POZNAN UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTING AND TELECOMMUNICATION
Institute of Computing Science

Master's thesis

HIGH-QUALITY MACHINE LEARNING-BASED MODELS FOR 3D RNA STRUCTURE QUALITY ASSESSMENT

Bartosz Adamczyk, 148163

Supervisor
dr hab. inż. Maciej Antczak, prof. PP

POZNAŃ 2025

Contents

Chapter 1

Introduction

3D RNA structures and fold into complex three-dimensional structures that determine their biological function. Despite significant advancements in research over the past several decades, our understanding of the fundamental origins of these structures remains limited. While the *CASP15* [1] and the *RNA-Puzzles* [2][3] competition confirmed we can predict where RNA nucleotides will form canonical base pairs with remarkable accuracy, the non-canonical and long-range interactions that give RNA its functional form remain largely invisible to computational models. This thesis goes deep into uncharted territory, aiming to understand the hidden interactions of RNA through the application of artificial intelligence to assess general quality of 3D RNA structures.

1.1 Motivation

Recent years have been fruitful with generative models able to generate biological 3D structures of proteins. *Google DeepMind* laboratory made a breakthrough in 2022, presenting a model which, for the first time, outperforms human predictors in a protein 3D structure modelling contest [4]. The accomplishments achieved are a result of extensive data collected over numerous years of dedicated research on proteins. The RNA case is quite the opposite; the number of distinctive RNA families with assigned experimentally determined 3D structures is narrow, with only a few candidates in each. These conditions introduced a serious problem for typical approaches such as transformer-based models, which require a high number of diverse and high-quality samples. The cost of experimentally deriving the complex 3D structure is significant, so any options to omit the expense are warmly welcomed. In recent years, the *RNA-Puzzles* contest has emerged as a platform to identify the most effective computational methods for predicting 3D RNA structures. To date, the competition has conducted 39 challenges (five editions) over its 10-year existence. That is a tiny drop in the ocean of demand for fast domain field evolution. Therefore, scientists seek a more efficient way of testing their computational approaches. The best solution assumes the existence of a referee, a tool, or an expert capable of such an assessment. Some methods try to fulfil the demand for such a tool, but unfortunately, the results are not satisfying.

1.2 Goal

The thesis focuses on fulfilling the demand for such a referee by providing a computational model that enables the assessment of 3D RNA models referencelessly using the RMSD metric. The simplest solution to this problem can be solved using the Root Mean Squared Deviation

metric. Still, the calculation is only possible when the target structure is available and consists of the same number of atoms (the same RNA sequence). The algorithm then considers atoms as separate entities for comparing 3D structures. Next, the algorithm takes atoms for both structures and superimposes them; later, the RMSD [5] is computed using the following formula:

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}$$

Where:

- $\{v_{ix}, v_{iy}, v_{iz}\}$ - corresponding x, y, z for i atom from structure v,
- $\{w_{ix}, w_{iy}, w_{iz}\}$ - corresponding x, y, z for i atom from structure w,
- n - number of atoms within structure.

The problem becomes more complex when only the 3D model is under evaluation. The challenge requires an algorithm with a deep understanding of the characteristics of 3D RNA structures, which is ideally suited for artificial intelligence applications. The objective is to create a computational model optimized for deployment on resource-constrained personal devices.

1.3 RNA as Data and Matter: Bridging Biological and Chemical Views

As a biochemical structure, RNA can be viewed as segments of data that encode biological functions, or as minuscule components, such as atoms, that form chemical compounds. By combining these characteristics, one can develop an interdisciplinary understanding of the subject. This approach has yet to be utilised in any other state-of-the-art solutions.

1.3.1 Biological view

Biological researchers use nucleotides to represent RNA sequences, which is the simplest possible representation of their structure. The bag of words of potential bases for nucleotides consists of four: Adenine (A), Cytosine (C), Guanine (G), and Uracil (U), of which the last one is an exclusive base for RNA. Modified nucleotides are also a notable exception, but they are a relatively rare occurrence and are therefore usually omitted.

Scientists always depict the beginning of the sequence as 5'-end and the end as 3'-end, as the Figure 1.1 shows. The nucleobases are connected to a sugar-phosphate backbone, creating a chain/strand that ultimately combines the spatial features of the 3D RNA structure.

```
Sequence (5' to 3'):
UGCUCCUAGUACGAGAGGAACGGAGUG
```

FIGURE 1.1: RNA sequence (1D) given for RNA-Puzzle 11 challenge.

The RNA secondary (2D) representation enriches sequences with spatial information related to base pairing, providing insight into the tertiary relationships between nucleotides. Base pairings are distinguished into canonical and non-canonical ones.

Watson–Crick–Franklin (Canonical) pairing

Well-known and explored standard Watson-Crick–Franklin base pairs representation almost always takes into consideration non-overlapping pairs of Adenine-Uracil and Guanine-Cytosine. The interaction involves standard, strong, and hard-to-dissolve hydrogen bonding contacts; breaking such interactions requires a high amount of energy. Canonical pairings introduce high-level motifs into structures such as helices, stems, and loops. Secondary structure can be represented in both graphical and textual ways. The most straightforward representation is to use dots for non-interacting nucleotides and brackets to indicate interaction with the corresponding open and closed residues that form a base pair.
Example:

```
Sequence (5' to 3'):
UGCUCCUAGUACGAGAGGAACGGAGUG
.(((((..............)))))
```

Figure 1.2: RNA secondary structure (2D) given for RNA-Puzzle 11 challenge.

Figure 1.2 illustrates the secondary structure of an RNA hairpin loop, where a double-stranded stem is formed by canonical base pairing. The stem extends from the second nucleotide at the 5' end to the sixth residue before the loop on one strand, and from residues 22 to 26 on the complementary strand. At the bottom of the Figure 1.2, a single-stranded hairpin loop contains 15 unpaired nucleotides that remain flexible. The green curve traces the sequential order of the RNA strand from 5' to 3', while blue circles represent individual nucleotides (A, U, G, C).

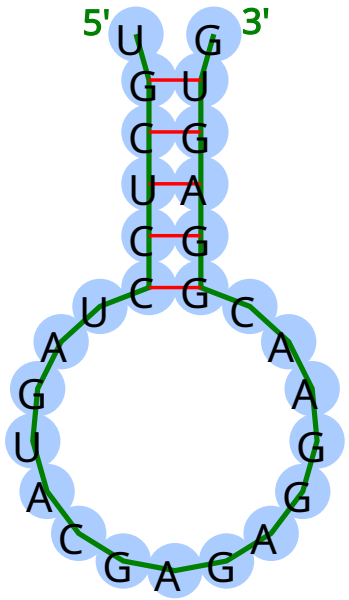


Figure 1.3: RNA secondary stucture given for RNA-Puzzle 11 challenge, graphical representation.

It is worth mentioning other properties that could be derived from the chart, such as sequence distance. The value is calculated based on the distance between residue serial numbers, as indicated by the green line between two selected nucleotides. It is primarily associated with pairing information to maintain the structural context.

To introduce a non-canonical pairing, it is necessary to introduce the chemical representation of RNA and the characteristics associated with molecular interactions.

1.3.2 Chemical view

RNA nucleobases can be represented purely using chemical compound descriptions. The chemical knowledge enables new approaches to extract features and interactions between biological components.

Basic units forming RNAs

Adeine

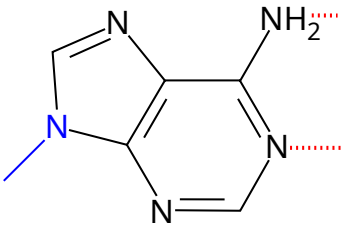


Figure 1.4: Chemical representation of Adeine; image from public domain.

Atoms within the compound form a pentagon with a hexagon two-ring structure. The compound consists of two atoms that are capable of forming hydrogen bonds.

Uracil

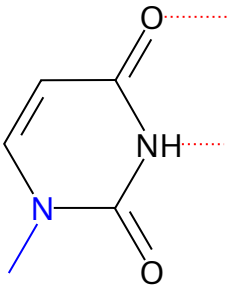


Figure 1.5: Chemical representation of Uracil; image from public domain.

In this case, the compound form consists of a single hexagonal ring. The compound consists of two atoms that are capable of forming hydrogen bonds.

Guanine

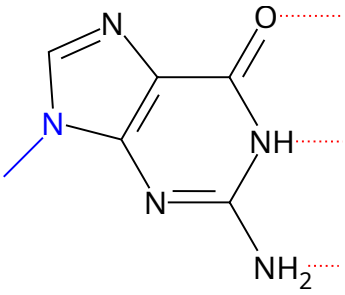


Figure 1.6: Chemical representation of Guanine; image from public domain.

The compound forms a similar structure to Adeine, but Guanine differs in having three atoms that are capable of forming stronger interactions, namely, hydrogen bonds.

Cythosine

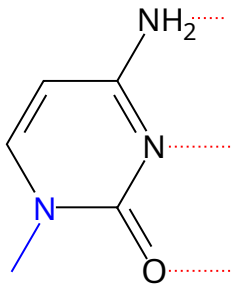


FIGURE 1.7: Chemical representation of Cythosine; image from public domain.

Cytosine shares the same structure as Uracil, creating a single hexagonal ring. The compound, having three free atoms for creating interactions, typically creates a strong connection with Guanine.

Sugar-phosphate backbone

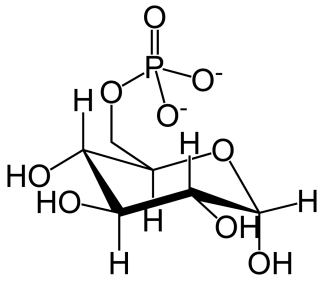


FIGURE 1.8: Chemical representation of Sugar-phosphate backbone; image from public domain.

Bonds

Each of the mentioned chemical compounds is made of atoms connected by bonds. The atom interactions can be distinguished into two groups. Covalent interactions are based on the sharing of electrons between atoms, creating the strongest possible known interactions. Alternatively, Ion interactions, which use electrostatic attraction between oppositely charged ions, create weak tangles.

The distances of such bonds can vary between pairs of atoms; therefore, properly modelling them is crucial for determining the final RNA structure fold.

Another type of bond (non-covalent one) is the hydrogen bond, where an H atom covalently bonded to an electronegative atom (N, O, or F) interacts with another electronegative atom, playing a crucial part in forming base pairs in biological structures.

Purines and Pyrimidine

The distinction was introduced to distinguish the direction of forming base pairing. Pyrimidine base nucleotides create a single-ring structure (a 6-membered ring, or hexagon), while purine creates a double-ring structure (a 6-membered ring, or hexagon, and a 5-membered ring, or pentagon) that shares one bond. In the RNA universe, nucleotides can be easily categorised into:

- Purins: Guanine (Figure 1.6) and Adenine (Figure 1.4),
- Pyrimidines: Uracil (Figure 1.5) and Cytosine (Figure 1.7).

RNA base pairs

Base pairing from a chemical point of view

From a Watson–Crick–Franklin base pairs point of view, nucleobases connect with their corresponding nucleobases, which are not identical, but have the same number of available atom slots for interaction.

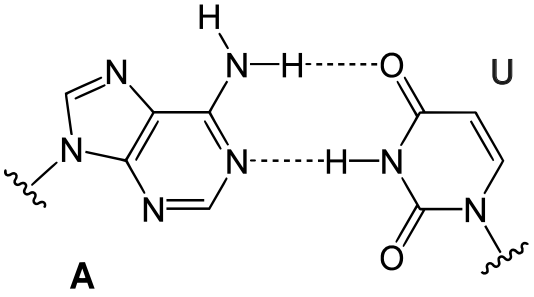


Figure 1.9: Formed pair of Adeine and Uracil [6].

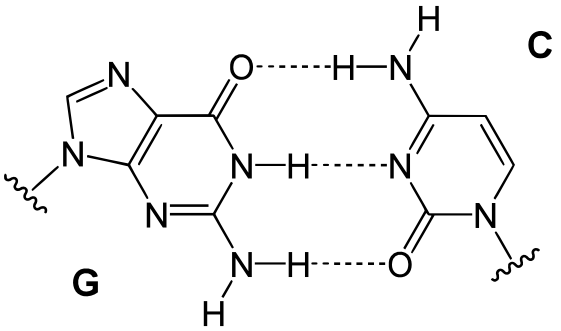


Figure 1.10: Formed pair of Guanine and Cythosine [6].

Figures 1.9 and 1.10 show formed hydrogen bonds (base pairing) between corresponding nucleobases. The waves symbolise attachment with a sugar backbone.

Non-canonical pairing

Non-canonical pairing extends the base knowledge that was associated with classic canonical pairing.

The Leontis-Westhof classification system [7] categorises non-canonical base pairs based on three main criteria:

- Edge combinations: Six possible pairings (W:W, W:H, W:S, H:H, H:S, S:S); Watson–Crick–Franklin (W), Hoogsteen (H), and Sugar (S).
- Glycosidic bond orientation: Cis or trans arrangements.
- Base identities: The specific nucleobases involved.

The rules governing non-canonical base pairing are more flexible than those of Watson–Crick–Franklin pairing. Nucleotides can form hydrogen bonds using alternative geometric arrangements and orientations, sometimes sharing atoms between multiple pairing interactions. This flexibility enables the formation of higher-order structures such as base triples (triads) and base quadruples (Figure 1.15). The distinction between purins and pyrimidines plays a crucial

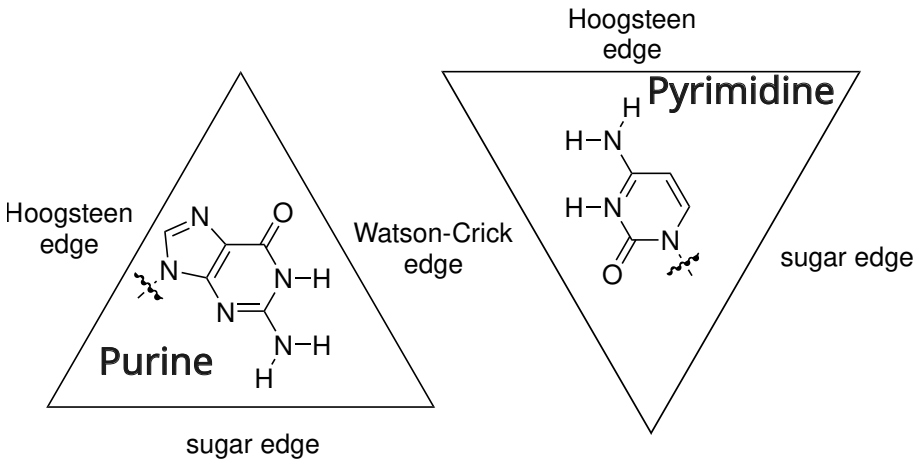


FIGURE 1.11: Edges of possible pairings for purines and pytimidines [8].

role. Base pairings can occur only between these groups. It is not possible to create hydrogen bonds between single purins and single pyrimidines.

Figure 1.11 presents the edges of nucleotides that can interact with each other, creating all six possible combinations of pairing.

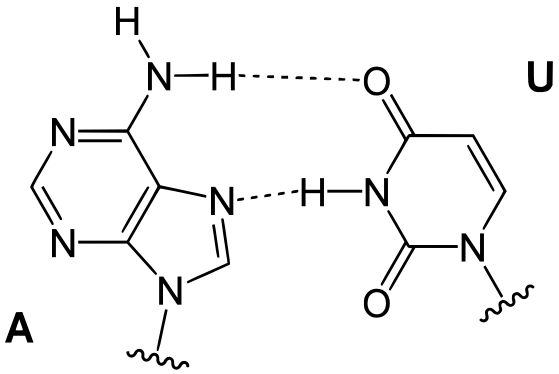


FIGURE 1.12: Adeine and Uracil base pairing creating H:W; Hoogsteen:Watson–Crick–Franklin; public domain.

The second property connected to non-canonical base pairing focuses on the orientation of pairing (cis and trans):

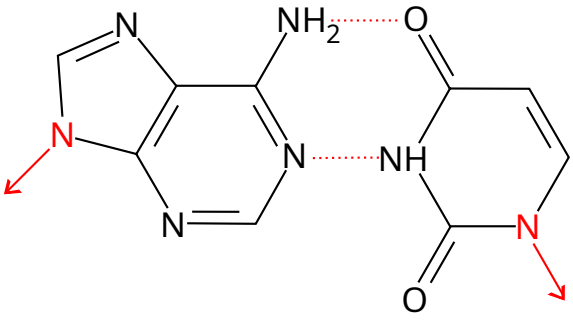


FIGURE 1.13: Cis (common) direction of base pair.

- cis (Figure 1.13): The glycosidic bonds of the two nucleotides are on the same side,
- trans (Figure 1.14): The glycosidic bonds are on opposite sides.

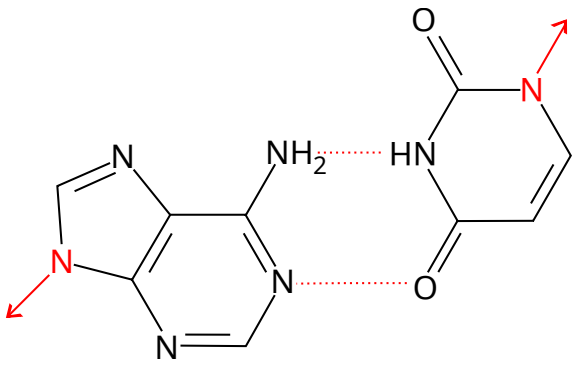


FIGURE 1.14: Trans (opposite) direction of base pair.

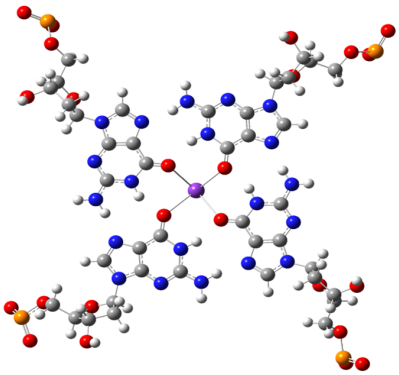


FIGURE 1.15: Example of quadruplex derived from PDB 1KF1 [6].

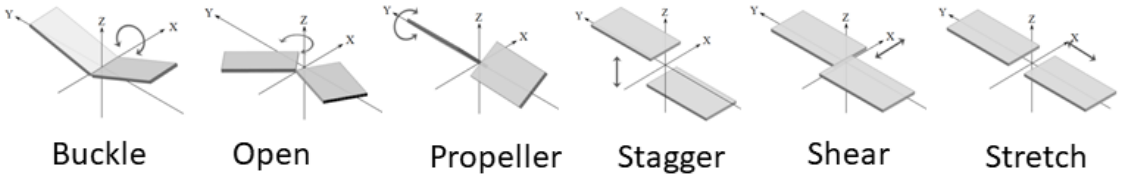


FIGURE 1.16: Parameters describing geometry of base pairs [6].

Rigid-body parameter relationships connected to angles within base pairing were explored and classified by the International Union of Pure and Applied Chemistry. They proposed parameters used to describe geometry (see Figure 1.16). The description provides a better understanding of the relationship without relying on an absolute Cartesian coordinate system.

The six parameters are organised into two fundamental categories reflecting their mathematical nature and structural significance. *Shear*, *Stretch*, and *Stagger* measure displacements in Ångström units, while *Buckle*, *Propeller*, and *Opening* quantify angular relationships in degrees [9].

Shear represents lateral displacement along the x-axis between paired bases, describing the sliding motion that optimises hydrogen bonding geometry. This parameter is particularly diagnostic for base pair classification. Watson-Crick–Franklin pairs show *Shear* values near zero, while G-U wobble pairs exhibit characteristic *Shear* values of -2.2 Å [10]. The parameter captures the fundamental geometric adjustment that allows non-canonical pairs to maintain stable hydrogen bonding despite altered base orientations.

Stretch measures separation along the y-axis, approximately corresponding to the direction of hydrogen bonding. This parameter reveals how bases adjust their proximity to optimise

electrostatic interactions. Watson-Crick–Franklin pairs maintain Stretch values near zero, but Hoogsteen A-U pairs show *Stretch* values around -3.5 Å [10], reflecting their dramatically altered hydrogen bonding geometry where bases approach from opposite sides of the major groove.

Stagger quantifies vertical displacement along the z-axis, measuring how bases are offset perpendicular to their planes. This parameter contributes to base pair non-planarity and affects stacking interactions with adjacent base pairs in the helical structure.

The rotational parameters describe angular relationships that influence RNA structure and stability. *Buckle* measures rotation about the x-axis, creating the "book-opening" motion that causes bases to be non-coplanar. *Propeller* describes rotation about the y-axis, generating the blade-like twisting that is crucial for optimising base stacking interactions. Watson-Crick–Franklin pairs typically show negative *Propeller* values around -11° [9], which enhance stacking interactions and contribute to helical stability.

Opening measures rotation about the z-axis [9], determining the angular relationship between bases in their plane. Watson-Crick–Franklin pairs maintain *Opening* values near zero, but Hoogsteen pairs can show *Opening* values around 66°[9], reflecting their fundamentally different hydrogen bonding geometry.

Angular representation RNAs

Bond angles are a group of structural parameters that describe the geometric configuration of RNA nucleotides, including their ribose sugar backbone and nitrogenous bases. Due to their relative nature, these measurements provide an effective way to characterise the three-dimensional conformation and local geometry of RNA structures. Bond angles are calculated from the spatial arrangement of three consecutive covalently bonded atoms.

Torsion angles are a more sophisticated description of nucleotides. The IUPAC standardised nomenclature system defines seven primary torsion angles per RNA nucleotide, each specified by four consecutive atoms. They require four adjacent bonded atoms to calculate two planes, and then the angle between them.

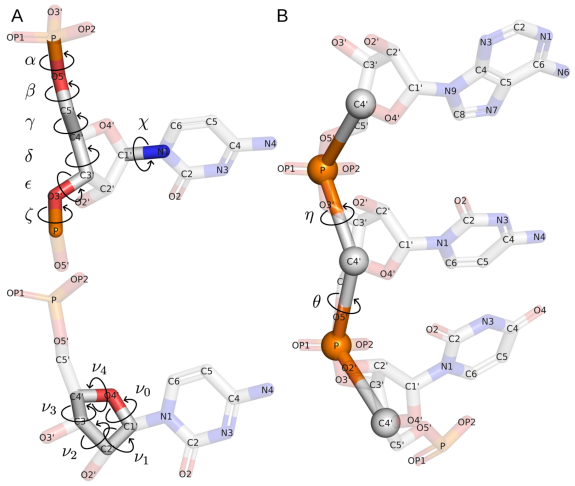


FIGURE 1.17: Torsion angles on the left (A) and pseudo-torsion angles on the right (B) in RNA structure [11].

Figure 1.17 illustrates the complete set of torsion angles in nucleic acid structures across three distinct panels. Panel A (upper left) displays the seven backbone torsion angles (α, β, γ,δ, ϵ, ζ) along the sugar-phosphate backbone chain, with the glycosidic angle χ defining the

TABLE 1.1: Torsion (TA) and pseudo-torsion (PA) angles defined for RNA [11].

TA	Atoms involved	TA	Atoms involved	PA	Atoms involved
α	O3'$_{n-1}$-P-O5'-C5'	χ	O4'-C1'-N1-C2 (pyrimidines)	η	C4'$_{n-1}$-P-C4'-P$_{n+1}$
β	P-O5'-C5'-C4'		O4'-C1'-N9-C4 (purines)	θ	P-C4'-P$_{n+1}$-C4'$_{n+1}$
γ	O5'-C5'-C4'-C3'	ν_0	C4'-O4'-C1'-C2'	η'	C1'$_{n-1}$-P-C1'-P$_{n+1}$
δ	C5'-C4'-C3'-O3'	ν_1	O4'-C1'-C2'-C3'	θ'	P-C1'-P$_{n+1}$-C1'$_{n+1}$
ϵ	C4'-C3'-O3'-O	ν_2	C1'-C2'-C3'-C4'		
ζ	C3'-O3'-P-O5'$_{n+1}$	ν_3	C2'-C3'-C4'-O4'		
		ν_4	C3'-C4'-O4'-C1'		

base-sugar orientation. The six backbone angles follow the phosphodiester linkage: α controls phosphate-sugar connectivity, β governs the O5'-C5' bond rotation, γ influences sugar pucker coupling, δ directly correlates with sugar conformation, ϵ links to the next phosphate, and ζ completes the backbone connection.

The glycosidic angle χ requires different atomic definitions for purines ($O4' - C1' - N9 - C4$) versus pyrimidines ($O4' - C1' - N1 - C2$). This angle determines whether bases adopt syn (-90° to 90°) or anti (90° to -90°) conformations relative to the sugar.

Panel B (right) illustrates the simplified pseudo-torsion angles θ and η, which provide two-dimensional representations of the backbone geometry. These virtual bond representations (built on nonconsecutive atoms) have proven essential for systematic conformational classification and structural validation [12]. The lower left panel presents the five sugar pucker angles (($\nu_0 - \nu_4$) that describe the endocyclic torsions within the ribose ring structure. Table 1.1 provides the precise atomic definitions for each angle, specifying which four consecutive atoms participate in each torsion measurement.

The comprehensive torsion angle parameters enable complete structural encoding of three-dimensional nucleic acid geometry through systematic description of conformational features. This torsional representation provides a sequential encoding framework that incorporates inter-nucleotide relationships, offering advantages over absolute Cartesian coordinate systems by capturing the intrinsic flexibility and relative positioning of structural elements within the molecular chain.

1.3.3 3D RNA structure determination methods

Only 4,422 RNA representative structures exist in the Protein Data Bank [13] compared to over 200,000 protein structures. There are two representation formats of 3D structures.

- **PDB**: legacy format; first presentation took place in the 80s, the format is characterised by strictly defined columns.
- **mmCIF/PDB**: new format; in contrast to the legacy format, it is flexible, enabling users to define their fields and tables with several columns aligned to user needs.

Three primary methods exist for experimentally determining 3D RNA structures. The cost of these operations remains significant to this day.

X-ray Crystallography

X-ray crystallography exploits the diffraction of electromagnetic radiation by periodic crystal lattices to determine atomic coordinates. When X-rays interact with electrons in a crystal, they undergo elastic scattering, producing diffraction patterns that are the Fourier transform

of the electron density distribution [14]. This provides the highest atomic precision, often achieving < 2 Å resolution.

Nuclear Magnetic Resonance Spectroscopy

NMR exploits the magnetic properties of atomic nuclei with non-zero spin angular momentum. Chemical shifts arise from electronic shielding effects, providing structural information about local environments. The method's outputs are the constraints that encompass the actual shape of the 3D structure coordinates [15].

Cryo-EM - Electron Microscopy

Cryo-EM preserves biological specimens in vitreous ice, maintaining near-native conformations while enabling high-resolution imaging. The contrast transfer function describes the relationship between the object wave function and the formation of an image [16]. Recent advances achieve near-atomic resolution for large macromolecular assemblies.

Comparative Analysis

Each method presents distinct trade-offs. Crystallography delivers superior atomic precision but requires challenging RNA crystallization, with expensive preparation yet established analysis protocols. NMR operates under native conditions and reveals dynamics but is limited to < 100 nucleotides, requiring costly isotopic labeling and sophisticated analysis. Cryo-EM avoids crystallization and handles large complexes with straightforward preparation, though demanding computational processing and struggling with molecules < 150 kDa.

1.4 Evaluation datasets

1.4.1 RNA-Puzzles

RNA-Puzzles is a collaborative scientific initiative designed as a collective experiment for blind 3D RNA structure prediction. The project operates by confidentially providing RNA sequences from solved structures to participating computational modelling groups, enabling them to predict the 3D structure without prior knowledge of the experimentally determined 3D structure. The primary goals include determining the capabilities and limitations of current computational methods for RNA structure prediction, assessing progress in the field, identifying bottlenecks that impede advancement, and promoting the availability of prediction tools to guide users in selecting appropriate methods.

1.5 Scope of the work

This study examines the potential of graph-based representations of three-dimensional RNA structures in machine learning applications, with a specific focus on their efficacy in training predictive models for structural analysis. It investigates the influence of contrastive learning techniques on enhancing model performance compared to conventional training methods, with an emphasis on evaluating model quality using the Root Mean Square Deviation (RMSD) metric. The process of assessing the accuracy of structural models when the actual 3D structure is unknown is referred to as General Model Quality Assessment. Current techniques for predicting 3D RNA structures assess models at various stages, but a singular, consistent

methodology proves ineffective. Consequently, researchers often face challenges in selecting the native-like 3D RNA structure from a multitude of available predictions.

The dataset utilised in this study comprises RNA structural data sourced from three distinct origins: two established, state-of-the-art databases and one proprietary dataset created explicitly for this research. The focus is solely on 3D RNA structures. This research utilises open-source software tools for extracting features from molecular structures and encoding them into graph representations.

The methodology involves developing a comprehensive computational pipeline that processes 3D RNA structures, converts them into graph format, and evaluates various machine learning approaches, including contrastive learning frameworks. Notably, the study does not address RNA folding prediction, molecular dynamics simulations, or experimental validation of structures. The theoretical framework is informed by literature on graph neural networks and structural bioinformatics, while computational constraints limit the analysis to structures containing fewer than 250 nucleotides.

The expected deliverable is a validated pipeline for RNA structure analysis that demonstrates improvements in RMSD performance metrics. Results will be tailored to the RNA structure domain and may not be directly applicable to other molecular structure prediction tasks.

Chapter 2

Related work

This chapter highlights the importance of identifying improved methodologies for managing and analyzing the 3D structure of RNA data.

2.1 Preliminaries

Preliminaries were divided into two groups:

- modelling - approach to model 3D RNA structure,
- quality assessment - comparing and scoring 3D RNA structures.

Both approaches need domain knowledge to be properly implemented.

2.2 SLR method

2.2.1 PICO

The Population, Intervention, Comparison, and Outcomes (PICO) framework proposed by Kitchenham and Charters [17] was developed to identify keywords during a literature review. The following keywords were used for querying knowledge sources

	main keyword	synonyms
Population	bioinformatics, machine learning, 3D RNA models	RNA geometry, RNA conformations,RNA models
Intervention	Quality assessment, 3D RNA modeling, structure validation	Quality control, 3D model analysis, conformational analysis
Comparison	Assessment metrics, RNA quality benchmarks, 3D structure comparison	Structural evaluation criteria, quality scoring, reference structures
Outcome	Assessment results, accuracy, structure quality scores	Evaluation outcomes, precision metrics, quality indices

2.2.2 Aim of the Systematic Literature Review (SLR)

The primary aim of this Systematic Literature Review (SLR) is to critically analyse and synthesise current methodologies, assessment metrics, and computational tools used for the quality evaluation of 3D RNA models.

This systematic literature review aims to map the existing body of research on 3D RNA quality assessment, addressing knowledge gaps. Additionally, this review aims to support the development of novel or improved methods for assessing the quality of the RNA model.

2.2.3 Research questions

To achieve these objectives, the SLR is guided by the following research questions:

RQ1 What state-of-the-art methods are used for the general quality assessment of 3D RNA structures?

RQ2 How do the current methods for RNA structure quality assessment compare in terms of accuracy, robustness, and ease of application?

RQ3 What approaches are utilised in various biochemical fields of study for applying Graph Neural Networks and regression models?

RQ4 Are there any challenges and limitations in the current approaches to 3D RNA quality assessment?

2.2.4 Search process

The niche related to this study is narrow. Therefore, using search engines to search for articles mainly yields irrelevant results or an empty list. Therefore, a more effective approach was to apply the snowball technique. The initial group of papers follows Rosetta [18], RASP [19], RNA KB potential [20], 3dRNAscore [21], DFIRE-RNA [22], RNA3DCNN [23], Atomic Rotationally Equivariant Scorer (ARES) [24], and lociPARSE [25].

The search process was completed with RNAGCN [26] being found. The work presents an approach similar to that presented in ARES, based on atomic interactions.

The method terminated once all findings had been exhausted.

Selection criteria

The application of the following criteria decides whether to consider such an entry/paper:

SC1 is it bioinformatics related [YES/NO]

SC2 is it motifs quality assessment/modelling [YES/NO]

SC3 is it based on graph representation [YES/NO]

SC4 is it accessible by PUT resources [YES/NO]

2.3 Current methods for RNA structure quality assessment compared in terms of accuracy, robustness, and ease of application?

RNAGCN does not provide instructions or weights for their graph neural network model.

ARES provides complete instruction with all the necessary tools to run the model.

The range of data quality used to train the ARES model has a peak shifted beyond the 5 Å RMSD, which depicts that the potential model will be best for analysing the model of the investigated range of interest, 0-10 Å.

lociPARSE also provides complete instructions with all the necessary tools to run the model. The lociPARSE training set clearly shows that the distribution mean is around 3 Å RMSD models in the training set. The conclusion could lead to the idea that lociPARSE will be better for assessing smaller RMSD due to a more precise dataset.

2.4 Graph Neural Network approaches in biochemical studies

Graph Neural Network framework [27] offers exceptional adaptability in diverse applications, from property prediction to molecular dynamics simulations [28]. Its hierarchical nature facilitates scaling to larger systems while maintaining local chemical accuracy. The ability to incorporate physical constraints and symmetries through specialised layers and loss functions represents a significant advantage over traditional methods.

Despite these advantages, several critical limitations warrant consideration. The first of these is the substantial data requirement for practical training, which is particularly challenging given the limited availability to replicating experimental datasets in materials science. The computational complexity of input preparation and model optimisation presents additional challenges, requiring significant computational resources and expertise.

Implementing GNNs requires expertise in both domain knowledge and deep learning. The approach suffers from limited interpretability of learned representations due to the non-linear nature of message passing.

2.5 Challenges and limitations of 3D RNA representations

Following studies presents different approaches in assessing 3D RNA structures:

- ROSETTA [18],
- RASP [19],
- RNA KB potential [20],
- 3dRNAscore [21],
- DFIRE-RNA [22],
- RNA3DCNN [23].

Those approaches utilise domain knowledge to determine whether 3D RNA structure is native-like. The main approach presented in the papers is the minimisation of the energy difference between atoms.

ROSETTA combines low-resolution conformational sampling through fragment assembly with a sophisticated full-atom refinement phase, leveraging a physically realistic energy function that includes critical atomic interactions. The approach is efficient for motifs up to 12 nucleotides in length, accurately capturing complex features such as non-canonical base pairs and irregular backbone conformations. Its limitation manifests through multiple symptoms of poor conformational sampling, including non-convergence of the lowest-energy models, an inability to sample conformations proximate to the native state, and failure to achieve energy levels comparable to those of the native state. These sampling inadequacies become progressively more pronounced as motif size increases, suggesting fundamental constraints in the method's ability to handle larger RNA structures.

RASP is a knowledge-based potential derived from a carefully curated set of non-redundant 3D RNA structures. *RASP*'s architecture incorporates both local and non-local interactions through an information theory-optimised framework, enabling accurate discrimination between native and non-native conformations. The full-atom variant of the potential (*RASP-ALL*) demonstrates remarkable sensitivity in capturing both canonical base pairs and non-canonical interactions that are prevalent in functional RNA molecules. *RASP* successfully identified structural perturbations in experimental systems, including the effects of destabilising mutations in the self-splicing group.

RASP has several noteworthy limitations. The primary among these is imposed by the relatively small set of experimentally determined RNA structures. The statistical potential's derivation from merely 85 non-redundant structures, necessitated by the technical challenges associated with RNA structure determination, potentially circumscribes its generalizability.

A significant methodological limitation lies in the absence of a solvation term within the current framework. This omission potentially compromises the accuracy in evaluating highly unfolded or non-compact RNA conformations. This limitation is particularly salient given the established importance of solvation effects in 3D RNA structures. Furthermore, the current implementation cannot minimise energy scores, a feature that has demonstrated substantial utility in alternative approaches, such as *ROSETTA* [18].

The RNA KB potential presents an approach to RNA structure evaluation through the development of fully differentiable knowledge-based (KB) potentials in both coarse-grained and all-atom representations. The authors derived these potentials from a carefully curated dataset of 77 high-resolution 3D RNA structures, employing sophisticated statistical methods to handle regions with low counts. The distinctive feature of the methodology lies in its distance-based potential framework, which implicitly incorporates various RNA interactions without requiring explicit parameterisation of individual components.

The methodology exhibits several notable constraints that warrant consideration. First, the relatively small training dataset of 77 3D RNA structures, although carefully curated, may limit the model's potential to capture the full spectrum of RNA structural diversity. This limitation is particularly significant given the growth in newly discovered RNA structures and motifs. Second, the implicit treatment of electrostatics and solvent effects, while computationally efficient, represents a simplified approach that may not fully capture the complex physicochemical interactions governing 3D RNA structure, particularly in cases involving specific ionic conditions or unusual solvent environments.

3dRNAscore is another knowledge-based potential for evaluating RNA tertiary structures that uniquely combines energy dependencies. The method was developed using a dataset of 317 non-redundant 3D RNA structures. The potential's distinctive feature is based on information from the backbone torsion angles alongside traditional distance-based measurements, enabling a more comprehensive evaluation of the structural characteristics of RNA.

The training set of 317 structures may not fully capture the entire spectrum of 3D RNA structural diversity, potentially leading to bias in the statistical potential. The method's treatment of electrostatic interactions also represents a significant simplification of the complex physicochemical reality. RNA molecules are susceptible to electrostatic interactions due to their negatively charged phosphate backbone. The approach is not able to accurately score such 3D structure examples.

DFIRE-RNA is an all-atom distance-dependent knowledge-based potential for RNA structure evaluation, derived from the distance-scaled finite ideal-gas reference state (DFIRE) framework. The method employs a sophisticated statistical approach utilising 405 non-redundant

RNA structures for training, with implementation of an ideal gas reference state that has proven successful in protein 3D structure prediction.

The primary limitation stems from the method's exclusive reliance on distance-dependent parameters while neglecting crucial orientation-dependent interactions.

RNA3DCNN is a deep learning approach for RNA structure quality assessment utilising 3D convolutional neural networks. The method was trained using 414 non-redundant RNA structures. The approach employs a unique strategy to evaluate 3D RNA structures at both local and global levels by examining individual nucleotides within their spatial context. Two distinct models were developed: one for evaluating near-native structures and one for assessing broader structural spaces. The approach innovatively transforms RNA structural data into a 3D grid representation with three channels (atomic occupation, mass, and charge), enabling direct processing by 3D CNNs without manual feature extraction.

lociPARSE is a deep learning method for scoring RNA 3D structure quality that adapts AlphaFold2's Invariant Point Attention (IPA) architecture with locality-aware modifications. Instead of predicting traditional RMSD metrics, the model trains on Local Distance Difference Test (lDDT) scores [29], which are superposition-free and capture the accuracy of the local atomic environment. The architecture uses k-nearest neighbour information and edge-biased attention to focus on spatial proximity between nucleotides, while maintaining invariance to global rotations and translations. The model predicts both nucleotide-wise and molecular-level quality scores, significantly outperforming existing statistical potentials and machine learning approaches on RNA structure prediction benchmarks, including CASP15 targets.

The central limitation of the method is linked to the increasing computational power required for processing 3D RNA structures, primarily due to the need to perform a grid cube calculation. The generalisation is based only on a small number of processed structures. Therefore, the performance is dependent on only a small subset of previously seen schemes.

Chapter 3

Data description and applied technologies

This chapter focuses on describing a feature extraction pipeline enhanced with third-party tools.

Proper feature extraction from biological data can be challenging due to the complexity of atomic interactions. The most important fact related to training data is that the model will be as good as the data on which it was trained. Therefore, a significant amount of effort in the thesis was devoted to preparing, testing, and validating the correctness of the data structures achieved. The goal was to prepare a comprehensive description of RNA nucleotides, encompassing all possible descriptions and features that can be extracted from the atomic coordinates.

3.1 Representative set of non-redudant 3D RNA structures

The challenge of preparing high-quality data for training purposes is connected to the proper distribution of RMSD values. Improper distribution of values implies improper results for the potential model. Root of issues connected with a wrongly understood 3D RNA structure pattern or associated with a higher RMSD value. Preparation of a dataset characterised by a constant distribution of RMSD values is a significant challenge [30][31].

3.1.1 Creating diverse dataset

Naive elements movement

The mmCIF/PDB file consists of atom coordinates forming a larger structure; however, this geometric manipulation approach is not proper for simulating real-world cases as it violates fundamental physical constraints including bond lengths and angles that create unrealistic covalent geometries, steric clashes when atoms are placed too close to each other, disrupted electrostatic interactions and hydrogen bonding patterns crucial for base-pairing, and broken sugar-phosphate backbone continuity. The approach also ignores the loss of structural cooperativity inherent in RNA folding, where cooperative interactions between distant regions and allosteric networks mean that local changes propagate throughout the entire structure, effects that geometric manipulation completely disregards. The only advantage of this approach is the control over RMSD values, which can be easily manipulated by adjusting the positions of structural elements; however, this benefit comes at a significant cost to physical realism.

Prediction tool use

Several scientists have investigated the problem of predicting 3D RNA structure. AlphaFold 3 [32], representing a significant breakthrough, extends its predictions beyond proteins to predict RNA, DNA, and protein-nucleic acid complexes accurately. Boltz-1 [33] is another recent deep learning model that shows promising results for RNA structure prediction. RNA-Composer [34] employs fragment assembly methods and has proven to be a reliable tool for generating RNA 3D structures from sequence data. Outputs from these modellers are most probably inputs for the potential scorer. Nevertheless, the approach comes with drawbacks. These predictors are resource-demanding; to generate a significant number of such models, it is necessary to employ high-computation processing machines. Another drawback is the constant and adjustable generation of the RMSD range. The models are deterministic computation tools. Our experiments showed no significant change in results after modifying the model's parameters.

Molecular Dynamics tools

The group of simulation tools that mimics the world of inter-atomic interactions. The representative of the group is OpenMM [35], a molecular dynamics simulation toolkit that creates a computational environment to simulate real-world interatomic interactions. The tool implements classical mechanics using force fields (AMBER [36], CHARMM [37], OPLS [38]) that mathematically model inter-atomic forces, including bonded interactions (bonds, angles), non-bonded interactions, and specialised terms for hydrogen bonding. The software propagates changes through time while conserving energy and momentum, and modifies chemical structures.

3.1.2 Considered datasets

During the work, experiments were conducted using three different datasets: lociPARSE, ARES, and RNAQuANet, which represent different training/validation sets.

lociPARSE dataset

Dataset comprising training and test sets of 30 independent RNAs sourced from trRosettaRNA [39], along with CASP15 [40] experimental structures and all submitted predictions downloaded from the CASP15 competition platform. Additionally, an their in-house curated set of 60 non-redundant RNA targets was employed for hyperparameter optimisation. To generate 3D RNA models for training and validation, the researchers used both deep learning and traditional methods, including six deep learning approaches (DeepFoldRNA [41] predicting six models, trRosettaRNA with ten models, RoseTTAFoldNA [42] with one model, RhoFold [43] with one model, and DRfold [44] with six models) and one molecular dynamics method. The dataset was further augmented through PyRosetta [45] perturbation techniques, where DeepFoldRNA decoys were perturbed using the FastRelax [46] program with both 5,000 and 10,000 iterations, generating two distinct perturbation variants per original decoy, resulting in a total of 12 additional decoys. This systematic approach yielded 37 models per sequence target, culminating in a total of 51,763 models for the comprehensive evaluation of 3D RNA structure prediction methods. This method combined two approaches and utilised molecular dynamics and prediction tools. The lociPARSE set exhibits the best distribution of decoy

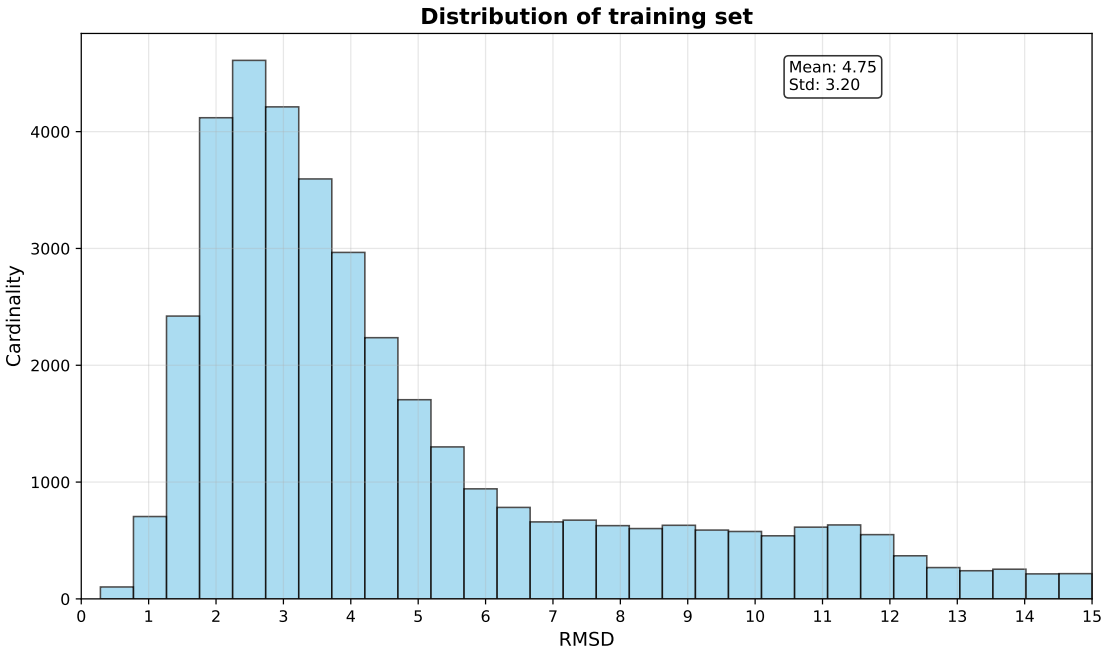


Figure 3.1: Distribution of lociPARSE training set.

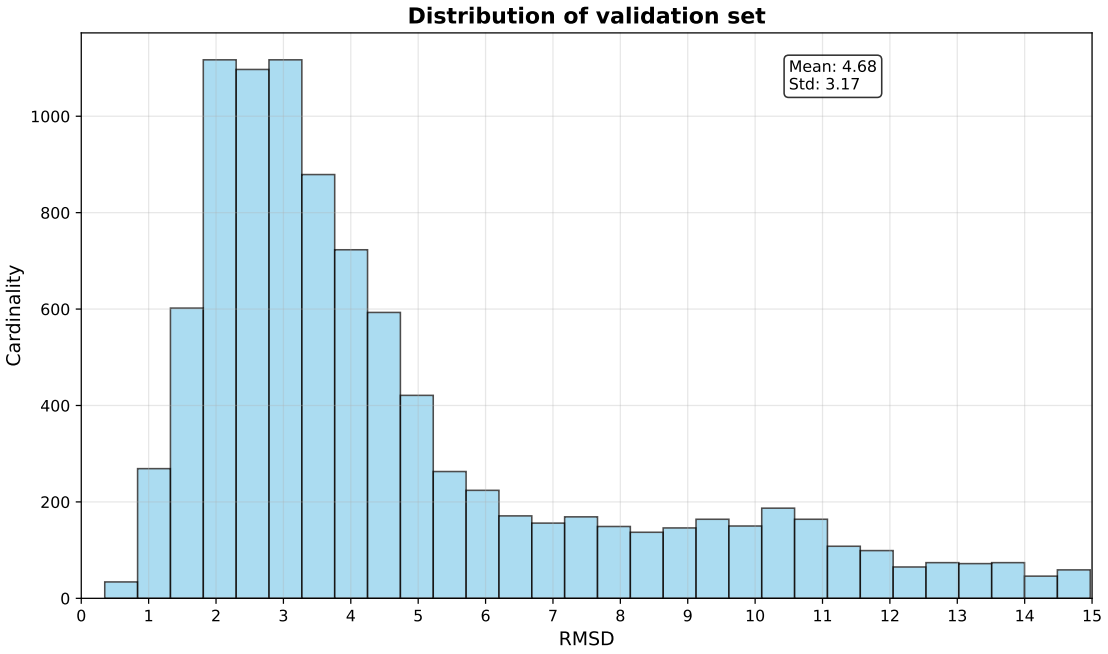


Figure 3.2: Distribution of lociPARSE validation set.

RMSDs, with the peak of the distribution located close to 2.5 Å, which is presented on Figures 3.1, 3.2.

ARES dataset

The researchers created their dataset using a remarkably minimal approach; ARES was developed based on only 18 experimentally determined 3D RNA structures. To make a diverse dataset of models with various RMSD values, they utilise the Rosetta FARFAR2 [47] sampling method. The method employs the Monte Carlo method, which generates a spectrum of solu-

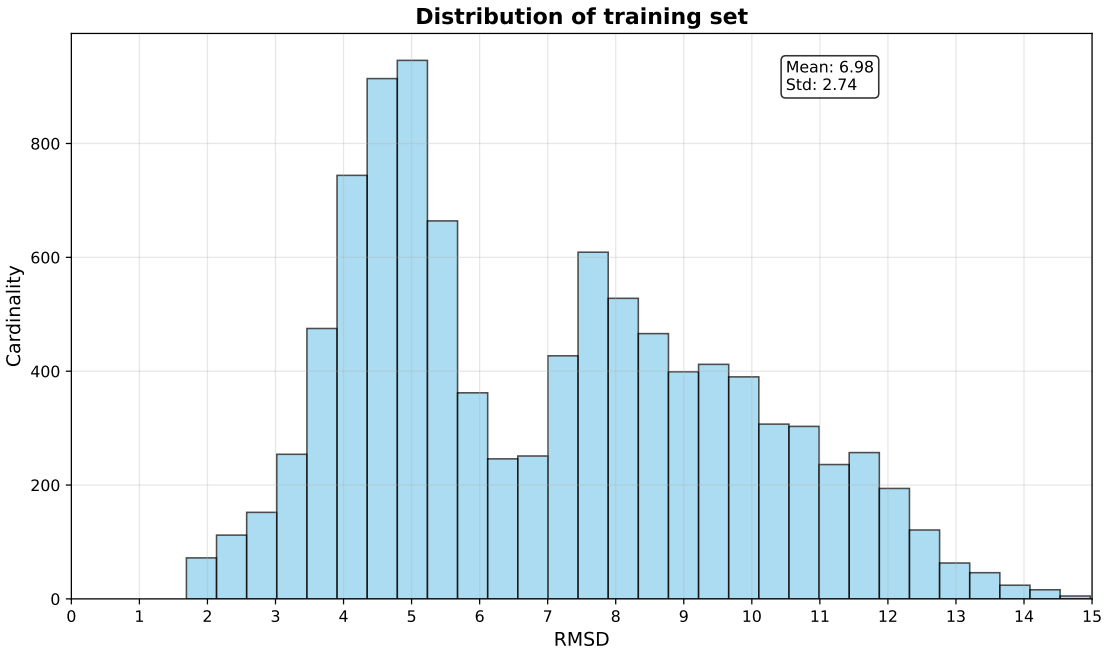


FIGURE 3.3: Distribution of ARES training set.

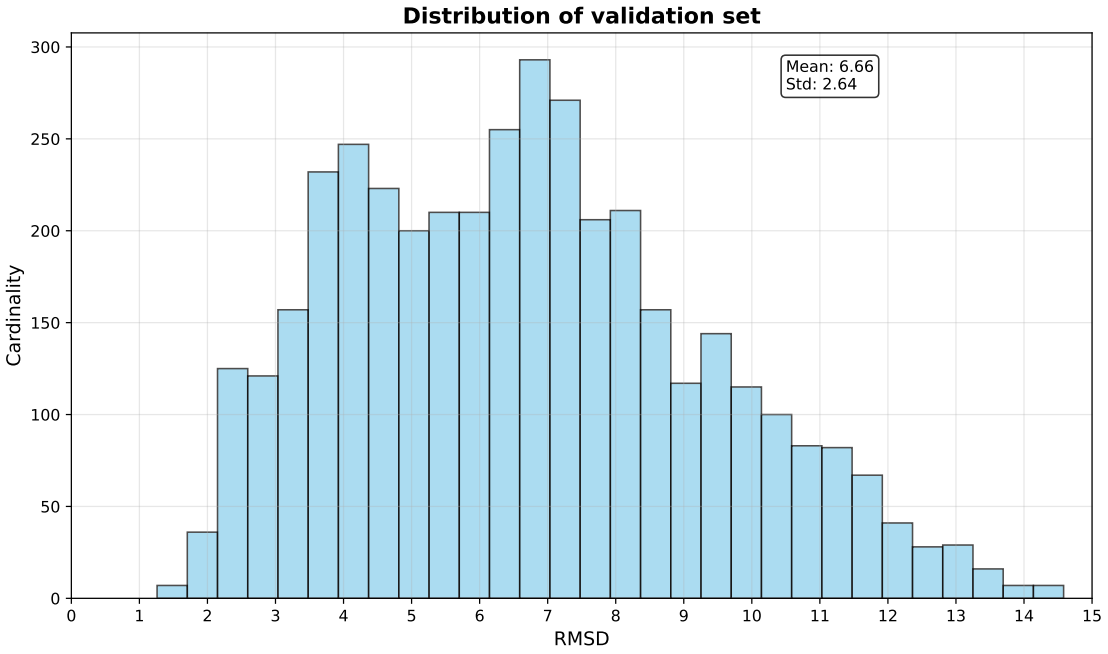


FIGURE 3.4: Distribution of ARES validation set.

tions aimed at minimising the internal score. The FARFAR2 is a fragment-based assembly method that, in addition to generating results, utilises a database of known fragments and combines them to create the final structure. Thanks to the approach, it is possible to predict a significant number of structures with various RMSD values. One drawback comes with limited accuracy. This method is not capable of creating a low RMSD accurate structure. The researchers from the ARES team generated 1,000 structural models of each experimentally determined 3D RNA structure by utilising this method. Thanks to the method's characteristic, model generation is conducted in a blind, reference-less manner. The result can create

both a native-like structure and a completely unrealistic one. The ARES set has the worst distribution of RMSD values. According to Figures 3.3, 3.4 the peak of the distribution lies around 7 Å, which does not include 3D RNA structures that are easy to compare with each other for finding the best conformation compromises.

RNAQuANet dataset

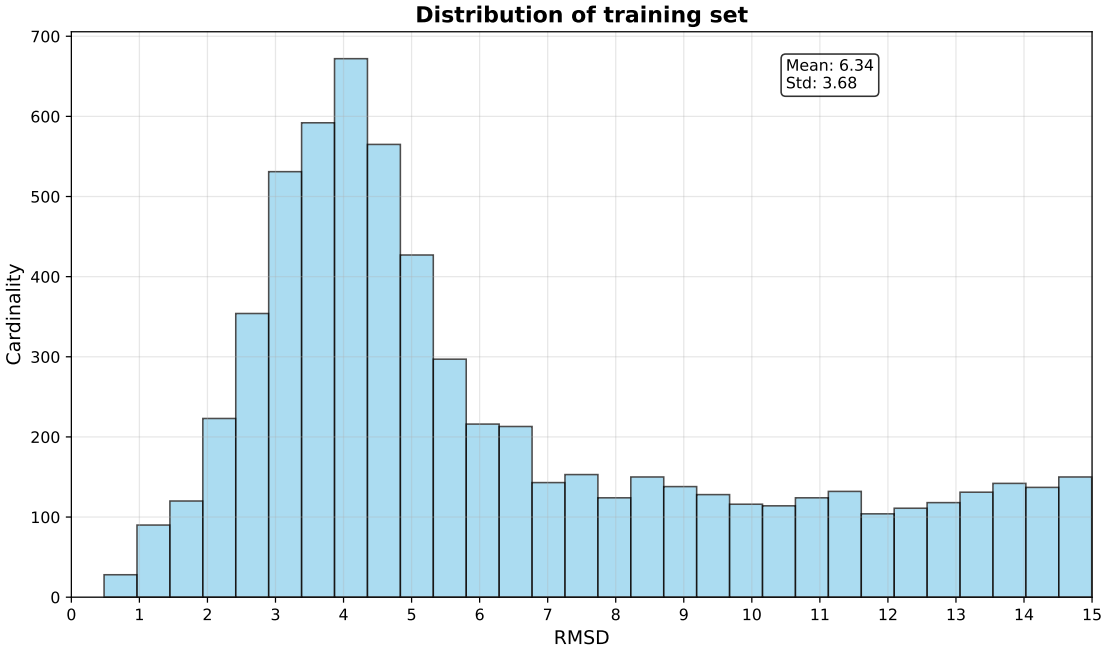


FIGURE 3.5: Distribution of RNAQuANet training set.

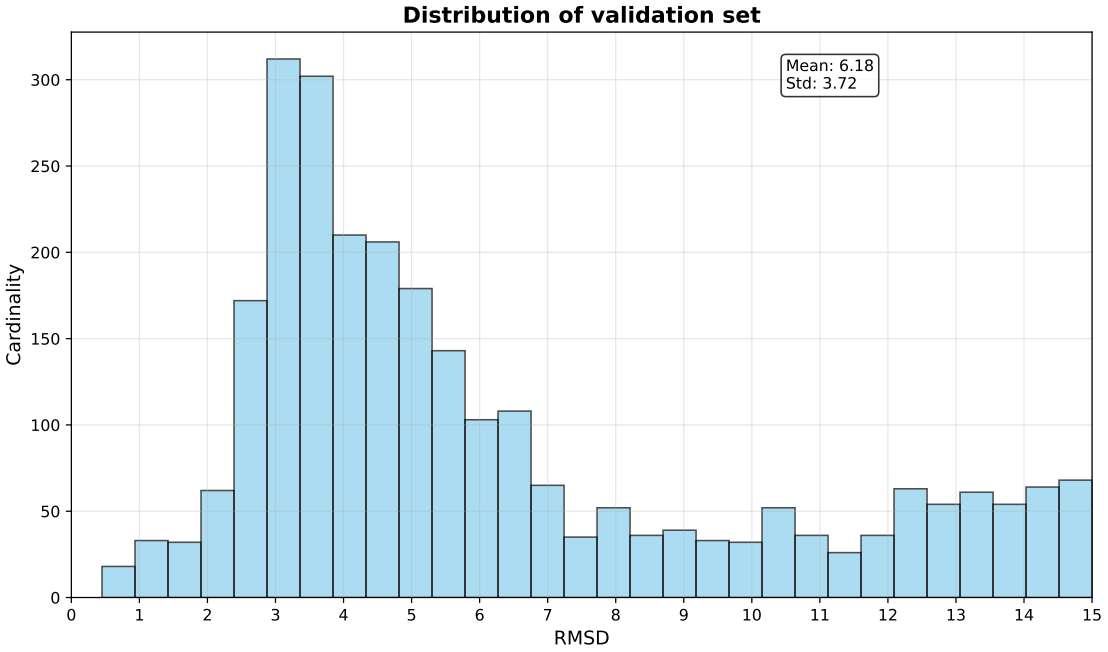


FIGURE 3.6: Distribution of RNAQuANet validation set.

The RNAQuANet dataset was compiled from a non-redundant high-resolution (<3 Å) col-

lection of 3D RNA structures extracted from the RNAsolo database [31], which contains RNA-cleaned 3D structures generated through various experimental methods. From an initial set of 1,840 structures, 737 passed through a rigorous filtering process that required: 3D RNA structure resolution $\leq$ 3 Å to ensure high quality, at least one base pair included, at least one unpaired nucleotide, the number of paired nucleotides being greater than or equal to unpaired nucleotides, and structure sizes between 10-190 nucleotides. Each reference structure was used to generate up to 30 alternative 3D RNA conformations using RNAComposer with diverse secondary structure predictors (RNAfold [48], CONTRAfold [49], ContextFold [50], CentroidFold [51], IPknot [52], RNAstructure [53], RNAshapes [54], HotKnots [55]), creating a dataset with RMSD values ranging from 0.268 to 58.753 Å and sequence lengths from 10 to 190 nucleotides. The final dataset was split into training (455 references, 10,890 3D models), validation (161 references, 3,968 models), and test sets (120 references, 2,932 models) using K-means clustering [56] to ensure balanced distribution of structural characteristics. The distribution of the RNAQuANet set is similar to that of lociPARSE, but is slightly shifted towards higher RMSD values, according to histograms presented on Figures 3.5, 3.6.

3.1.3 Expectations

The lociPARSE dataset appears to be the most promising for training purposes due to its wide distribution of RMSD values. The smaller the RMSD value, the better the model is at recognising subtle relationships between nucleotides. With a higher value comes a broader spectrum of possible solutions and a weaker typical pattern, which is a key clue in the training process. After numerous training attempts, the lociPARSE dataset was identified as the most effective source of knowledge.

3.2 Feature extraction process

The final prepared input to the RNAQuANet model consists of a graph with **35** edge features and **79** node features.

The primary challenge in generating a pipeline for feature extraction was determining how to handle varying numbers and modalities across different nucleotide feature sets. There were investigated method that solves that problem. The regular concatenation of features could result in a very long feature vector, which could be challenging for the trainer due to the curse of dimensionality. This approach could also introduce a problem with defining a proper null value for such fields, which do not describe particular nucleotides. In most cases, 0 is a relevant value for many features, such as angles and distances. Therefore, setting the proper value could be crucial for the entire model development, but the potential model should be able to address the underrepresentation of particular fields. This approach, rather than a solution, could cause the model to ignore fields written in these fields.

Another approach that was opted out of was compressing features into a smaller feature vector by replacing unused features by particular nucleobase for each nucleotide configuration. That could partially solve the previous problem, but it could introduce a new one where the model assumes that these features are somehow related to each other. Finally, the feature could be ignored again, due to a change in the feature domain.

The most promising idea introduced to the final solution involves splitting node features into two sub-feature spaces. The common for each nucleotide, and specific to the particular nucleobases.

- 59 Common features:
  - Bond distances: C1'-C2', C5'-C4', C2'-O2', C2-N3, C3'-O3', C6-N1, O4'-C1', O5'-C5', C4'-C3', P-O5', C3'-C2', C4'-O4', N3-C4, OP2-P, N1-C2, O3'-P, OP1-P, C5-C6,
  - Bond angles: O3'-P-O5', O3'-P-OP1, C4'-O4'-C1', C1'-C2'-C3', C6-N1-C2, C1'-C2'-O2', C3'-O3'-P, N1-C2-N3, O5'-C5'-C4', C5'-C4'-O4', P-O5'-C5', C5-C6-N1, O4'-C1'-C2', N3-C4-C5, C2'-C3'-C4', C4'-C3'-O3', C5'-C4'-C3', O3'-P-OP2, O5', C3', OP1, C5', C4', OP2, P, O3',
  - Torsion angles: C4'-O4'-C1'-C2', C4-C5-C6-N1, C3'-O3'-P-OP2, O3'-P-O5'-C5', C2-N3-C4-C5, C1'-C2'-C3'-C4', C5'-C4'-C3'-O3', O5'-C5'-C4'-C3', C6-N1-C2-N3, N3-C4-C5-C6, O5'-C5'-C4'-O4', C3'-O3'-P-OP1, C3'-C4'-O4'-C1', C2'-C3'-C4'-O4', C5'-C4'-O4'-C1', N1-C2-N3-C4, C3'-O3'-P-O5', C5-C6-N1-C2, P-O5'-C5'-C4', O4'-C1'-C2'-O2', C4'-C3'-O3'-P, C5'-C4'-C3'-C2', O4'-C1'-C2'-C3',
  - Structure size (the number of nucleotides),
  - One-hot encoded nucleobase [AUGC].
- 31 Adeine-specific features:
  - Bond distances: C1'-N9, C6-N6, C4-N9, C5-C4, N7-C5, N9-C8, C8-N7,
  - Bond angles: N3-C4-N9, C1'-N9-C8, O4'-C1'-N9, C5-C4-N9, C1'-N9-C4, C8-N7-C5, N9-C8-N7, N7-C5-C6, C5-C6-N6, N7-C5-C4,
  - Torsion angles: N7-C5-C6-N1, C1'-N9-C8-N7, C4-N9-C8-N7, C5-C4-N9-C8, N9-C8-N7-C5, O4'-C1'-N9-C4, N7-C5-C6-N6, C2-N3-C4-N9, C8-N7-C5-C6, N7-C5-C4-N9, C1'-N9-C4-N3, O4'-C1'-N9-C8, C4'-O4'-C1'-N9, C8-N7-C5-C4,
- 18 Cithosine-specific features:
  - Bond distances: C2-O2, C4-C5, C1'-N1, C4-N4,
  - Bond angles: N1-C2-O2, C4-C5-C6, O4'-C1'-N1, C1'-N1-C6, C1'-N1-C2, C2-N3-C4, N3-C4-N4,
  - Torsion angles: C1'-N1-C2-O2, O4'-C1'-N1-C6, C2-N3-C4-N4, C1'-N1-C6-C5, C1'-N1-C2-N3, C4'-O4'-C1'-N1, O4'-C1'-N1-C2.
- 18 Uracil-specific features:
  - Bond distances: C2-O2, C4-O4, C4-C5, C1'-N1,
  - Bond angles: N1-C2-O2, C4-C5-C6, N3-C4-O4, O4'-C1'-N1, C1'-N1-C6, C1'-N1-C2, C2-N3-C4,
  - Torsion angles: C1'-N1-C2-O2, O4'-C1'-N1-C6, C1'-N1-C6-C5, C1'-N1-C2-N3, C4'-O4'-C1'-N1, C2-N3-C4-O4, O4'-C1'-N1-C2.
- 34 Guanine-specific features:
  - Bond distances: C1'-N9, C6-O6, C4-N9, C5-C4, N7-C5, C2-N2, N9-C8, C8-N7,
  - Bond angles: N3-C4-N9, C1'-N9-C8, C5-C6-O6, O4'-C1'-N9, C5-C4-N9, C1'-N9-C4, C8-N7-C5, N9-C8-N7, N7-C5-C6, N1-C2-N2, N7-C5-C4,
  - Torsion angles: N7-C5-C6-O6, N7-C5-C6-N1, C1'-N9-C8-N7, C4-N9-C8-N7, C5-C4-N9-C8, N9-C8-N7-C5, O4'-C1'-N9-C4, C2-N3-C4-N9, C8-N7-C5-C6, N7-C5-C4-N9, C1'-N9-C4-N3, O4'-C1'-N9-C8, C6-N1-C2-N2, C4'-O4'-C1'-N9, C8-N7-C5-C4.

The primary challenge in generating a pipeline for feature extraction was determining how to handle varying numbers and modalities across different nucleotide feature sets. There were investigated method that solves that problem. The regular concatenation of features could result in a very long feature vector, which could be challenging for the trainer due to the curse of dimensionality. This approach could also introduce a problem with defining a proper null. The current solution assumes an input vector of standard features. The problem was solved using an autoencoder, an architecture that attempts to embed a feature vector into a shorter one and then expand it back to its original size. Based on the RNAQuANet dataset, four autoencoders were prepared, each for a different nucleobase. This approach, in result of training optimisation, gives a vector of 7 elements, which are later concatenated with the standard part of the feature vector. This solution not only compresses the input vector but also preprocesses and identifies correlations in the given set of features, which facilitates final model training.

In addition to the aforementioned features, one-hot encoding was used for each nucleobase and the structure size.

The approach resolved all possible nullable features within the description vector and provided stable values for all feature components.
The following tools were used to extract features from mmCIF files:

- RNAgrowth tool to extract nucleobase features and all of the angles [57],
- x3DNA-DSSR tool to extract base pairing characteristics and classification [58],
- PDBfixer to solve the problem with atom incompleteness [35].

The pipeline is operating using an HTTP server packed into a single Docker image file. It can therefore be run smoothly without any dependency issues. The result of the feature extraction is encoded into a single file graph declaration suitable for use within PyTorch Geometric and numpy.

Chapter 4

Architecture - diverse approaches

The literature mentioned in Chapter 2 presents a broad view of possible approaches to solve the referenceless 3D RNA quality assessment problem. The older methods mentioned, such as RNA KB, and RASP, employ an analytical approach to this problem by modelling the physical environment around RNA nucleotides and simulating fundamental atomic forces present between atoms. These physics-based methods evaluate structural quality by calculating energy-like scores based on known principles of molecular interactions, including electrostatic interactions, hydrogen bonding, and steric clashes. Essentially, these tools assess the realism of 3D RNA structures by determining whether the predicted conformation is energetically favourable according to established physical and chemical principles. This examination involves factors such as appropriate atomic distances, proper hydrogen bond formation, realistic bond angles and lengths, and overall structural stability. The passage appears to be establishing a foundation for contrasting these traditional physics-based approaches with newer methodologies that may take more data-driven approaches to the same general quality assessment challenge.

4.1 Deep learning approach

4.1.1 3D convolutional neural network

The newer approaches presented in Chapter 2 utilise machine learning methods. Convolution in deep learning is a fundamental mathematical operation that forms the backbone of convolutional neural networks (CNNs) [59]. At its core, convolution involves sliding a small matrix, known as a filter or kernel, across an input (such as an image) and computing dot products at each position to produce a feature map. The filter acts as a pattern detector.

RNA3DCNN utilises a 3D convolutional neural network, which was the first published model based on convolution for the problem considered [23]. In general, convolutional neural network layers are found to be the most effective use case within the field of image and video processing. CNN's main advantage is its invariant context, which is one of the key requirements that need to be addressed during the 3D RNA assessment. The RNA3DCNN model employs a small architecture comprising four 3D convolutional layers with 8, 16, 32, and 64 filters, respectively. The first two layers utilise a 5x5x5 kernel, while the last two use a 3x3x3 kernel. Following the first two consecutive convolutional layers, a max-pooling layer with a stride of 2 is applied. The network then includes one fully connected layer with 128 hidden units, followed by the final output layer that produces a single nucleotide unfitness score. The input to the network is a 32×32×32 voxel 3D image 4.1 with three channels representing the

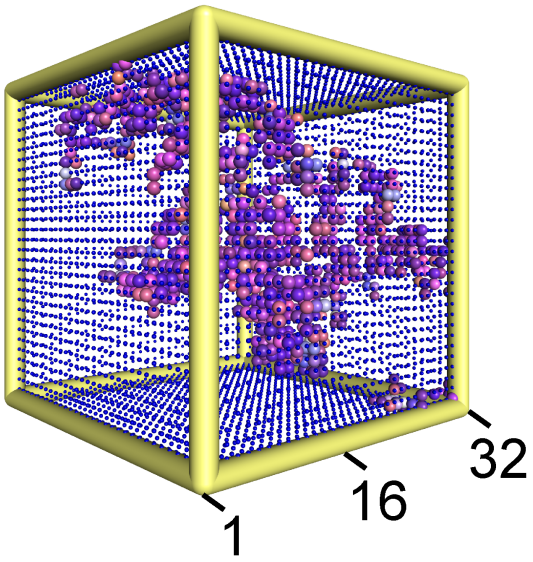


Figure 4.1: Visualisation of input data for RNA3DCNN [23].

atomic occupation number, mass, and charge within a local environment surrounding each nucleotide. All hidden layers use ReLU activation functions, and the output layer is linearly activated. The network contains a total of over 4 million parameters. The proposed model takes as input directly a 3D grid representation without any feature extraction. This method represents a very simple approach; nevertheless, the technique was superior to the state of the art at the time.

4.1.2 Convolutional Graph Neural Network

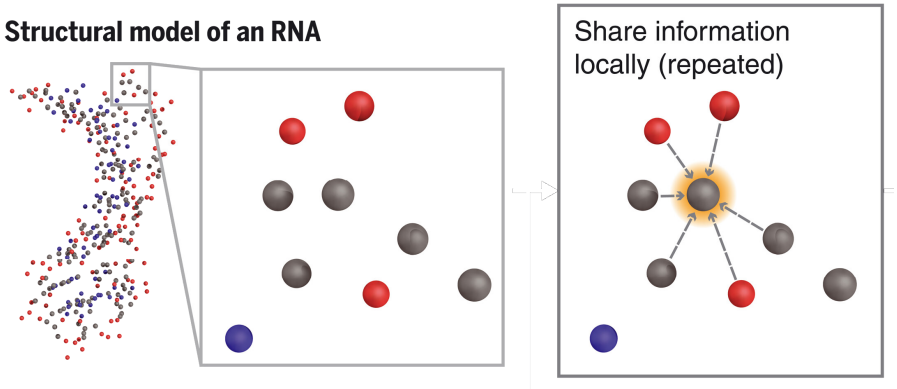


Figure 4.2: Visualisation of input data for ARES [24].

Graph Neural Networks (GNNs) are a type of deep learning architecture that utilises nodes and edges to model relationships between data entities. Data can be stored on both edges and nodes. The primary difference between regular convolution and graph convolution lies in the way they model neighbourhoods. Graphs do not limit the number of connected nodes, enabling the description of specific relationships between data nodes. In graphs, not only do features play an essential role, but paths do as well. A single graph can consist of many connected subgraphs. While data flow between subgraphs is less significant, it still occurs, creating separate environments that can interact with each other while remaining primarily focused on internal interactions. This characteristic contributes to the success of the ARES architecture

[24]. The researchers proposed an architecture that treats a raw cloud of points as nodes, with edges representing distances between atoms. This approach successfully identifies the existence of nucleotides and their relationships, such as base pairing. However, this architecture suffers from inadequate analysis of larger structures due to the fine granularity of data. Information can only be mixed within neighbourhoods, which works well until we realise that most data flow remains confined within individual nucleotides. A natural next step would be to adopt a coarser-grained data representation.

4.1.3 Transformer architecture

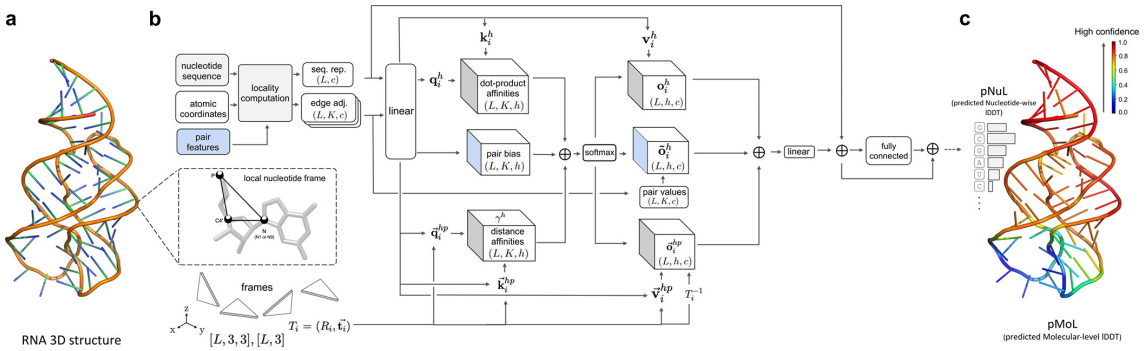


FIGURE 4.3: lociPARSE architecture [25].

The advancement of large language models (LLMs) raised interest in implementing transformers in other research areas. One of them was bioinformatics, as Google's DeepMind announced a model capable of predicting 3D structure of proteins, known as AlphaFold 2 [60]. The core of the model is invariant point attention, represented as transformer adaptation for predicting the next position of an element in a sequence, based on its previous position. The researchers from the lociPARSE [25] team got an idea to implement this approach in the 3D RNA world. Unlike the original AlphaFold2 implementation, lociPARSE introduces locality-aware geometry and edge-biased attention to convert nucleotide pair features to edge adjacencies by considering a set of K-nearest neighbour nucleotides (k=20) to capture the local atomic environment of each nucleotide based on Euclidean distances between atoms between nucleotide pairs. The model utilises coordinates from the point cloud, specifically focusing on three key atoms for nucleotide frame construction. Local nucleotide frames are defined from the Cartesian coordinates of P, C4', and glycosidic N atoms. They took this element from the AlphaFold system and trained it on their created dataset of 3D RNA data. Unlike existing machine learning methods that estimate superposition-based root-mean-square deviation (RMSD), lociPARSE estimates Local Distance Difference Test (lDDT) [29] scores capturing the accuracy of each nucleotide and its surrounding local atomic environment in a superposition-free manner, before aggregating information to predict global structural accuracy. The lDDT metric is normalised, where 1 represents an ideal structure and 0 represents the opposite. The lDDT offers several advantages over RMSD, being superposition-free. The results demonstrated that lociPARSE significantly outperforms existing statistical potentials (rsRNASP [61], cgRNASP [62], DFIRE-RNA [63], and RASP [64]) and machine learning methods (ARES and RNA3DCNN). However, the result appeared promising, but it requires much more computational power to not only train but also to run the system, as evidenced by the need for training on an 80-GB NVIDIA A100 GPU for 50 epochs.

4.2 RNAQuANet Architecture

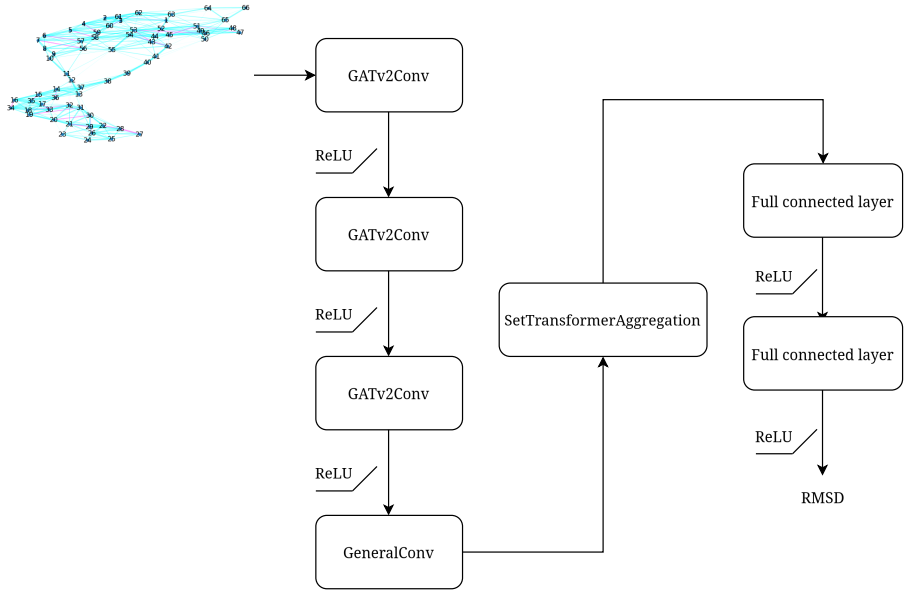


Figure 4.4: RNAQuANet architecture.

RNAQuANet architecture represents a hybrid of the models mentioned above, combining solutions from both the transformer-based architecture presented in the lociPARSE and the convolutional graph neural network demonstrated in the ARES. The architecture is split into two complementary paths.

The architecture was built using the PyTorch Geometric library [65], which provides a significant part of the well-known graph neural network layers.

The first path utilises Graph Convolutional Network (GCN) layers, which are the most common approach to model chemical compounds. In terms of the 3D RNA general quality assessment, it is the most promising use case. The general concept uses message passing for exchanging data between units, both nodes and edges. The result of the process modifies each entity in the graph: nodes and edges. This part begins with batch normalisation to achieve more regular training. The Figure 4.5 presents the concept of message passing. Each node

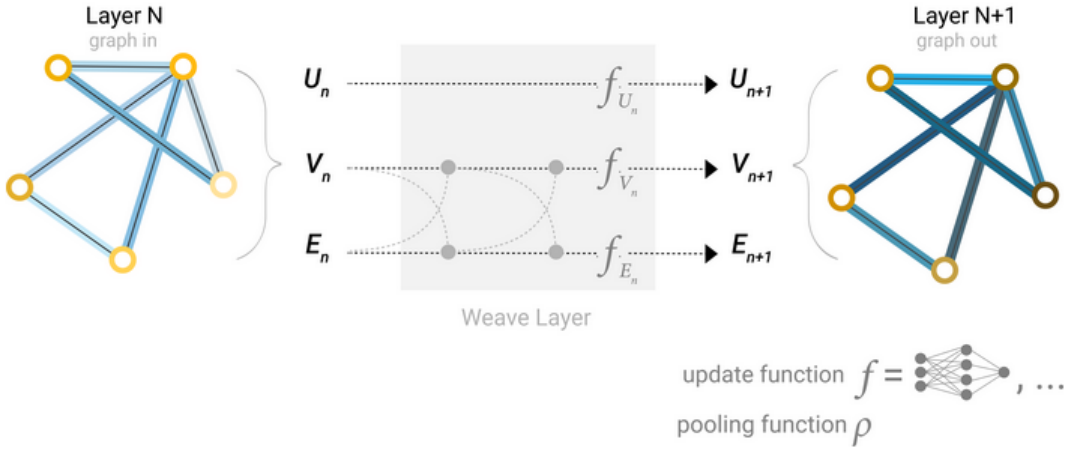


Figure 4.5: The concept of GNN [66].

and edge aggregates its neighbourhood, and later, using a specific trained function, applies

changes to itself.

The foundational component of this path consists of the first three layers implemented as GATv2Conv [67] modules, which represents an improved version of the original Graph Attention Network (GAT) that systematically addresses key limitations inherent in the original attention mechanism design. The attention mechanism is the key element of these layers, introducing dynamic relationship modelling capabilities that fundamentally transform how the network processes graph-structured data. Unlike traditional graph neural networks that apply uniform aggregation schemes, the GATv2Conv adaptation enables the model to distinguish the relative importance and value of all nodes within the graph structure and selectively treat neighbouring nucleotides based on their computed relevance scores. The selective attention mechanism enables the network to adaptively focus on the most significant neighbouring nucleotides and suppress less relevant connections. Furthermore, the enhanced attention computation in GATv2Conv enables better capture of long-range interactions between distant nucleotides in the 3D RNA structure, which is crucial for understanding global structural patterns and conformational stability that may not be apparent through local neighbourhood analysis alone. The layer takes three elements on its input: **node features**, in the form of a vector for each node; **edges**, in the form of pairs of nodes; **edge features**, in the form of a vector for each nodes pair in distance of 16 Å from C5' atoms. The last layer is a regular convolution layer without attention, which processes only node features and spatial data. The modification of the previous layer was motivated by the need to increase focus on edge features; this change compressed the computational effort into a smaller calculation space.

The middle part of architecture brings aggregation, which is crucial in Graph Neural Network-based architectures. The primary purpose is to flatten the graph space information into a single feature vector. Due to the main advancement of GNN, which is the ability to process graphs of various sizes, creating a stable function for aggregation is not a trivial task. The well-known examples, like mean, max/min (which find representatives), and sum (which learn structural properties) [68], do not perform very well for more complex calculations. Therefore, in recent years, new aggregation functions have emerged. The RNAQuANet architecture was tested using many examples of aggregation functions. The best results were achieved with the use of SetTransformerAggregation [69]. The aggregation architecture is specifically designed for learning on sets while maintaining permutation invariance. In simple terms, it works by treating graph nodes as a set of items and utilising a modified transformer architecture to aggregate their features into a single, unified graph representation. Unlike traditional attention mechanisms used in sequence modelling, Set-Transformer removes positional encoding since the order of nodes in a graph shouldn't matter for the final prediction. The architecture consists of two main components: an encoder that processes the input node features through multiple attention blocks to capture dependencies between different nodes, and a decoder that uses a special "seed vector" within attention blocks to create an initial readout vector, which is then further processed through self-attention modules and feedforward layers to produce the final graph-level representation. This approach is efficient because it can exploit dependencies between nodes when learning permutation-invariant representations, rather than simply embedding each node independently and then combining them as simpler methods do. The Set-Transformer essentially learns which nodes to pay attention to and how to incorporate their information optimally for the specific task, making it more adaptive and expressive than basic pooling functions like sum, mean, or max, while still ensuring that the same graph will produce the exact representation regardless of how its nodes are ordered.

The final path consists of two simple, fully connected layers. The primary purpose is to

expand the solution's architecture capacity; the input to these fully connected layers remains constant.

The main advantage of the RNAQuANet architecture is its size, which comprises approximately 647,000 parameters, significantly lower than those of other state-of-the-art methods. It is possible to train the model using a consumer GPU in under 1 hour. The interference takes seconds and does not require a specific GPU for computation.

Listing 4.1: PyTorch log of architecture layers parameters number.

```
   | Name          | Type                          | Params |
-------------------------------------------------------------
0  | norm1         | BatchNorm                     | 158    |
1  | GATconv1      | GATv2Conv                     | 62.3 K |
2  | GATconv2      | GATv2Conv                     | 212 K  |
3  | GATconv3      | GATv2Conv                     | 212 K  |
4  | GCN2Conv1     | GeneralConv                   | 52.9 K |
5  | fc1           | Linear                        | 6.3 K  |
6  | fc2           | Linear                        | 80     |
7  | dropout       | Dropout                       | 0      |
8  | activation    | ReLU                          | 0      |
9  | aggregation   | SetTransformerAggregation     | 102 K  |
-------------------------------------------------------------
647 K     Trainable params
```

Table 4.1: RNAQuANet Layer Architecture.

Layer Name	**Node Input Size**	**Node Output Size**	**Edge Features**
BatchNorm	79	79	-
GATv2Conv1 (heads=4)	79	316	35
ReLU + Dropout	316	316	-
GATv2Conv2 (heads=4)	316	316	35
ReLU + Dropout	316	316	-
GATv2Conv3 (heads=4)	316	316	35
ReLU + Dropout	316	316	-
GeneralConv	316	79	35
ReLU + Dropout	79	79	-
SetTransformerAggregation	79	79	-
Linear (fc1)	79	79	-
ReLU + Dropout	79	79	-
Linear (fc2)	79	1	-
Final ReLU	1	1	-

Chapter 5

Learning process description

The training is the crucial part of machine learning model development. To achieve plausible results, it is necessary to balance the training set to avoid possible skewness towards the most populous representatives. In terms of 3D RNA structures, it is particularly challenging due to the size of the training set. Most experimentally determined 3D RNA structures belong to two RNA families: tRNA and rRNA are usually redundant.

The learning process was divided into two methods: classical training and contrastive learning, and utilised a selected dataset comprising lociPARSE-based, RNAQuANet-based, and ARES-based approaches.

5.1 Classical training

The classical training approach employed in this study follows a conventional supervised learning methodology where each sample/3D RNA model from the dataset is utilised exactly once during each training epoch, with each input paired with its corresponding target value. To mitigate the effects of gradient accumulation that can occur when similar data points are processed sequentially, a common artefact arising from the systematic nature of dataset construction, the training data undergoes random shuffling before each epoch. The dataset utilised in this training process consists of clusters of structural decoys, which are computationally generated three-dimensional RNA structure models predicted for every reference structure considered. These decoys serve as negative examples that challenge the model to distinguish between accurate and inaccurate structural predictions, thereby enhancing the robustness of the learned representations. A critical component of the training methodology is centered on the strategic use of the RMSD score. The implementation of this metric required careful consideration of meaningful threshold boundaries to ensure effective learning. While the distinction between structures with RMSD values of 2 Å and 5 Å represents a significant difference in terms of practical utility, the quality of structures with RMSD values of 30 Å and 40 Å is equally unsuitable for meaningful biological interpretation. Through RNA-Puzzles challenges and analysis, combined with domain expertise, an optimal RMSD threshold of 15 Å was established as the boundary for training significance. This threshold creates a densely continuous filled training space where differences between models reinforce small distances, ensuring that the model learns proper distinctions between structures of varying quality. This approach maximises the learning potential within the biologically relevant range while avoiding the computational overhead associated with learning distinctions between uniformly poor-quality structures, which can be classified simply as unusable. The

model training was limited by the early stopping method, which analyses 10 epochs backwards and decides to stop training if there is no improvement in Mean Absolute Error results during that period. It was dictated by better generalisation and the exclusion of outliers, which were introducing bias to the final result.

5.1.1 Training on dataset proposed for RNAQuANet

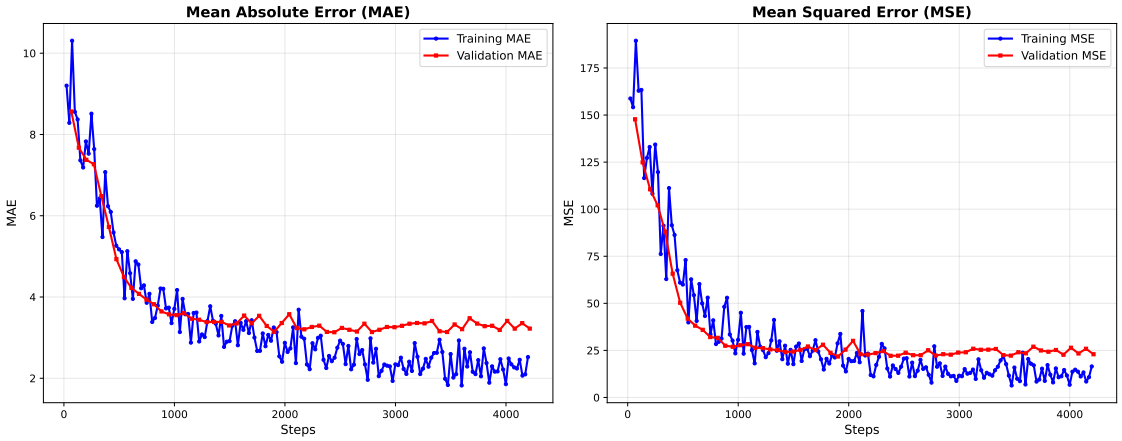


Figure 5.1: Performance of RNAQuANet-based model training.

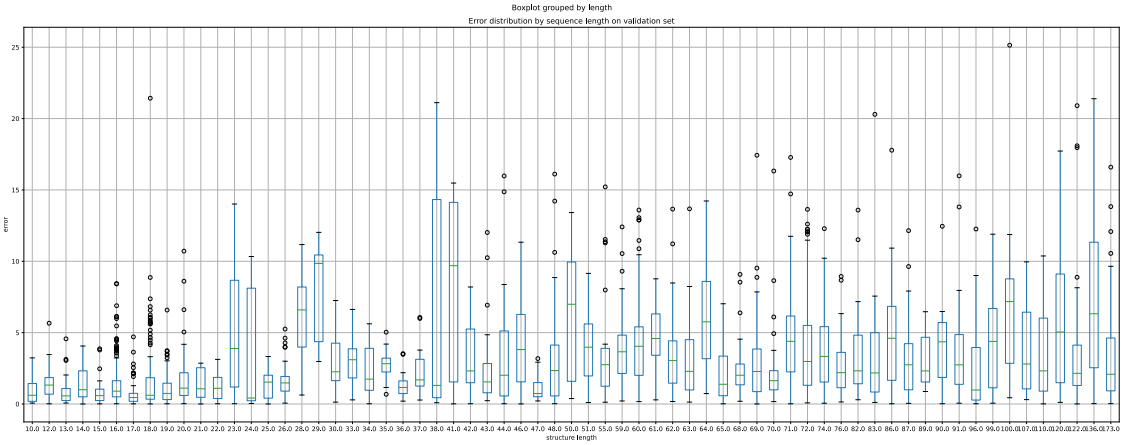


Figure 5.2: RNAQuANet-based validation prediction error distribution.

The Figure 5.1 presents the results of model training using the RNAQuANet dataset. The training loss is clearly noisy but steadily decreasing over time. The overfitting is barely visible, but a global optimum was achieved very quickly. The best validation metrics were a Mean Absolute Error of 3.4 Å and a mean squared error of 24 Å.

The Figure 5.2 presents a boxplot for the prediction error by each structure size in the validation set. The results are very promising, despite some outliers, such as in 23, 38, and 41 nts. The mean value of errors is placed below 5 Å and does not grow significantly with the increase in the number of nucleotides.

5.1.2 Training on dataset proposed for ARES

The Figure 5.3 presents the results of model training using the ARES dataset. In contrast to RNAQuANet-based training, the trend is not noisy and decreasing in training for both MAE

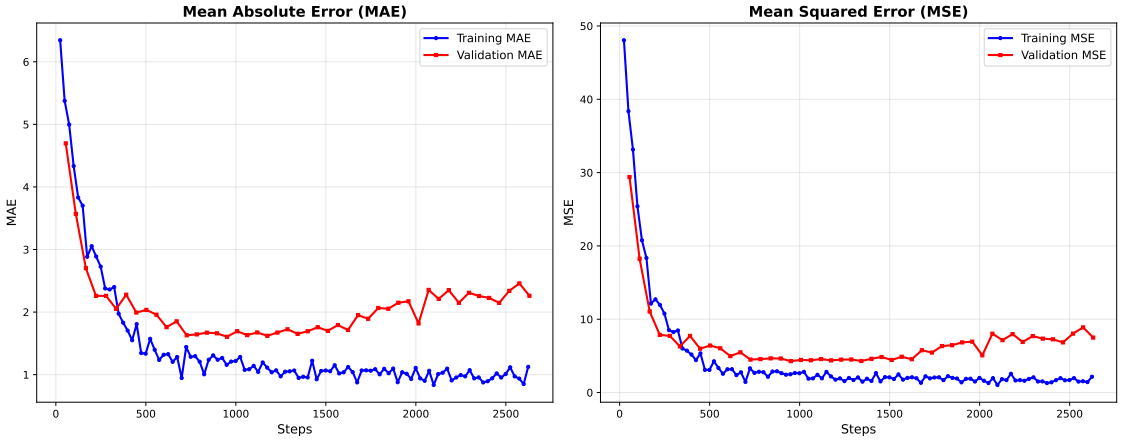


FIGURE 5.3: Performance of ARES-based model training.

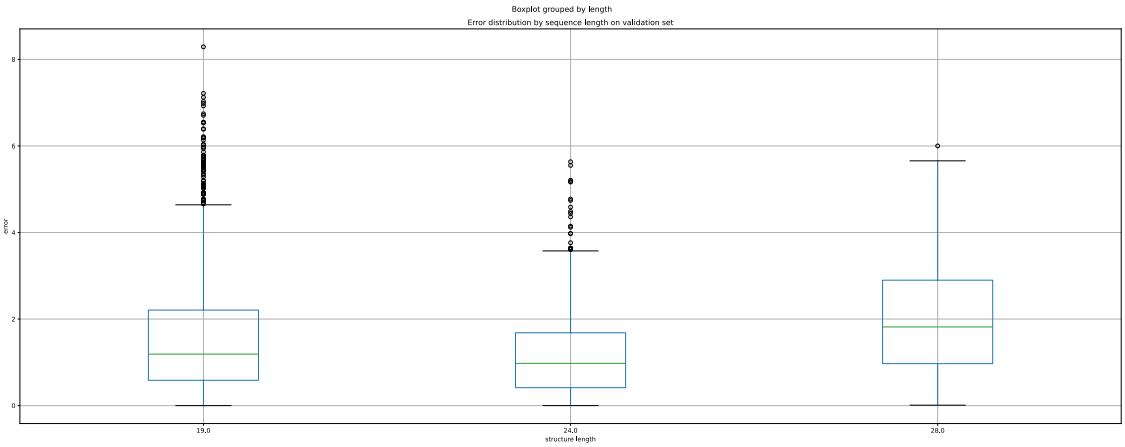


FIGURE 5.4: ARES-based validation set prediction error distribution.

and MSE. The validation trend shows very quick and significant overfitting. It was caused by an undifferentiated training set. The best metrics for validation are: MAE 1.6 and MSE 4.28.

The Figure 5.4 Figure [reference] shows a boxplot displaying prediction errors for different structure sizes in the validation dataset. The results remain consistent across structure sizes, with the exception of 19-nucleotide structures, which exhibit a notably high number of outlying values.

5.1.3 Training on dataset proposed for lociPARSE

The Figure 5.5 presents the results of model training using the lociPARSE dataset. It is characterised by significant noise in the training loss. In contrast to validation, which presents as a smooth, slowly decreasing line. It was the longest-lasting training, which took approximately 30 minutes. The best metrics for validation are: MAE 1.24 and MSE 4.57.

The Figure 5.6 presents a boxplot for model perfomance by each structure size in the validation set. Despite the noisy beginning of the plot, which is caused by the cardinality of the given groups, the rest of the part, despite minor outliers, looks satisfactory. This training was chosen later as an indicator for later improvement.

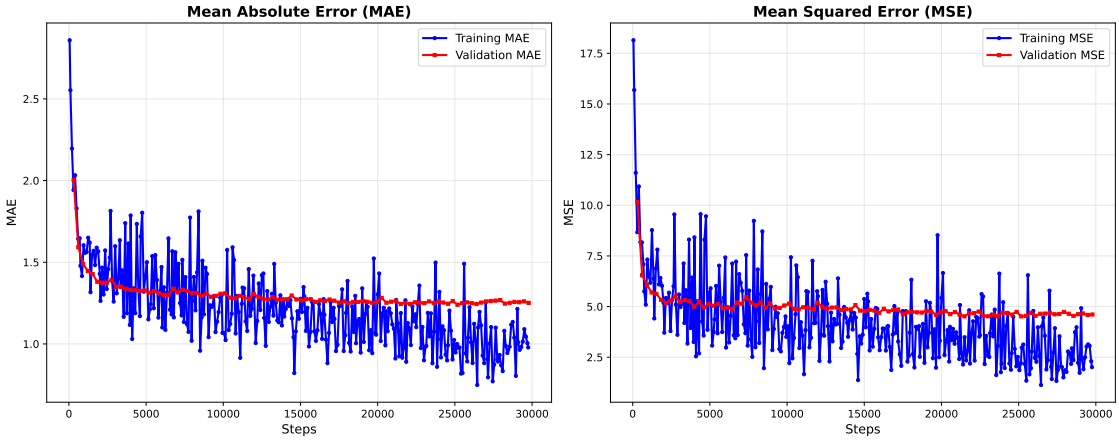


FIGURE 5.5: Performance of lociPASRSE-based model training.

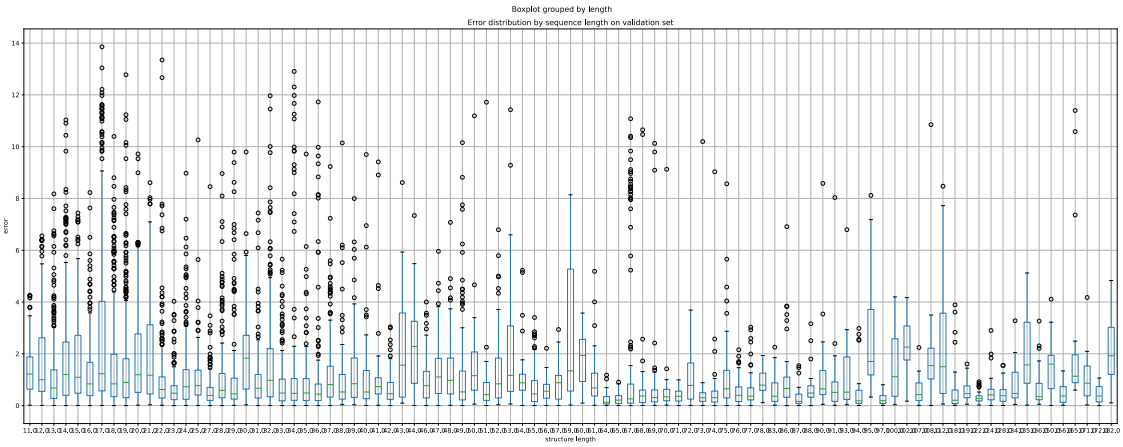


FIGURE 5.6: lociPASRSE-based validation set prediction error distribution.

5.2 Contrastive learning

The classical learning assumes the model's best effort to pay attention to details that can be later generalised for other 3D RNA structures. Even a small dataset imbalance can cause bias and overshadow less numerous examples. Another problem that arose during evaluation was the model's inability to distinguish small changes. To understand the grounds, it is necessary to examine the loss function more closely. It is clearly visible that the model during training is penalised with the value of the distance between the expected result and the prediction. One of the evaluation approaches assesses the quality of 3D RNA structure models by ranking them from the best to the worst structure, based on the RMSD value. The potential model needs to be able to predict values that will correspond to the proper order of expected results. The classical training focuses on a single value result, assuming that the context will emerge during the training process. That can be impossible with the few distinctive structures available. It was necessary to improve the loss function to put attention on the differences between models from the given model group.

There was no such loss function proposition in the literature. The first attempt was made to introduce Spearman's correlation to the MAE loss function. Due to the non-differentiable nature of the loss and the sophisticated method of calculation, the training process is characterised as insensitive. The Figure 5.7 presents the value of Spearman correlation across the training process. The value barely ever exceeds 0.3, which can be classified as noise.

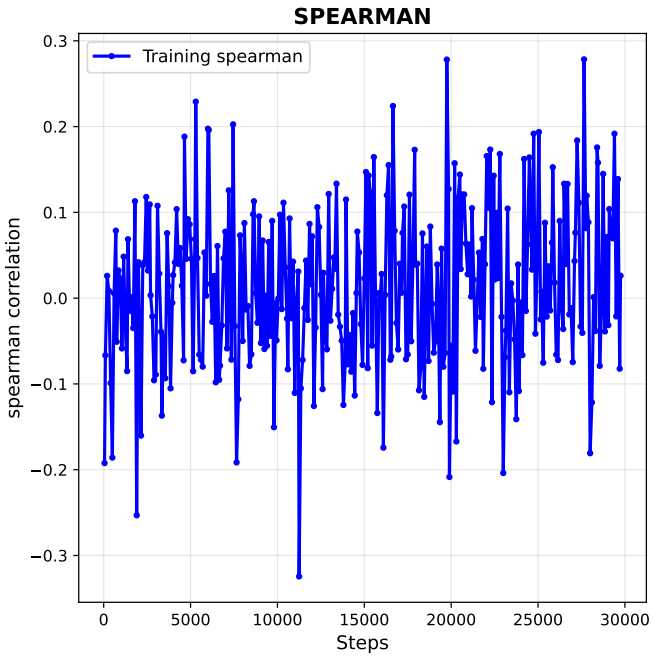


Figure 5.7: Spearman correlation in lociPARSE training.

5.2.1 Simple contrastive learning

The approach is based on a simple assumption. In each 3D RNA model group, the trainer takes each possible pair of entities, performs training steps and compares the results. If the prediction value depicts that a better 3D RNA model has a lower RMSD value, then the loss is simply the average of the MAE metric from both structures. If not, then add 1 to the calculated MAE value. This penalty mechanism effectively guides the model toward learning the correct structural quality hierarchy by increasing the loss when predictions contradict the RMSD-based ranking system. The pairwise comparison strategy ensures that the model encounters diverse structural variations within each training iteration, exposing it to subtle conformational differences that might otherwise be overlooked in standard training protocols. Additionally, by operating on pairs rather than individual structures, the method inherently captures relative structural quality assessments, which are often more reliable than absolute quality predictions in the context of 3D RNA structure evaluation. The approach may seem trivial, but it significantly improves the model's understanding of the contrast between 3D RNA models.

5.2.2 lociPARSE training

The Figure 5.8 presents the results of model training using the lociPARSE dataset. It is visible that after the first iteration, a global optimum loss was achieved. The green spikes depict the frequency of the swapped order of the two models. Overfitting is less significant in this example than in the RNAQuANet example, but the MAE/MSE values are lower than those obtained from the classical approach. The primary difference from RNAQuANet's contrastive training is the steady MAE/MSE validation value.

The Figure 5.9 displays a boxplot showing prediction errors across different structure sizes in the validation dataset. The results are nearly identical to those obtained using the classical method, suggesting significant potential for improved ranking in the final assessment.

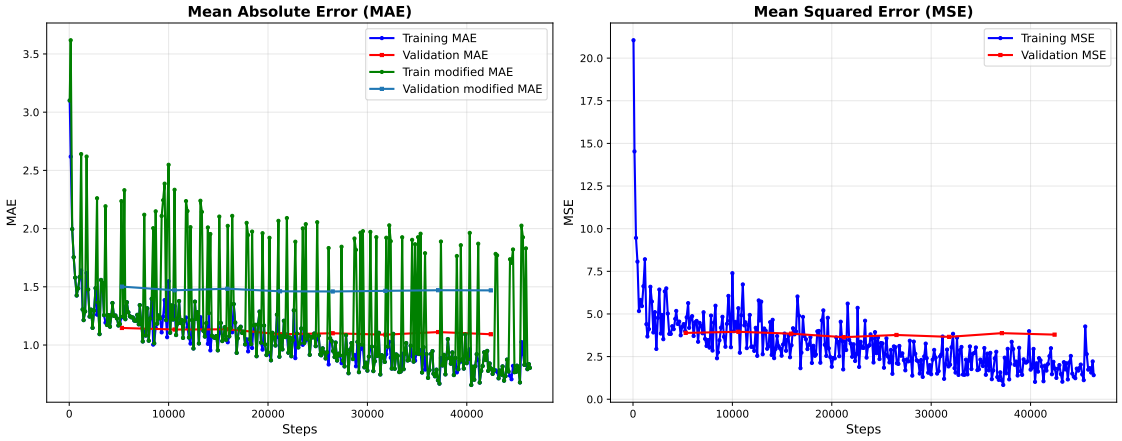


FIGURE 5.8: Performance of lociPASRSE-based model training.

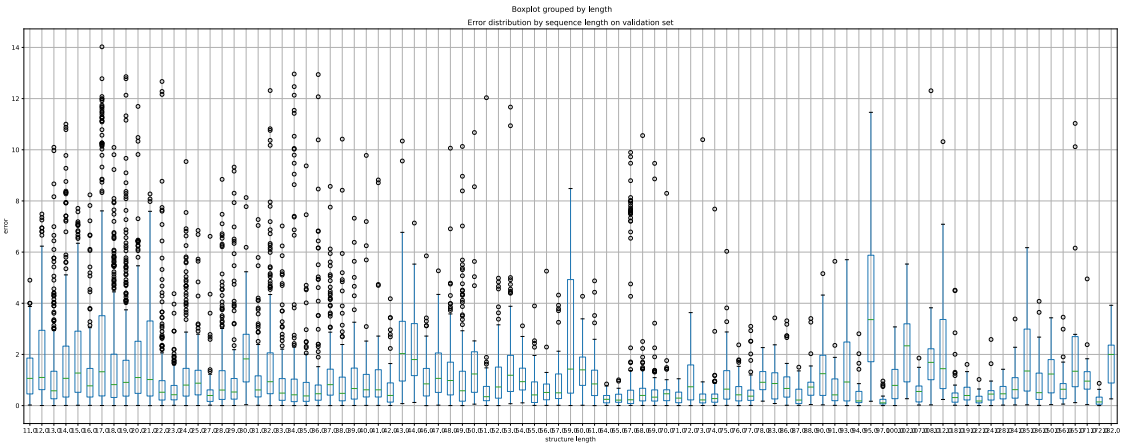


FIGURE 5.9: lociPASRSE-based validation prediction error distribution.

5.2.3 ARES training

This training was omitted due to the lack of a viable approach for such a group cardinality. The classes of similar structures consist of 1000 decoys; therefore, the method is not applicable here.

5.2.4 RNAQuANet training

The Figure 5.10 presents the results of model training using the RNAQuANet dataset. It is visible that after the first iteration, a global optimum loss was achieved. The green spikes depict the frequency of the swapped order of the two models. Overfitting is significant, but the MAE/MSE values are lower than those from the classical approach.

The Figure 5.11 presents prediction errors for various structure sizes in the validation data using a boxplot format. The results demonstrate higher levels of noise, most notably within the central portion of the structure length distribution.

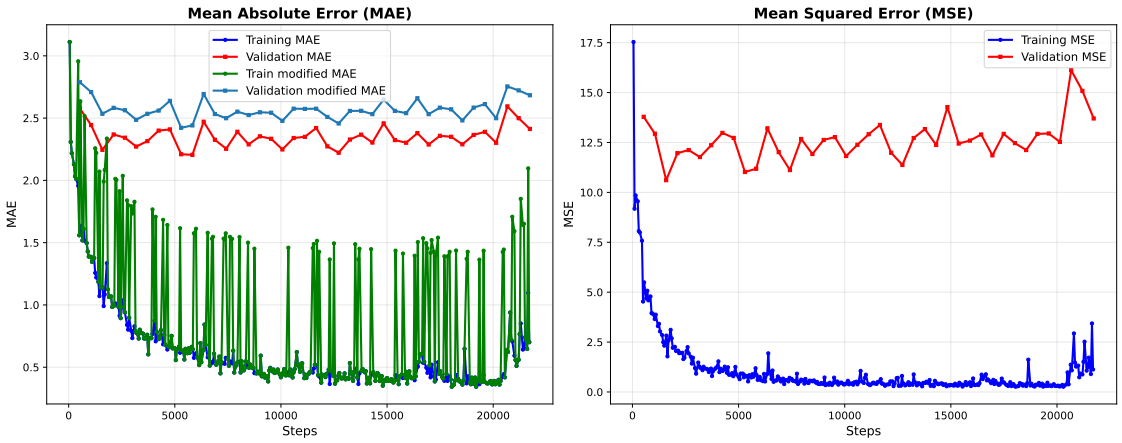


FIGURE 5.10: Performance of RNAQuANet-based model training.

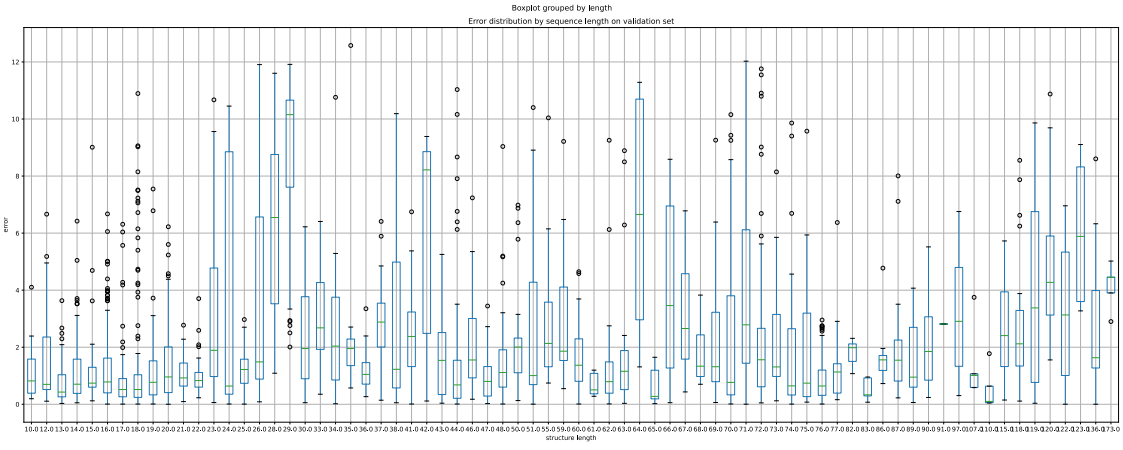


FIGURE 5.11: RNAQuANet-based validation prediction error distribution.

Chapter 6

Evaluation of the proposed models

The evaluation was conducted based on the published challenges considered in the five rounds of RNA-Puzzles competition. The contest repository contains models submitted by the community for challenges that have been considered. The most time-consuming part was connected with preprocessing 3D structures. Most of the 3D predictions were incomplete; therefore, we needed to refine them and reject any disputes that could not be resolved.

The evaluation was conducted on three model families: RNAQuANet, lociPARSE, and ARES. The last two were additionally retrained using the final lociPARSE dataset to ensure clean training without the influence of the evaluation dataset.

6.1 Overall models performance

The Figure 6.1 presents particular ranking places in the Spearman Correlation coefficient-based ranking for each RNA-Puzzles challenges. The first place, which is the most important measure, is nearly evenly split between all of the models, with a slight favour to lociPARSE. ARES is the best model in 7 competitions, lociPARSE in 9 and RNAQuANet in 6. Second place is more in favour of ARES and lociPARSE, while third place sees about half of the results coming from RNAQuANet. The Figure 6.2 presents each model's ranking position across all evaluations of the RNA-Puzzles challenges (the lower value the better). It is clearly visible that ARES was trained on lociPARSE, and lociPARSE performs better in terms of single-model prediction. The RNAQuANet model, which was trained using a contrastive approach, presents the best opportunity for future improvement.

6.2 Best opportunity for RNAQuANet

This section focuses on three selected RNA-Puzzles challenges where the RNAQuANet model presented sufficient results. These editions highlight RNAQuANet's most significant capabilities.

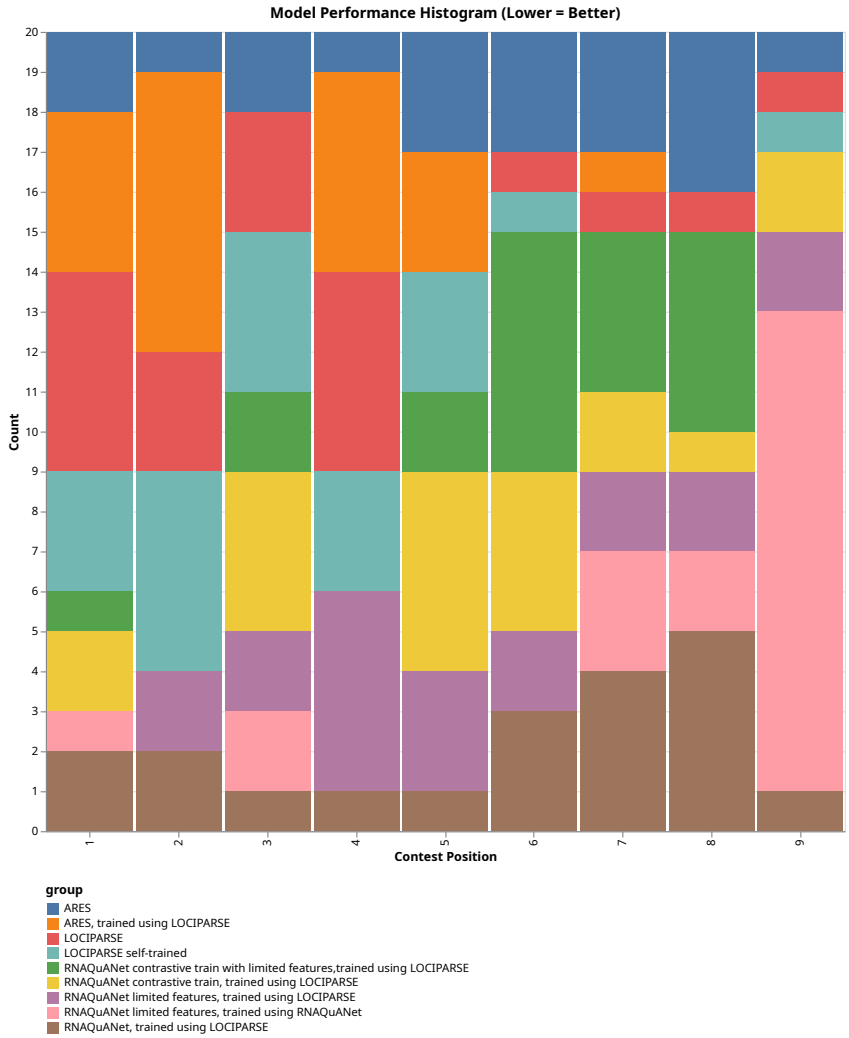


Figure 6.1: Overall position of models in ranking for each of the 22 RNA-Puzzles challenges.

6.2.1 RNA-Puzzles 18th challenge

Thirteen teams participated in the particular challenge, submitting 62 predicted models. The structure length is 71 nucleotides. The RMSD mean of all structures is 14.52 Å, and the standard deviation is 6.50.

The Figure 6.3b presents the range of models submitted by contestants. The set was quite rich with various RMSD values.

The table shows that ARES usually outputs better MAE/MSE results; however, RNAQuANet, which was trained in a contrastive manner, also produces a reasonable ranking.

The Figure 6.4 shows the expected/achieved ranking of 3D RNA models. It is clearly visible that a significant correlation exists in real-life use cases, which could be sufficient to distinguish between good and bad examples.

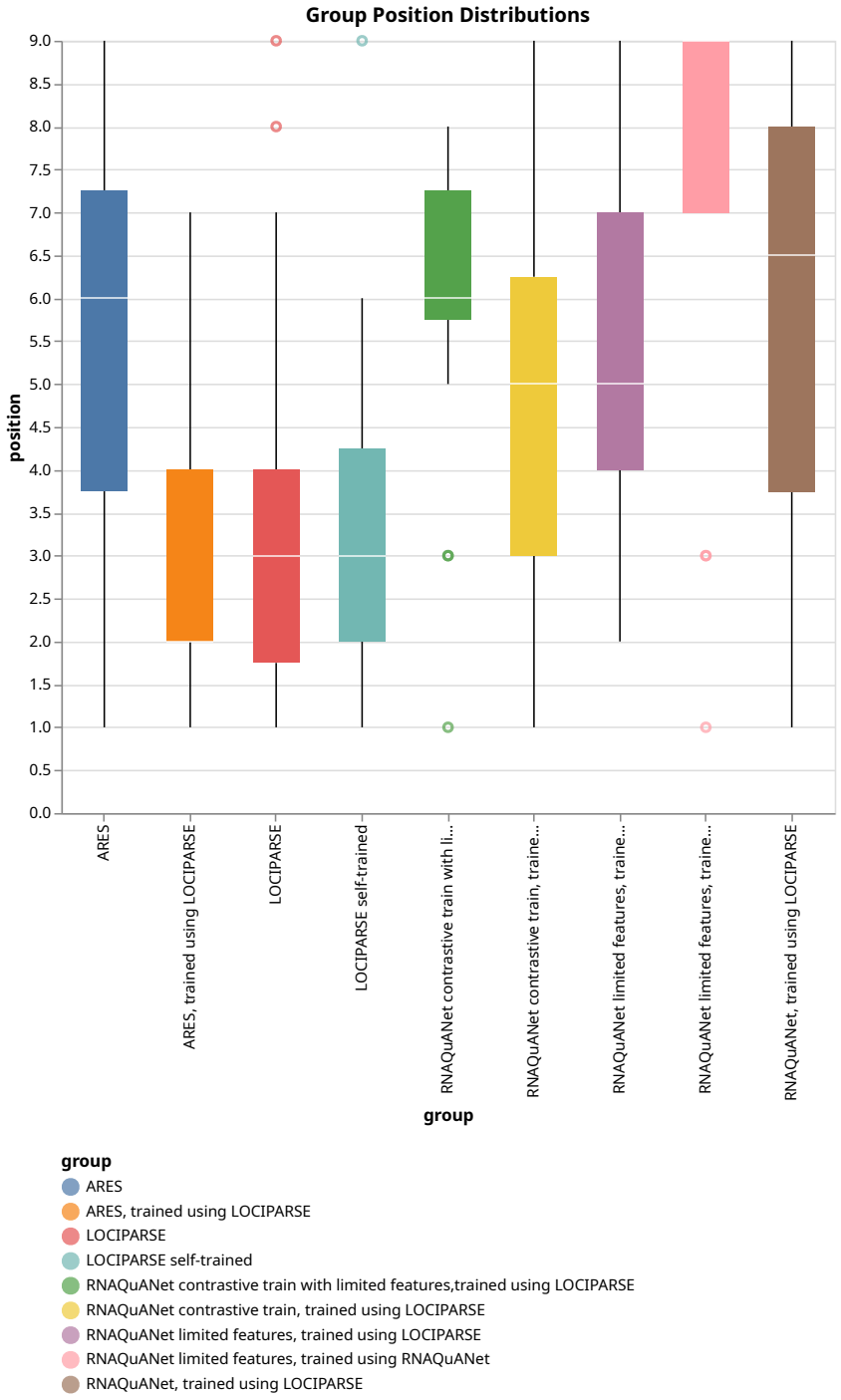
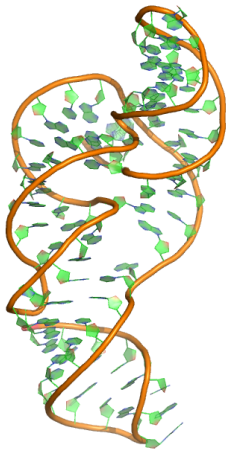


FIGURE 6.2: Overall position of models in ranking for each RNA-Puzzles challenges.

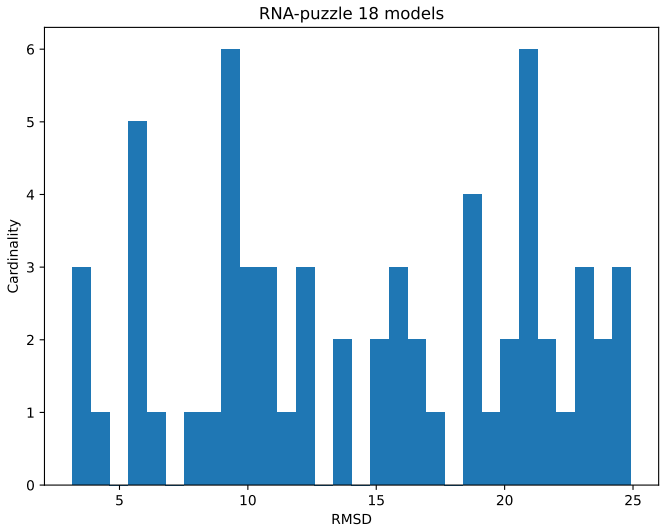
6.2.2 RNA-Puzzles 25th challenge

Nine teams participated in the particular challenge, submitting parsable 47 predicted models. The structure length is 69 nucleotides. The RMSD mean of all structures is 5.75 Å, and the standard deviation is 3.85. This dataset is narrower than the previous one. Definitely, it is one of the most difficult sets due to the high number of structures from the narrow, native-like models.

In the 25th challenge, RNAQuANet was not the best model, but the quality was not far from lociPARSE. The Spearman correlation coefficient of the best-trained RNAQuANet model



(A) Reference 3D structure, for challenge 18th, visualization prepared using PyMOL.



(B) Number of 3D predictions vs their RMSD scores distribution for 18th challenge.

TABLE 6.1: Developed models comparison for 18th challenge from RNA-Puzzles competition.

Model	**Spearman**	**MAE**	**MSE**
RNAQuANet contrastive train, trained using LOCIPARSE dataset	0.67	11.44	160.64
ARES, trained using LOCIPARSE dataset	0.44	9.21	116.02
RNAQuANet limited features, trained using RNAQuANet dataset	0.36	**6.81**	**65.21**
LOCIPARSE	0.36	-	-
LOCIPARSE self-trained	0.36	-	-
RNAQuANet contrastive train with limited features, trained using LOCIPARSE dataset	0.30	11.67	174.16
RNAQuANet limited features, trained using LOCIPARSE dataset	0.26	12.13	185.45
ARES	0.18	8.04	95.68
RNAQuANet, trained using LOCIPARSE dataset	0.14	11.95	182.65

ranked second overall.

The figure 6.6 shows the expected/achieved ranking of 3D RNA models. The correlation is sparse in the middle of the ranking, which can depict model problems with assessing quality in a 2-4 Å value range.

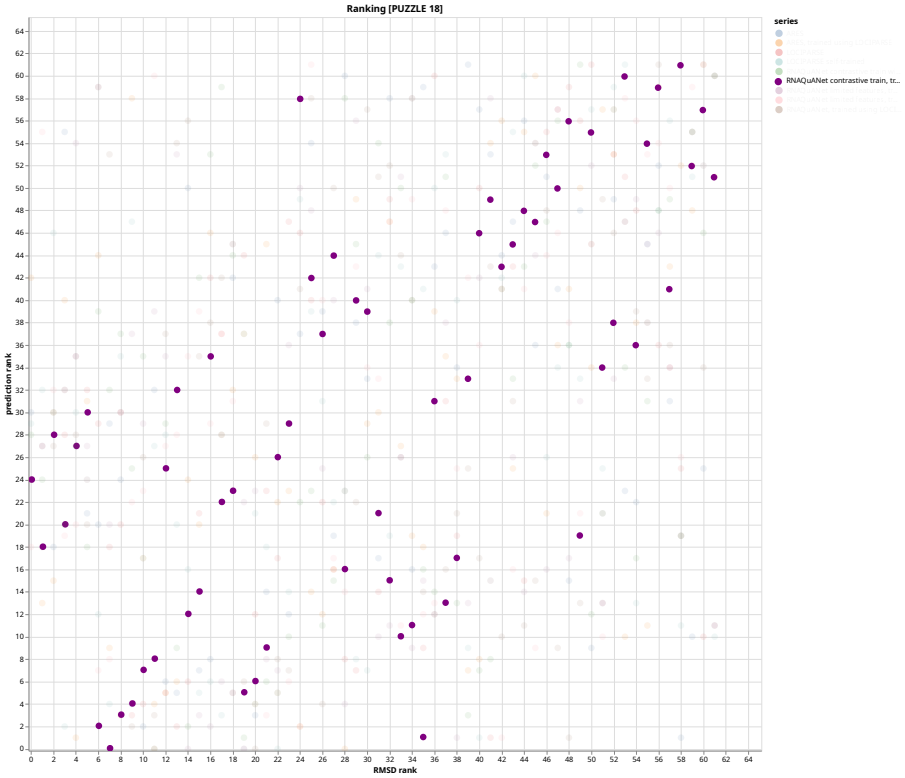
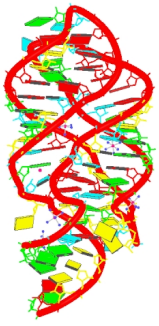
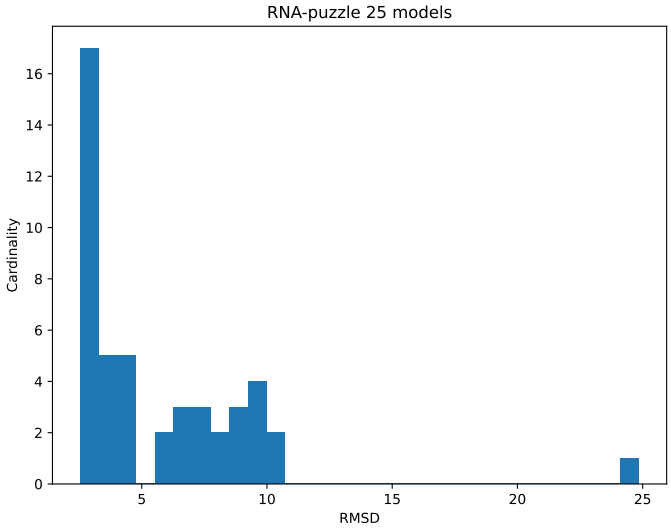


FIGURE 6.4: Constrastive RNAQuANet model vs ground truth for 18th challenge in RNA-Puzzles competition.



(A) Reference 3D structure, for challenge 25th, visualization prepared using PyMOL.



(B) Number of 3D predictions vs their RMSD scores distribution for 25th challenge.

6.2.3 RNA-Puzzles 32nd challenge

Fifteen teams participated in the particular challenge, submitting 106 predicted models. The structure length is 49 nucleotides. The RMSD mean of all structures is 15.23 Å, and the standard deviation is 6.68. The RMSD space is wide and evenly distributed. The accuracy of the RMSD prediction in terms of ordering significaly rise with assessing worse group of 3D RNA models 6.8.

Table 6.2: Developed models comparison for 25th challenge in RNA-Puzzles competition.

Model	Spearman	MAE	MSE
LOCIPARSE	0.71	-	-
RNAQuANet limited features, trained using LOCIPARSE dataset	0.64	3.98	27.67
LOCIPARSE self-trained	0.63	-	-
ARES, trained using LOCIPARSE dataset	0.62	**2.65**	**18.01**
RNAQuANet contrastive train, trained using LOCIPARSE dataset	0.42	3.65	26.32
RNAQuANet, trained using LOCIPARSE dataset	0.39	3.50	25.81
ARES	0.29	3.66	20.97
RNAQuANet contrastive train with limited features, trained using LOCIPARSE dataset	0.26	3.51	26.26
RNAQuANet limited features, trained using RNAQuANet dataset	-0.25	5.31	45.27

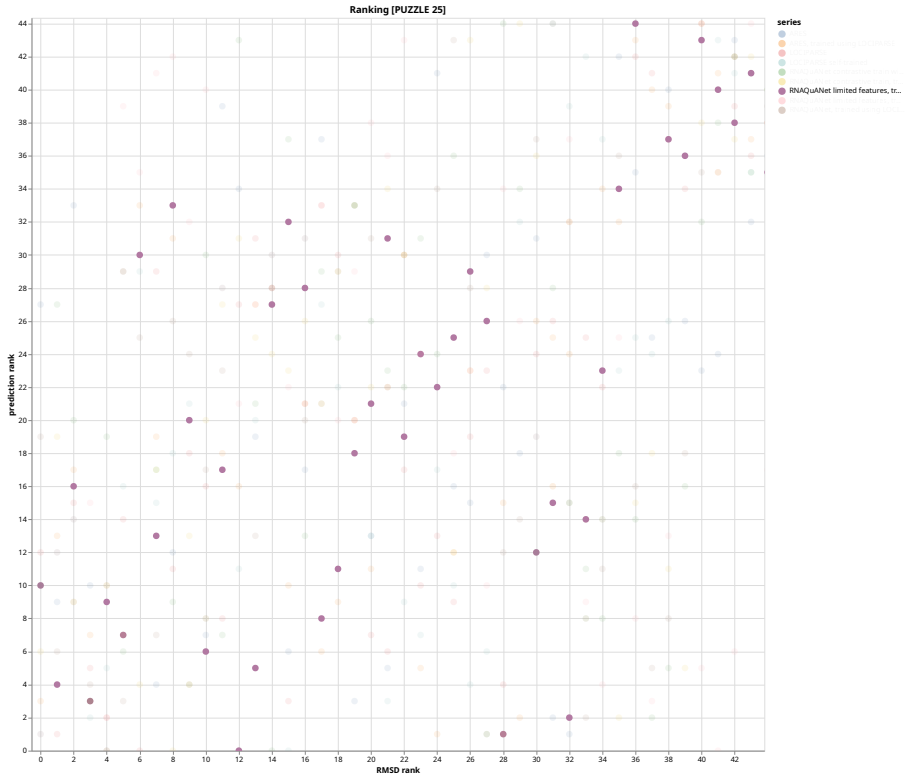


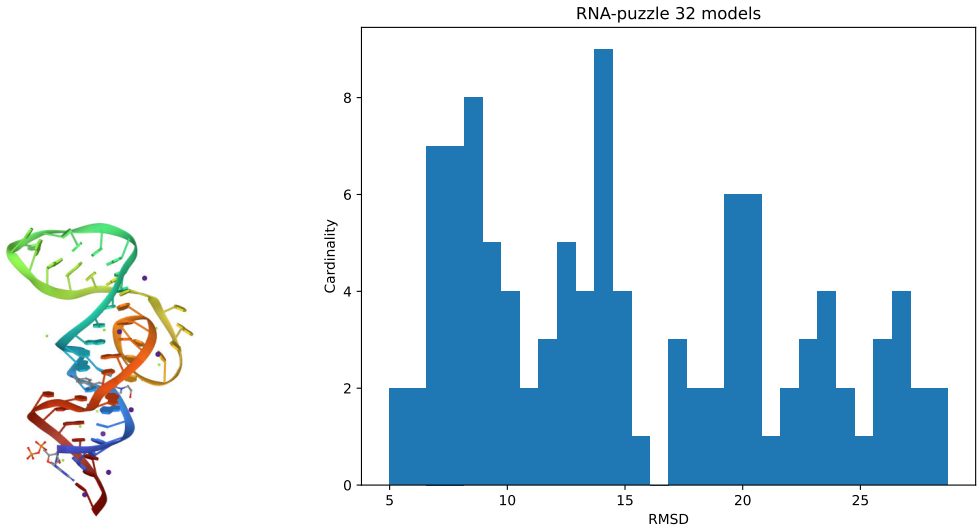
Figure 6.6: Limited features (common only) RNAQuANet model vs ground truth for 25th challenge in RNA-Puzzles competition.

6.3 Further improvement of RNAQuANet

This section focuses on contest editions in which RNAQuANet performs the worst. The 28th edition was selected as the most interesting edition due the RMSD ranking placement.

6.3.1 RNA-Puzzles 28th challenge

Eleven teams participated in the particular challenge, submitting 72 predicted models. The structure length was 77 nucleotides. The RMSD mean of all structures is 6.19 Å, and the standard deviation is 4.46. The RMSD space was concentrated in the range 0-5 Å, with some



(A) Reference 3D structure, for challenge 32nd, visualization prepared using PyMOL.

(B) Number of 3D predictions vs their RMSD scores distribution for 32nd challenge.

TABLE 6.3: Developed models comparison for 32nd challenge from RNA-Puzzles competition.

Model	**Spearman**	**MAE**	**MSE**
RNAQuANet contrastive train, trained using LOCIPARSE	0.51	11.10	154.78
RNAQuANet, trained using LOCIPARSE	0.50	11.32	162.25
RNAQuANet contrastive train with limited features, trained using LOCIPARSE	0.44	11.34	166.86
ARES, trained using LOCIPARSE	0.44	10.08	142.16
RNAQuANet limited features, trained using LOCIPARSE	0.17	11.85	179.91
LOCIPARSE	0.02	-	-
ARES	-0.09	8.38	113.34
RNAQuANet limited features, trained using RNAQuANet	-0.09	**7.84**	**86.89**
LOCIPARSE self-trained	-0.15	14.89	266.17

outliers beyond 10 Å.

Table 6.4 clearly shows that RNAQuANet was unable to handle this competition properly.

The problem can be explained by Figure 6.10, which clearly shows that the beginning of the ranking is anticorrelated. It is highly probable that models learn some 3D RNA motifs as invalid. The ranking was accurately predicted to end.

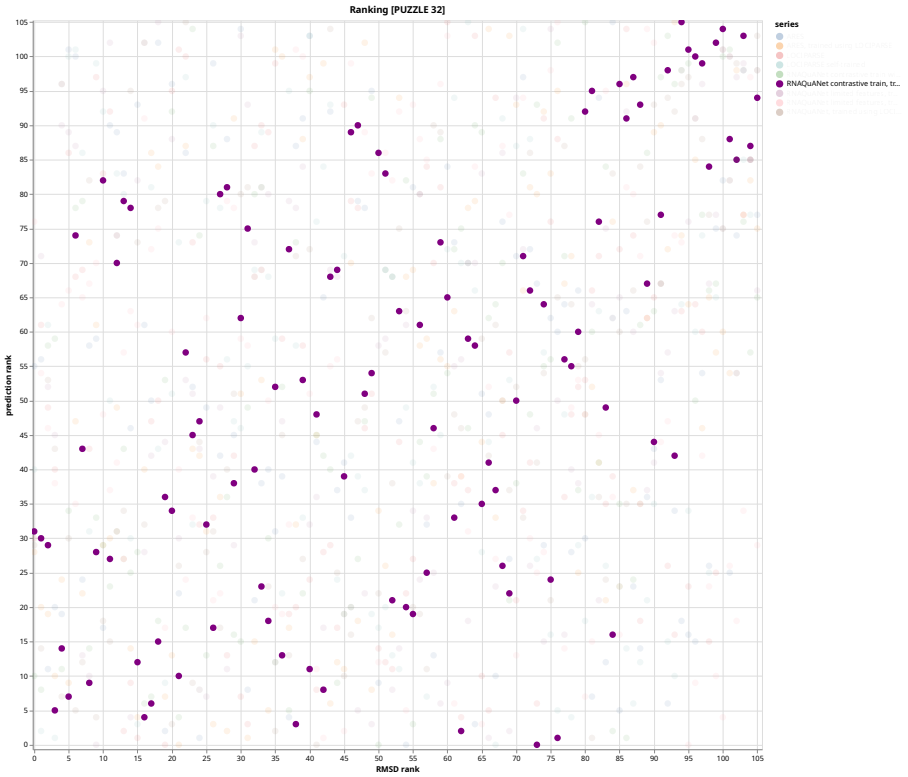
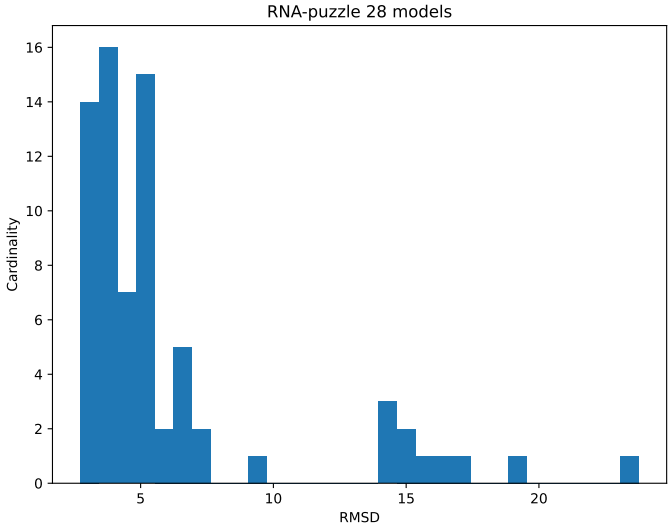


FIGURE 6.8: Constrastive RNAQuANet model vs ground truth for 32nd challenge in RNA-Puzzles competition.

(A) Reference 3D structure, for challenge 28th, visualization prepared using PyMOL.



(B) Number of 3D predictions vs their RMSD scores distribution for 28th challenge.

TABLE 6.4: Developed models comparison for 28th challenge in RNA-Puzzles competition.

Model	**Spearman**	**MAE**	**MSE**
LOCIPARSE	0.82	-	-
LOCIPARSE self-trained	0.78	-	-
ARES	0.75	3.84	**20.22**
ARES, trained using LOCIPARSE	0.33	**2.77**	21.30
RNAQuANet limited features, trained using LOCIPARSE	0.13	3.13	25.09
RNAQuANet contrastive train, trained using LOCIPARSE	0.07	2.97	24.31
RNAQuANet, trained using LOCIPARSE	-0.03	2.91	24.40
RNAQuANet contrastive train with limited features, trained using LOCIPARSE	-0.13	3.07	25.52
RNAQuANet limited features, trained using RNAQuANet	-0.13	6.24	48.78

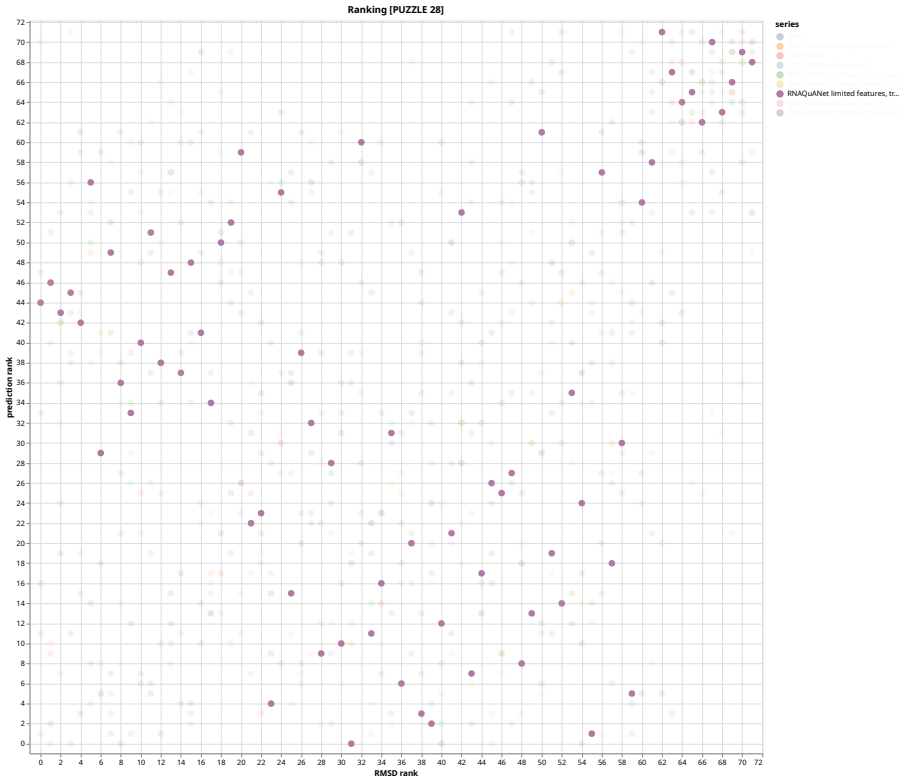


FIGURE 6.10: Limited features RNAQuANet model vs ground truth for 28th challenge in RNA-Puzzles competition.

Chapter 7

Summary

The proposed RNAQuANet architecture hybrid system, with only 647,000 parameters, is significantly more efficient than existing state-of-the-art methods. The model can be trained on consumer GPUs in under an hour and performs inference in seconds. The innovative contrastive learning approach penalises incorrect structural quality rankings, improving the model's ability to distinguish between structures with similar RMSD values. Here, a comprehensive feature extraction pipeline that processes 79 node features and 35 edge features from RNA structures, including bond distances, planar angles, torsion angles, and base-pairing characteristics are developed. To handle variable feature sets across different nucleotides, the system employs autoencoders that compress nucleotide-specific features while preserving structural information.

RNAQuANet demonstrated competitive performance, achieving the best Spearman correlation (0.67) for challenge 18 of RNA-Puzzles competition and showing particular strength in ranking structures with diverse RMSD values. However, the model struggled with certain structural motifs, as evidenced by its not entirely satisfactory performance, particularly in challange 28, suggesting potential overfitting to specific structural patterns.

The model development process faced limited availability of high-resolution RNA structures (only around 700 representative structures) and constrained training data diversity. The software infrastructure proved problematic, as non-robust tools like FR3D [70] created bottlenecks due to faulty in the feature extraction pipeline, while many source structures contained incomplete or invalid atomic coordinates that required extensive preprocessing. The heterogeneous quality of existing RNA structural analysis tools necessitated the development of custom solutions and workarounds.

Despite resource limitations and technical challenges, this work successfully demonstrates that graph neural networks can effectively assess RNA structure quality referenceless. The combination of architectural design, training strategies, and feature engineering provides a foundation for future developments in computational RNA structural biology. The project's success, achieved with limited computational resources and problematic software infrastructure, highlights both the potential of machine learning approaches in structural biology.

Chapter 8

Future work

In future work, it is essential to conduct comprehensive ablation studies to systematically determine which features are responsible for the final result and quantify their individual contributions to model performance. This analysis will likely lead to reducing the number of features, which was the case with unsatisfactory model performance, as redundant or noisy features often degrade predictive accuracy by introducing irrelevant variance into the learning process. A structured approach involving recursive feature elimination, SHAP analysis, and correlation studies would help identify the minimal feature set that maintains or improves performance while reducing computational overhead and potential overfitting.

A promising avenue would be to introduce higher-level motifs and structural patterns to the feature space, moving beyond simple nucleotide-based representations. Currently, the model operates only on nucleotide-based knowledge at the sequence level; however, it may require additional hierarchical support to fully comprehend the implications of torsion angle changes. Incorporating known regulatory elements, secondary structure motifs, binding domains, and conserved regions could provide the model with biologically meaningful context, connecting sequence information to functional outcomes. This multi-scale approach would enable the model to recognise patterns at different levels of biological organisation, from local base-pair interactions to global structural domains.

Another compelling direction would be to introduce physics-aware neural networks that not only operate on measurable experimental data but also are explicitly constrained by fundamental physical principles. These physics-informed architectures could incorporate differential equations governing molecular dynamics, empirical force fields describing atomic interactions, and statistical mechanical principles that govern conformational sampling. By embedding physical laws directly into the network architecture or loss function, the model would be prevented from making predictions that violate established scientific principles while potentially requiring less training data to achieve robust performance.

Finally, there is another pragmatic option to explore significantly larger neural network architectures to determine whether the current solution's memorisation and pattern recognition capabilities are fundamentally insufficient for the complexity of the underlying biological system. This scaling approach would involve investigating transformer-based architectures, graph neural networks capable of representing molecular topology, or ensemble methods that combine multiple specialised models.

Bibliography

[1] Rhiju Das, Rachael C. Kretsch, Adam J. Simpkin, Thomas Mulvaney, Phillip Pham, Ramya Rangan, Fan Bu, Ronan M. Keegan, Maya Topf, Daniel J. Rigden, Zhichao Miao, and Eric Westhof. Assessment of three-dimensional RNA structure prediction in CASP15. April 2023.

[2] Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Michał J. Boniecki, Janusz Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Yi Cheng, Fang-Chieh Chou, Rhiju Das, Nikolay V. Dokholyan, Feng Ding, Caleb Geniesse, Yangwei Jiang, Astha Joshi, Andrey Krokhotin, Marcin Magnus, Olivier Mailhot, Francois Major, Thomas H. Mann, Paweł Piątkowski, Radoslaw Pluta, Mariusz Popenda, Joanna Sarzynska, Lizhen Sun, Marta Szachniuk, Siqi Tian, Jian Wang, Jun Wang, Andrew M. Watkins, Jakub Wiedemann, Yi Xiao, Xiaojun Xu, Joseph D. Yesselman, Dong Zhang, Yi Zhang, Zhenzhen Zhang, Chenhan Zhao, Peinan Zhao, Yuanzhe Zhou, Tomasz Zok, Adriana Żyła, Aiming Ren, Robert T. Batey, Barbara L. Golden, Lin Huang, David M. Lilley, Yijin Liu, Dinshaw J. Patel, and Eric Westhof. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, 26(8):982–995, May 2020.

[3] Zhichao Miao, Ryszard W. Adamiak, Maciej Antczak, Robert T. Batey, Alexander J. Becka, Marcin Biesiada, Michał J. Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, Fang-Chieh Chou, Adrian R. Ferré-D'Amaré, Rhiju Das, Wayne K. Dawson, Feng Ding, Nikolay V. Dokholyan, Stanisław Dunin-Horkawicz, Caleb Geniesse, Kalli Kappel, Wipapat Kladwang, Andrey Krokhotin, Grzegorz E. Łach, François Major, Thomas H. Mann, Marcin Magnus, Katarzyna Pachulska-Wieczorek, Dinshaw J. Patel, Joseph A. Piccirilli, Mariusz Popenda, Katarzyna J. Purzycka, Aiming Ren, Greggory M. Rice, John Santalucia, Joanna Sarzynska, Marta Szachniuk, Arpit Tandon, Jeremiah J. Trausch, Siqi Tian, Jian Wang, Kevin M. Weeks, Benfeard Williams, Yi Xiao, Xiaojun Xu, Dong Zhang, Tomasz Zok, and Eric Westhof. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, 23(5):655–672, January 2017.

[4] Joana Pereira, Adam J. Simpkin, Marcus D. Hartmann, Daniel J. Rigden, Ronan M. Keegan, and Andrei N. Lupas. High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1687–1699, July 2021.

[5] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, September 1976.

[6] Dhananjay Bhattacharyya and Abhijit Mitra. Non-canonical base pairing. *WikiJournal of Science*, 6(1):X, 2023.

[7] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001.

[8] Russell S. Hamilton, Graeme Ball, and Ilan Davis. *A Multidisciplinary Approach to RNA Localisation*, page 213–233. Springer New York, 2012.

[9] Dr. Xiang-Jun Lu; x3dna.org. X3DNA-DSSR Homepage – Nucleic Acid Structures — x3dna.org. `https://x3dna.org/highlights/simple-base-pair-parameters`. [Accessed 09-08-2025].

[10] Prakash Ananth, Gunaseelan Goldsmith, and Narayanarao Yathindra. An innate twist between Crick's wobble and Watson-Crick base pairs. *RNA*, 19(8):1038–1053, July 2013.

[11] Marta Mackowiak, Bartosz Adamczyk, Marta Szachniuk, and Tomasz Zok. RNAtango: Analysing and comparing RNA 3D structures via torsional angles. *PLOS Computational Biology*, 20(10):e1012500, October 2024.

[12] Dr. Xiang-Jun Lu; x3dna.org. X3DNA-DSSR Homepage – Nucleic Acid Structures — x3dna.org. `https://x3dna.org/highlights/torsion-angles-of-nucleic-acid-structures`. [Accessed 09-08-2025].

[13] RNAsolo2 | Repository of cleaned PDB-derived RNA 3D structures — rnasolo.cs.put.poznan.pl. `https://rnasolo.cs.put.poznan.pl/`. [Accessed 09-08-2025].

[14] MS Smyth and JHJ Martin. x Ray crystallography. *Molecular Pathology*, 53(1):8, 2000.

[15] Frank A Bovey, Peter A Mirau, and HS Gutowsky. *Nuclear magnetic resonance spectroscopy*. Elsevier, 1988.

[16] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. How cryo-EM is revolutionizing structural biology. *Trends in biochemical sciences*, 40(1):49–57, 2015.

[17] Barbara Ann Kitchenham and Stuart Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 07 2007.

[18] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, 7(4):291–294, Apr 2010.

[19] Emidio Capriotti, Tomas Norambuena, Marc A. Marti-Renom, and Francisco Melo. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, 27(8):1086–1093, 02 2011.

[20] Julie Bernauer, Xuhui Huang, Adelene Y.L. Sim, and Michael Levitt. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, 17(6):1066–1075, April 2011.

[21] Jian Wang, Yunjie Zhao, Chunyan Zhu, and Yi Xiao. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Research*, 43(10):e63–e63, February 2015.

[22] Tongchuan Zhang, Guodong Hu, Yuedong Yang, Jihua Wang, and Yaoqi Zhou. All-Atom Knowledge-Based Potential for RNA Structure Discrimination Based on the Distance-Scaled Finite Ideal-Gas Reference State. *Journal of Computational Biology*, 27(6):856–867, June 2020.

[23] Jun Li, Wei Zhu, Jun Wang, Wenfei Li, Sheng Gong, Jian Zhang, and Wei Wang. RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLOS Computational Biology*, 14(11):e1006514, November 2018.

[24] Raphael J. L. Townshend, Stephan Eismann, Andrew M. Watkins, Ramya Rangan, Maria Karelina, Rhiju Das, and Ron O. Dror. Geometric deep learning of RNA structure. *Science*, 373(6558):1047–1051, August 2021.

[25] Sumit Tarafder and Debswapna Bhattacharya. lociPARSE: A Locality-aware Invariant Point Attention Model for Scoring RNA 3D Structures. *Journal of Chemical Information and Modeling*, 64(22):8655–8664, November 2024.

[26] Chengwei Deng, Yunxin Tang, Jian Zhang, Wenfei Li, Jun Wang, and Wei Wang. RNAGCN: RNA tertiary structure assessment with a graph convolutional network. *Chinese Physics B*, 31(11):118702, October 2022.

[27] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[28] Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1), November 2022.

[29] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

[30] Neocles B Leontis and Craig L Zirbel. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. *RNA 3D Structure Analysis and Prediction*, pages 281–298, 2012.

[31] Bartosz Adamczyk, Maciej Antczak, and Marta Szachniuk. RNAsolo: a repository of cleaned PDB-derived RNA 3D structures. *Bioinformatics*, 38(14):3668–3670, June 2022.

[32] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, Jun 2024.

[33] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pages 2025–06, 2025.

[34] Marcin Biesiada, Katarzyna J Purzycka, Marta Szachniuk, Jacek Blazewicz, and Ryszard W Adamiak. Automated RNA 3D structure prediction with RNAComposer. In *RNA Structure Determination: Methods and Protocols*, pages 199–215. Springer, 2016.

[35] Peter Eastman, Raimondas Galvelis, Raúl P Peláez, Charlles RA Abreu, Stephen E Farr, Emilio Gallicchio, Anton Gorenko, Michael M Henry, Frank Hu, Jing Huang, et al. OpenMM 8: molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*, 128(1):109–116, 2023.

[36] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The Amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[37] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, Houyang Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.

[38] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the american chemical society*, 118(45):11225–11236, 1996.

[39] Wenkai Wang, Chenjie Feng, Renmin Han, Ziyi Wang, Lisha Ye, Zongyang Du, Hong Wei, Fa Zhang, Zhenling Peng, and Jianyi Yang. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nature communications*, 14(1):7266, 2023.

[40] Andriy Kryshtafovych, Maciej Antczak, Marta Szachniuk, Tomasz Zok, Rachael C. Kretsch, Ramya Rangan, Phillip Pham, Rhiju Das, Xavier Robin, Gabriel Studer, Janani Durairaj, Jerome Eberhardt, Aaron Sweeney, Maya Topf, Torsten Schwede, Krzysztof Fidelis, and John Moult. New prediction categories in CASP15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1550–1557, June 2023.

[41] Serge Thill. DeepFoldRNA: A Graph Neural Network for RNA 3D Structure Prediction. *Bioinformatics and Code*, 1(1), 2025.

[42] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature methods*, 21(1):117–121, 2024.

[43] Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 21(12):2287–2298, 2024.

[44] Yang Li, Chengxin Zhang, Chenjie Feng, Robin Pearce, P Lydia Freddolino, and Yang Zhang. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nature Communications*, 14(1):5745, 2023.

[45] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, 2010.

[46] Firas Khatib, Seth Cooper, Michael D Tyka, Kefan Xu, Ilya Makedon, Zoran Popović, David Baker, and Foldit Players. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47):18949–18953, 2011.

[47] Andrew Martin Watkins, Ramya Rangan, and Rhiju Das. FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure*, 28(8):963–976, 2020.

[48] Ivo L Hofacker. Vienna RNA secondary structure server. *Nucleic acids research*, 31(13):3429–3431, 2003.

[49] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[50] Shay Zakov, Yoav Goldberg, Michael Elhadad, and Michal Ziv-Ukelson. Rich parameterization improves RNA structure prediction. *Journal of Computational Biology*, 18(11):1525–1542, 2011.

[51] Kengo Sato, Michiaki Hamada, Kiyoshi Asai, and Toutai Mituyama. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic acids research*, 37(suppl_2):W277–W280, 2009.

[52] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, 2011.

[53] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292, 2004.

[54] Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

[55] Jihong Ren, Baharak Rastegari, Anne Condon, and Holger H Hoos. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11(10):1494–1504, 2005.

[56] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, CA, 1967. University of California Press.

[57] Maciej Antczak, Tomasz Zok, Maciej Osowiecki, Mariusz Popenda, Ryszard W Adamiak, and Marta Szachniuk. RNAfitme: a webserver for modeling nucleobase and nucleoside residue conformation in fixed-backbone RNA structures. *BMC bioinformatics*, 19(1):304, 2018.

[58] Xiang-Jun Lu, Harmen J Bussemaker, and Wilma K Olson. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*, 43(21):e142, 2015.

[59] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2304, 1998.

[60] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021.

[61] Ya-Lan Tan, Xunxun Wang, Ya-Zhou Shi, Wenbing Zhang, and Zhi-Jie Tan. rsRNASP: A residue-separation-based statistical potential for RNA 3D structure evaluation. *Biophysical Journal*, 121(1):142–156, 2022.

[62] Ya-Lan Tan, Xunxun Wang, Shixiong Yu, Bengong Zhang, and Zhi-Jie Tan. cgRNASP: coarse-grained statistical potentials with residue separation for RNA structure evaluation. *NAR Genomics and Bioinformatics*, 5(1):lqad016, 2023.

[63] Tong Zhang, Guangyao Hu, Yifei Yang, Jian Wang, and Yaoqi Zhou. All-Atom Knowledge-Based Potential for RNA Structure Discrimination Based on the Distance-Scaled Finite Ideal-Gas Reference State. *Journal of Computational Biology*, 27(6):856–867, 2020.

[64] Emidio Capriotti, Tomas Norambuena, Marc A Marti-Renom, and Francisco Melo. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, 27(8):1086–1093, 2011.

[65] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[66] Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alexander B. Wiltschko. A gentle introduction to graph neural networks. *Distill*, 2021. https://distill.pub/2021/gnn-intro.

[67] Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks?, 2022.

[68] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks?, 2018.

[69] David Buterez, Jon Paul Janet, Steven J. Kiddle, Dino Oglic, and Pietro Liò. Graph Neural Networks with Adaptive Readouts, 2022.

[70] Michael Sarver, Craig L Zirbel, Jesse Stombaugh, Ali Mokdad, and Neocles B Leontis. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of mathematical biology*, 56(1):215–252, 2008.