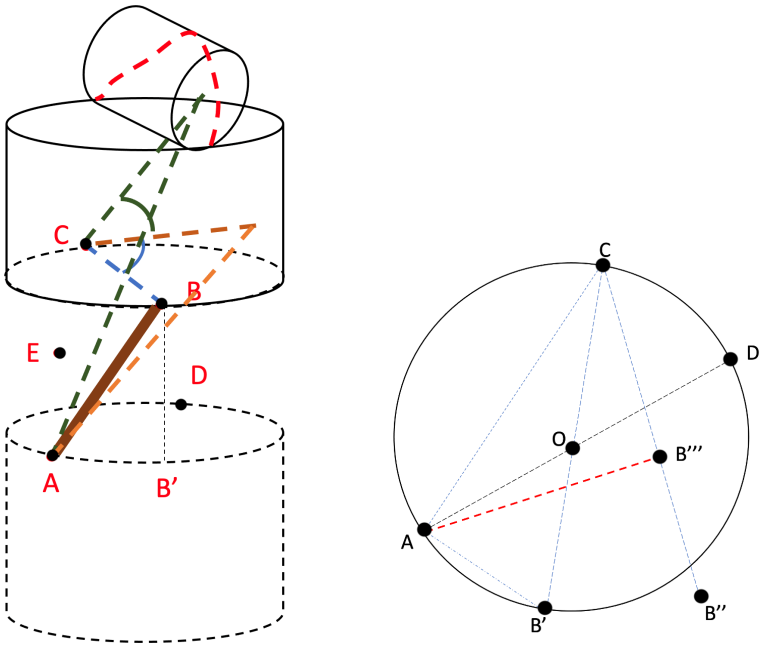
Figure 10. (**a**) A hairpin structure with a bulge part; nucleotides A and B are the two ends of the bulge, and nucleotide E is the one on the bulge. Nucleotide C and nucleotide D are the base-pairing partners of A and B, respectively. The yellow curve is the bulge, the red is the hairpin loop, and the blue and green correspond to the double helix curves of the stems. (**b**) The bulge's nucleotides start from the surface of the bottom double helix's cylinder and then return to the surface, taking a precise turn. This projection shows the rotation angle of each nucleotide.

The angle γ should equal 0, which implies that it does not rotate around the Y-axis. The reason is that if it rotates around the BC-axis (i.e., the Y-axis), it will not affect the distance between nucleotide B and nucleotide A. Thus, we exclude this variable. Angle δ must be >0 because the distances between two enclosing nucleotides in known RNA bulges are all greater than 5.8 (length b). Therefore, we increase the distance by rotating the Z-axis counterclockwise. Also, for the same reason, for angle θ, we conclude that $\theta > 0$.

The height of the bulge cylinder can be calculated as follows (Figure 11):

$$\begin{cases} \angle B'CB'' = \delta, \text{ the angle around the Z-axis;} \\ \angle ACB = \frac{\angle AOB}{2}, \text{ where } \angle AOB = \frac{360}{11} = 32.7272^\circ; \\ B'C = d, \text{ the diameter we calculated previously;} \\ AC = d * \cos(\angle ACB), \text{ and } AB' = d * \sin(\angle ACB); \\ BB' = h = \sqrt{AB^2 - AB'^2}, \text{ where } AB = b; \\ BB''' = h + d * \sin\theta, \text{ the vertical distance from nucleotide } B \text{ to plane } ADB'; \\ CB''' = d * \cos\theta; \\ \text{So, } AB''' = \sqrt{AC^2 + CB'''^2 - 2 * AC * CB''' * \cos(\angle ACB' + \delta)}; \\ \text{The new height for the bulge cylinder is Hbulge } = \sqrt{BB'''^2 + AB'''^2}. \end{cases}$$

Now, we obtain the height of the bulging cylinder concerning the two rotation angles, which allows us to perform relevant tests on the bulge model.



(**a**) (**b**)

Figure 11. (**a**) Three separate rotations of nucleotide B around nucleotide C. The distance between nucleotide A and nucleotide B changes as B rotates. (**b**) Nucleotide B with rotation; the projection on the plane changes from B′ to B″ and then to B‴.

4. Implementation

This section will provide some details on an implementation of the proposed model, which was tested for RNA 3D structure prediction.

4.1. Data Extraction from PDB

The data structure forms provided by PDB files are SMCRA hierarchic data structures [37], which have a unique order (structure/model/chain/residue/atom). Usually, crystal structures from the PDB database only have one model. Models are numbered from 0; chains are identified with capital letters like *A*, *B*, etc. The following code extracts O5′ coordinates from the first residue in chain A of the first model of the structure with the PDB id 1Q75.

```
structure = parser.get_structure(1Q75) #choose the PDB 1Q75
model = structure[0] #choose model0
chain_A = model['A'] #choose chain A
res=chain_A[1] #choose residue 1
atom = res["O5'"] #choose atom O5'
```

4.2. Translation and Rotation Matrices

We now describe how to place the model in all the desired 3D positions. For this, we need rotation and translation operations [45]. In particular, we used the following translation and rotation matrices [46] in this implementation.

The following matrix translates coordinates (x, y, z) to (x_1, y_1, z_1):

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ dx & dy & dz & 1 \end{pmatrix} \tag{1}$$

And (x, y, z) and (x_1, y_1, z_1) have the following relationship: $(x_1, y_1, z_1, 1) = (x, y, z, 1) \cdot T$.

The translation matrix is relatively simple. However, rotation around a point in space needs to move that point to the origin before applying the following rotation matrices; after that, the center of rotation is moved back to the original location.

Three angles are needed to describe a rotational process. In this research, we use Eulerian angles, three angles around the three different axes, in the following order.

(1) Matrix for rotation around the Z-axis: This is to rotate coordinates (x, y, z) around the Z-axis counterclockwise by angle θ to coordinates (x_1, y_1, z_1):

$$R_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2}$$

The rotation is expressed as $(x_1, y_1, z_1, 1) = R_z \cdot (x, y, z, 1)^T$.

(2) Matrix for rotation around the X-axis: This is to rotate coordinates (x, y, z) around the X-axis counterclockwise by angle θ to (x_1, y_1, z_1):

$$R_x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3}$$

The rotation is expressed as $(x_1, y_1, z_1, 1) = R_x \cdot (x, y, z, 1)^T$.

(3) Matrix for rotation around Y-axis:

This is to rotate coordinates (x, y, z) around the Y-axis counterclockwise by angle θ to (x_1, y_1, z_1):

$$Ry = \begin{pmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4}$$

The rotation can be expressed as $(x_1, y_1, z_1, 1) = R_y \cdot (x, y, z, 1)^T$.

According to Euler's rotation theorem, any rotation can be described by only three angles around three axes [47]. As a result, these matrices can apply to the plotting of any helix structure, just in a different order.

4.3. Output Format

The input for the prediction task is a query RNA sequence, along with its known or predicted secondary structure (created using a third-party tool). The output of the prediction is an XYZ file containing a set of 3D atomic coordinates for all the nucleotides in the query sequence. The XYZ file format is a chemical file format supported by many programs. Although there is a formal standard, several variations can be found on the internet. All XYZ files typically have several segments, as follows (illustrated with example 1Q75.xyz). The first line contains the number of atoms in the file; in our case, this was also

the number of nucleotides in a given RNA sequence. The second line will usually be a title, or output filenames, or any preferred meaningful information; here, we filled this line with a PDB ID. The subsequent lines are the atoms and corresponding Cartesian coordinates (Figure 12).

```
[15]
[1q75]
O5' -9.383 3.564 6.205
O5' -11.106 2.229 1.404
O5' -9.538 1.633 -4.01
O5' -5.76 0.728 -8.034
O5' -0.285 -0.641 -9.72
O5' 2.99 -4.823 -9.494
O5' 7.595 -8.845 -8.477
O5' 6.624 -11.453 -4.987
O5' 4.632 -11.097 0.086
O5' 2.796 -7.747 5.931
O5' 5.374 -3.212 7.377
O5' 5.796 1.874 3.669
O5' 5.246 6.876 0.319
O5' 1.834 11.231 -1.912
O5' -3.481 13.774 -3.505
```

Figure 12. Coordinates of O5′ atoms of RNA with PDB ID 1Q75 in XYZ file format.

5. Performance Evaluation

In structural bioinformatics, the RMSD (Root Mean Square Deviation) is the most common way to compare two structures [48], e.g., it is used in biomolecule structure prediction as well as in the experimental determination of biomolecule structures. The RMSD is often measured in Angstroms, Å, and calculated using the following formula [49] (4.5).

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left\|x_i^{S_1} - x_i^{S_2}\right\|^2\right)} \tag{5}$$

where N is the number of residues, and $x_i^{S_1}$ and $x_i^{S_2}$ are the ith atoms' coordinates from two structures, S_1 and S_2, respectively.

In general, if two structures are identical, the RMSD value should be 0. Since often the two structures to be compared may belong to two different coordinate systems, directly computing their RMSD may result in a very high RMSD value. So, RMSD computation is based on transformation between the two coordinate systems to obtain the actual minimal RMSD value [50]. Biopython's built-in RMSD calculation method and Pymol's alignment function already apply such an algorithm to obtain the best RMSD value. We used the RMSD to measure the accuracy of our model predictions. Pymol also makes it possible to visualize the two structures and their superimposition.

5.1. Performance on Stem–Loops Without Bulge

Our model was used to predict 3D structures from given RNA sequences. In the preliminary work, 30 small RNAs with stem–loop structures without bulges from the PDB were tested, and the model's performance was evaluated based on the RMSDs calculated for the predicted and known native structures. Table 1 shows two RMSD values for each RNA because the top loop helix of the stem–loop was adjusted to the appropriate angles. For each of these RNAs, the RMSD_mean is the RMSD averaged over these adjustments, while the minimum RMSD is the lowest of them all.

The data in the line graph shown in Figure 13, which corresponds to Table 1, show additional information on the performance. The blue plot represents the lowest RMSD value obtained by adjusting the top loop segments to a position appropriate for each RNA sequence; the orange plot represents the RMSD value obtained by averaging over the values for the two angles since they were nearly Gaussian distributions. Moreover, from this graph, we can see that the two curves overlap, strong evidence to support our hypothesis for the top loop model.

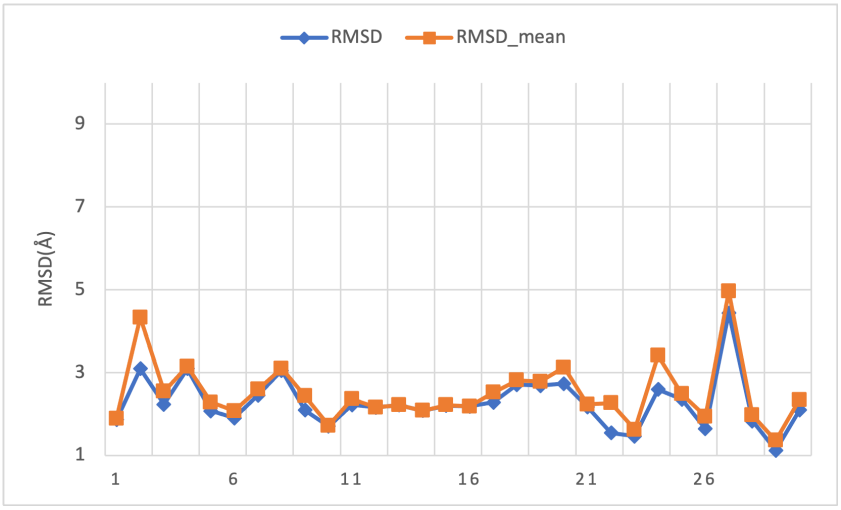


Figure 13. The plots of the mean RMSDs (orange) and minimum RMSDs (blue) over all adjustments of the top loop segments to a position appropriate for each RNA sequence.

For example, for the RNA with the PDB ID 1Q75, Pymol allowed us to further examine the closeness of our predicted model to the actual model. We can see from Figure 14 that our predicted model (red) basically overlaps with the actual model (green), when viewed from all three different perspectives.

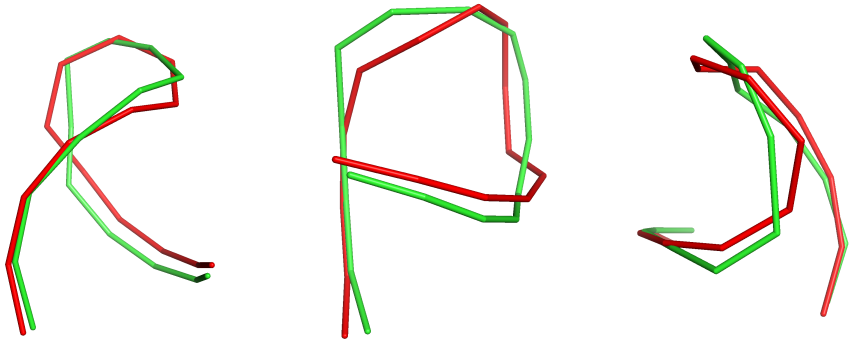


Figure 14. Views from 3 different perspectives of the superimposition of the predicted (red) and native (green) structures of the RNA with the PDB ID 1Q75, generated using Pymol [39]: predictions for 1Q75 had a minimum RMSD of 1.86 and a mean RMSD of 1.89.

5.2. Performance on Stem–Loops with Bulge

Regarding the 3D structure prediction of stem–loops with a bulge, due to time constraints, we only tested four RNA sequences (1NBR, 1BVJ, 1TXS, 1MKF), and the results are shown in Table 3. For all four, the RMSDs showed that decent results were achieved by our model, especially for the irregular bulge structures. We illustrate this with the RNA with the PDB ID 1NBR in Figure 15. In the prediction for 1NBR, the RMSD value between the prediction and native models was 2.46 Å; we attribute this good RMSD value to the fact that not only did the double helices and top loops between the two fit well, but the

bulges in the two models also appeared to be closely aligned. The test result suggests that our choices of bulge parameters were appropriate if not perfect. In spite of the limited test cases, from them, we observed that the rotation angle of the X-axis was equal to 0, which allowed us to hypothesize that the bulging part only had two parameters determining the rotations around the Z-axis and around the Y-axis.

Table 3. Evaluation of 3D structure prediction for 4 RNA molecules with a structure with a bulge in its double helix. The zero rotation angle of the X-axis allowed us to further hypothesize that the bulging segment only had two parameters determining the rotations around the Z-axis and around the Y-axis.

X-Angle (°)	Z-Angle (°)	Rotation_Itself (°)	PDB_ID	RMSD (Å)
0	18	19	1NBR	2.46
0	12	10	1BVJ	2.88
0	15	−15	1TXS	3.35
0	21	50	1MKF	3.28

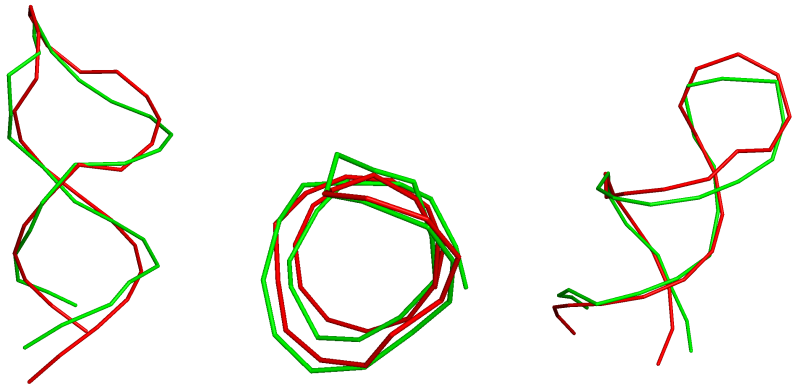


Figure 15. Superimposition of predicted (red) and native (green) structures of RNA with PDB ID 1NBR, viewed from three perspectives and generated using Pymol [39] with RMSD of 2.46 Å. 1NBR contains bulge in its double helix.

6. Conclusions

We have introduced a mathematical model of the 3D structure of RNA as a collection of helices in the spatial space. In particular, we have shown that a stem–loop structure actually consists of a double helix with an apex single helix connecting the two strands that together form the double helix. In addition, bulges present in the double helix can also be modeled with a small, single helix. We demonstrated that all these helices have their own sets of parameters that can be effectively learned from the available native structures.

Our work was motivated by the recent developments in machine learning-based structure prediction. Due to the limited volume of available native RNA 3D structures, pure machine learning-based methods for RNA 3D prediction have lagged far behind those for protein structure prediction. In particular, fine-tuning RNA LLMs for advancing structure prediction requires the generation of a large number of plausible 3D structures. The precise mathematical modeling of helices, the presumably elementary components of RNA 3D structures, would make it possible to efficiently assemble synthetic yet plausible 3D models for effective modeling and training with machine learning.

To the best of our knowledge, this is the first such effort in the pure mathematical modeling of RNA 3D structures. While this paper only presents tests of stem–loop modeling, the ability of our method to use single helices to model irregular shapes like top loops and

bulges well suggests some mathematics fundamentals underlying RNA 3D structures, a topic that deserves to be extensively explored.

Author Contributions: Conceptualization, L.C.; Methodology, S.Z. and L.C.; Writing—review & editing, L.C.; Project administration, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gingeras, T.R. Current frontiers in RNA research. *Front. RNA Res.* **2023**, *1*, 1152146. [CrossRef]
2. Liu, Y.; Hao, Y. Exploring the Possibility of RNA in Diverse Biological Processes. *Int. J. Mol. Sci.* **2023**, *24*, 10674. [CrossRef] [PubMed]
3. Spitale, R.C.; Incarnato, D. Probing the dynamic RNA structurome and its functions. *Nat. Rev. Genet.* **2023**, *24*, 178–196. [CrossRef] [PubMed]
4. Leontis, N.B.; Westhof, E. *RNA 3D Structure Analysis and Prediction (Nucleic Acids and Molecular Biology)*; Springer: Berlin/Heidelberg, Germany, 2012.
5. Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide: An Interdisciplinary Guide (Interdisciplinary Applied Mathematics)*; Springer: New York, NY, USA, 2010.
6. Bernard, C.; Postic, G.; Ghannay, S.; Tahi, F. State-of-the-RNArt: Benchmarking current methods for RNA 3D structure prediction. *NAR Genom. Bioinform.* **2024**, *6*, lqae048. [CrossRef]
7. Parisien, M.; Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **2008**, *452*, 51–55. [CrossRef] [PubMed]
8. Petrov, A.I. RNA 3D Hub: A database of RNA structural motifs and their interactions. *Nucleic Acids Res.* **2013**, *41*, D315–D321.
9. 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*, e112. [CrossRef]
10. Rother, K.; Rother, M. RNA structure prediction tools and their usage in homologous modeling. *Methods Mol. Biol.* **2018**, *1704*, 315–336.
11. Keating, K.S.; Pyle, A.M. Ribonucleic acid structure: Folding and modeling. *Curr. Opin. Struct. Biol.* **2012**, *22*, 317–324.
12. Zhao, Y.; Gong, Z.; Zhang, H.; Zhou, Y. 3dRNA: 3D structure prediction from linear to tertiary structures. *Nucleic Acids Res.* **2012**, *40*, e112.
13. Cheng, C.Y.; Chou, F.C.; Das, R. Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol.* **2015**, *553*, 35–64. [PubMed]
14. Das, R.; Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 14664–14669. [CrossRef] [PubMed]
15. Zhou, L.; Wang, X.; Yu, S.; Tan, Y.; Tan, Z. FebRNA: An automated fragment-ensemble-based model for building RNA 3D structures. *Biophys. J.* **2022**, *121*, 3381–3392. [CrossRef] [PubMed]
16. 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*, e63. [CrossRef]
17. Case, D.A.; Cheatham, T.E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The AMBER biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688. [CrossRef]
18. Zhang, C.; Sun, X.; Baturka, M.; Case, D.A. Force field optimization for RNA simulations with explicit solvent and metal ions. *RNA* **2021**, *27*, 687–702.
19. Xu, X.; Zhao, P.; Chen, S. Vfold: A Web Server for RNA Structure and Folding Thermodynamics Prediction. *PLoS ONE* **2014**, *9*, e107504. [CrossRef]
20. Xu X.; Chen, S. Physics-based RNA structure prediction. *Biophys. Rep.* **2015**, *1*, 2–13. [CrossRef]
21. Li, Y.; Zhang, C.; Feng, C.; Pearce, R.; Freddolino, O.L.L.; Zhang, Y. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nat. Commun.* **2023**, *14*, 5745. [CrossRef]
22. Mukherjee, S.; Moafinejad, S.N.; Badepally, N.G.; Merdas, K.; Bujnicki, J.M. Advances in the field of RNA 3D structure prediction and modeling, with purely theoretical approaches, and with the use of experimental data. *Structure* **2024**, *32*, 1860–1876. [CrossRef]
23. Wang, W.; Feng, C.; Han, R.; Wang, Z.; Ye, L.; Du, Z.; Wei, H.; Peng, F.Z.Z.; Yang, J. trRosettaRNA: Automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **2023**, *14*, 7266. [CrossRef] [PubMed]

24. Wang, X.; Yu, S.; Lou, E.; Tan, Y.-L.; Tan, Z.-J. RNA 3D structure prediction: Progress and perspective. *Molecules* **2023**, *28*, 5532. [CrossRef] [PubMed]
25. Szikszai, M.; Wise, M.; Datta, A.; Ward, M.; Mathews, D.H. Learning models for RNA secondary structure prediction (probably) do not generalise across families. *Bioinformatics* **2022**, *38*, 3892–3899. [CrossRef]
26. Yuan, J.; Tang, R.; Jiang, X.; Hu, X. Large language models for healthcare data augmentation: An example on patient-trial matching. *AMIA Annu. Symp. Proc.* **2024**, *2023*, 1324–1333. [PubMed]
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019.
28. Wang, N.; Bian, J.; Li, Y.; Li, X.; Mumtaz, S.; Kong, L.; Xiong, H. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nat. Mach. Intell.* **2024**, *6*, 548–557. [CrossRef]
29. Shen, T.; Hu, Z.; Sun, S.; Liu, D.; Wong, F.; Wang, J.; Chen, J.; Wang, Y.; Hong, L.; Xiao, J.; et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat. Methods* **2024**, *21*, 2287–2298. [CrossRef]
30. Townshend, R.J.L.; Eismann, S.; Watkins, A.M. Geometric deep learning of RNA structure. *Science* **2021**, *373*, 1047–1051. [CrossRef]
31. Krahn, N.; Fischer, J.T.; Söll, D. Naturally occurring tRNAs with non-canonical structures. *Front. Microbiol.* **2020**, *2020*, 596914.
32. Cech, T.R.; Steitz, J.A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **2014**, *157*, 77–94. [CrossRef]
33. Doudna, J.A.; Cech, T.R. The chemical repertoire of natural ribozymes. *Nature* **2002**, *418*, 222–228. [CrossRef]
34. Leontis, N.B.; Lescoute, A.; Westhof, E. The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.* **2006**, *16*, 279–287. [CrossRef] [PubMed]
35. Simons, R.W.; Grunberg-Manago, M. (Eds.) *RNA Structure and Function*; Cold Spring Harbor Monograph Series; Cold Spring Harbor Laboratory Press: New York, NY, USA, 1998.
36. Westhof, E.; Leontis, N.B. An RNA-centric historical narrative around the Protein Data Bank. *J. Biol. Chem.* **2023**, *296*, 100555. [CrossRef] [PubMed]
37. 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.; et al. 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.* **2020**, *49*, D437–D451. [CrossRef] [PubMed]
38. RCSB: PDB-101. PDB101: Guide to Understanding PDB Data: Missing Coordinates and Biological Assemblies. Available online: https://www.rcsb.org (accessed on 1 November 2024).
39. Schrödinger, L. The PyMOL Molecular Graphics System, Version 3.1. Available online: https://www.pymol.org (accessed on 1 November 2024).
40. Westhof, E. Isostericity and tautomerism of base pairs in nucleic acids. *FEBS Lett.* **2014**, *588*, 2464–2469. [CrossRef]
41. Warden, M.S.; Tonelli, M.; Cornilescu, G.; Liu, D.; Hopersberger, L.J.; Ponniah, K.; Pascal, S.M. Structure of RNA Stem Loop B from the Picornavirus replication platform. *Biochemistry* **2017**, *56*, 2549–2557. [CrossRef]
42. Pettersen, E.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.; Ferrin, T. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef]
43. *MATLAB 9.10.0*, R2021a; The MathWorks Inc.: Natick, MA, USA, 2021.
44. Hermann, T.; Patel, D.J. RNA bulges as architectural and recognition motifs. *Structure* **2000**, *8*, R47–R54. [CrossRef]
45. Evans, P. Rotations and rotation matrices. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2001**, *57*, 1355–1359. [CrossRef]
46. Arfken, G. *Mathematical Methods for Physicists*, 3rd ed.; Academic Press: New York, NY, USA, 1985.
47. Palais, B.; Palais, R.; Rodi, S. A disorienting look at Euler's theorem on the axis of a rotation. *Am. Math. Mon.* **2009**, *116*, 892–909. [CrossRef]
48. Carugo, O. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *J. Appl. Crystallogr.* **2003**, *36*, 125–128. [CrossRef]
49. Nguyen, M.N.; Sim, A.Y.L.; Wan, Y.; Madhusudhan, M.S.; Verma, C. Topology independent comparison of RNA 3D structures using the CLICK algorithm. *Nucleic Acids Res.* **2016**, *45*, e5. [CrossRef] [PubMed]
50. Kromann, J.C. *Charnley/rmsd*. GitHub. Available online: https://github.com/charnley/rmsd (accessed on 1 November 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.