AI Medicine
https://www.sciltp.com/journals/aim

Article

ML-Based RNA Secondary Structure Prediction Methods: A Survey

Qi Zhao [1], Jingjing Chen [1], Zheng Zhao [2], Qian Mao [3], Haoxuan Shi [1] and Xiaoya Fan [4,*]

[1] School of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110000, China
[2] School of Artificial Intelligence, Dalian Maritime University, Dalian 116000, China
[3] Department of Food Science and Engineering, College of Light Industry, Liaoning University, Shenyang 110000, China
[4] School of Software, Dalian University of Technology, Key Laboratory for Ubiquitous Network and Service Software, Dalian 116000, China
* Correspondence: xiaoyafan@dlut.edu.cn

How To Cite: Zhao, Q.; Chen, J.; Zhao, Z.; Mao, Q.; Shi, H.; Fan, X. ML-Based RNA Secondary Structure Prediction Methods: A Survey. *AI Medicine* **2024**, *1*(1), 6. https://doi.org/10.53941/aim.2024.100006.

Received: 6 May 2024
Revised: 17 October 2024
Accepted: 22 October 2024
Published: 29 October 2024

Abstract: The secondary structure of noncoding RNAs (ncRNA) is significantly related to their functions, emphasizing the importance and value of identifying ncRNA secondary structure. Computational prediction methods have been widely used in this field. However, the performance of existing computational methods has plateaued in recent years despite various advancements. Fortunately, the emergence of machine learning, particularly deep learning, has brought new hope to this field. In this review, we present a comprehensive overview of machine learning-based methods for predicting RNA secondary structures, with a particular emphasis on deep learning approaches. Additionally, we discuss the current challenges and prospects in RNA secondary structure prediction..

Keywords: RNA secondary structure prediction; machine learning; deep learning

1. Introduction

Ribonucleic acid (RNA) is an essential substance in most living organisms and plays a nonnegligible role in regulating proteins and biological processes [1]. The RNA molecule comprises a specific sequence of nucleotides arranged in a 5' to 3' direction. Nucleotides include four kinds of bases which are adenine (A), cytosine (C), guanine (G), or uracil (U), and pair up through hydrogen bonds to form the secondary structure [2]. Typically, each base pairs up with only one other base, the most common examples are, the Watson-Crick base pairs (A-U and G-C) and the wobble base pair (G-U). These base pairings often result in a nested structure, where multiple stacked base pairs form a helix, and unpaired base pairs form loops (Figure 1a). It's worth noting that there are three types of special base pairings [2] commonly found in native RNA secondary structures: noncanonical base pairs (Figure 1b), base triples (Figure 1c), and pseudoknots (Figure 1d). Noncanonical base pairs refer to base pairing other than the Watson-Crick and the wobble base pairs, constituting around 40% of all base pairs in structured RNAs [3]. Base triples also widely exist in RNA structures, which involve three bases interacting jointly to stabilize various RNA tertiary interactions [4, 5]. Pseudoknots [6] occur when bases from different loops pair up and then create a non-nested structure between two separated bases. Though pseudoknots only represent a few base pairs in known secondary structures, they play an important role in RNA function [7].

RNA has long been believed to serve only as a messenger between DNA and proteins until the discovery of non-coding RNAs (ncRNAs). It is found that less than 2% of the human genome belongs to protein-coding regions and the rest is transcribed into ncRNAs [8]. NcRNAs are RNAs that do not encode proteins and play functions depending on their structures [9, 10]. Their function includes catalysis, translation, RNA modification, RNA stability, protein synthesis, expression regulation, and protein degradation [11–16]. Moreover, they are important in various human diseases such as cancer, diabetes, and atherosclerosis [13, 17]. Therefore, recognizing ncRNA structure and its involvement in both normal biological processes and pathological conditions has opened new avenues

Copyright: © 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

for researches and potential therapeutic interventions.

Generally, ncRNA molecules form higher-order structures. Unlike proteins, which fold globally driven by hydrophobic forces, RNA folding follows a hierarchical process [18]. RNA in the linear primary structure is folded to form a secondary structure rapidly, resulting in a significant energy loss [19, 20]. Then the secondary structure further forms the tertiary structure at a much slower speed. Though increasing amounts of abundant ncRNA sequences are public [21], most ncRNA structures remain unknown. This structural information deficiency makes it challenging to infer their functions, thus, gaining valid information on ncRNA structures holds great research value. High stability and variety of secondary structures within cells contribute to the crucial role it plays in ncRNA function [22, 23]. Therefore, even without knowing the higher-order structures, the secondary structure alone is often sufficient for inferring function and practical applications [23].

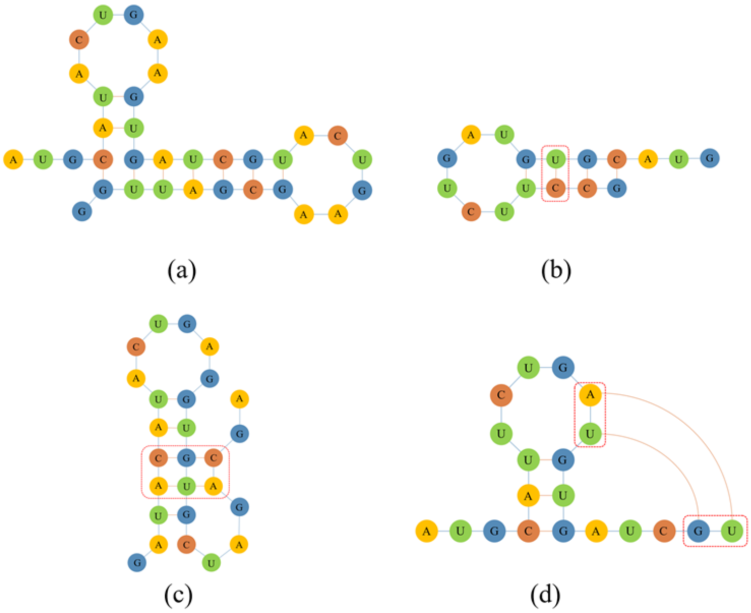
(a) (b) (c) (d)

Figure 1. The yellow, green, blue, and orange circles represent A, U, G, and C, respectively. Yellow lines represent hydrogen bonds, and blue lines represent covalent bonds. (**a**) Common base pair. (**b**) Noncanonical base pair. (**c**) Base triple. (**d**) Typical pseudoknot.

Since the 1970s, a spring of prediction methods has been developed, and computational methods have become the dominant approach for RNA secondary structure prediction. However, the development of both the accuracy and processing speed have stagnated in recent years. Machine learning (ML)-based approaches emerge to address these limitations. Initially, ML-based prediction methods were overlooked due to their simple models and limited accuracy resulting from data scarcity. However, with the rise of RNA datasets and advancements in deep learning (DL), ML methods now surpass traditional approaches in accuracy and applicability, which paves the way for the development of next-generation RNA structure prediction tools [24].

This paper focuses on reviewing methods for predicting RNA secondary structures based on ML and offers a detailed discussion of their pros and cons. Though other previous reviews have covered RNA secondary structure prediction [24–27], there is a lack of reviews focusing especially on DL techniques. This review aims to assist researchers in understanding the current status of RNA secondary structure prediction based on ML and DL while gaining insight into the existing challenges and prospects in this area.

2. Traditional Prediction Methods

RNA secondary structure prediction has been advancing over the past two decades with a variety of methods. Traditional prediction methods include wet-lab experimental experiments and computational predicting methods by algorithm (Figure 2). For wet-lab experiments, Nuclear Magnetic Resonance (NMR) [28] and X-ray crystallography [29] are the two most accurate methods, but they are time-consuming, costly, and limited in applicability. Chemical probing [30, 31] or enzymatic [32, 33] with next-generation sequencing [34, 35] are mainly suitable for in vitro studies but not for in vivo conformation. Only a small fraction of known ncRNAs have been experimentally determined up to now [36], thus, computational methods are accessible alternatives for predicting RNA secondary structures. Comparative sequence analysis [37, 38] is considered the most accurate computational method, which relies on the conservation of RNA secondary structures across evolution compared to primary sequences and identifies base pairs that covary to maintain Watson-Crick and wobble base pairs [39]. Several algorithms have been designed to improve the performance [40–45] and process pseudoknots [46–48]. However, comparative sequence analysis requires a large set of homologous sequences as a basis [49]. When homologous sequences are lacking, another method, namely the score-based method, has been widely used in this field. These methods assume that the native RNA structure has a minimum or maximum total score, transforming the prediction problem into an optimization problem. The scoring schemes for these methods [50–53] typically involve multiple parameters and are based on free energy calculations using the nearest neighbor model. Several methods have been developed to predict structures with pseudoknots [54–59]. However, accurately predicting special base pairs in RNA structures remains a challenging task. In addition, the folding mechanism hypotheses they rely on may not always hold, and the computational cost is extremely high for longer RNA sequences.

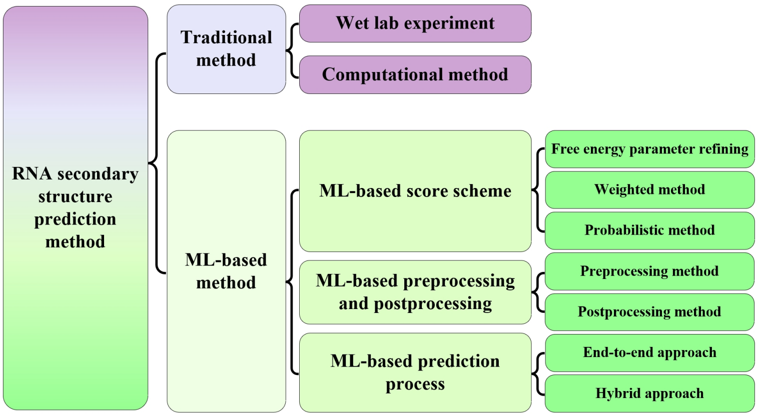

Figure 2. Categories of RNA secondary prediction methods.

3. ML-Based Methods

We classify ML-based RNA secondary structure prediction methods into three categories (Figure 2), namely ML-based score scheme, ML-based preprocessing and postprocessing, and ML-based prediction process, partly referencing the method from our previous paper [24]. These categories are defined based on the subprocess ML applied, and their advantages and limitations are summarized in Table 1. All these methods are trained using supervised learning [60]. These models are trained to learn functions that map input features (such as free energy parameters, encoded RNA sequences, sequence patterns, and evolutionary information) to outputs (including continuous values such as free energy or classification labels such as paired or unpaired bases) by adjusting their parameters using known data. The supervised training process enables ML-based methods to learn patterns from the available data, empowering them to make predictions on unseen data. When a new input is provided, the model can assign a corresponding label or predict a corresponding value based on the learned mapping function [60].

Table 1. Advantages and Limitations of Three Categories of ML-based RNA Secondary Structure Prediction Methods.

Categories	Advantages	Limitations
ML-based score scheme	Provide available parameters for the traditional prediction algorithms to improve the prediction accuracy.	Limited prediction accuracy, particularly for noncanonical base pairs, base triples, and pseudoknots.
ML-based preprocessing and postprocessing	Simplify the prediction process and compatible with traditional prediction algorithms.	The prediction accuracy relies heavily on the intermediate RNA secondary structure prediction model.
ML-based prediction process	Greatly improve the speed and accuracy of prediction, and can predict noncanonical base pairs, base triples, and pseudoknots.	Poor interpretation, high computing costs in model training, and not guaranteed generalization ability on new types of RNA.

3.1. ML-Based Score Scheme

ML-based score scheme aims to train models capable of generating new score schemes, replacing the traditional score scheme (Figure 3). While ML-based methods enhance accuracy through parameter estimation in score schemes, structural prediction remains an optimization problem, where the estimated scores scheme is used for evaluating the potential conformations. ML-based score schemes can be categorized into three types (Figure 2) based on the meaning of scores: the free energy parameter-refining approach, the weighted approach, and the probabilistic approach. The models based on these approaches are detailed in the Supplementary Table S1.

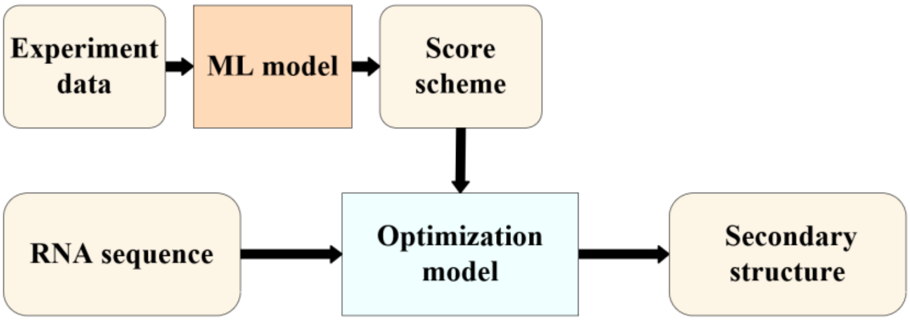

Figure 3. Framework of ML-based score scheme methods. RNA sequences are input into the ML model, and the scoring scheme evaluates the scores of potential conformations and picks out the optical results to output the RNA secondary structures.

3.1.1. ML-Based Free Energy Parameter Refining

Since the publication of the free energy theory, the free energy-focused approach has been widely adopted in score schemes, particularly in assigning free energy values to elements of RNA structures. Among these, Turner's NN model [53] is widely accepted due to its accuracy in approximating free energy. However, determining the multiple thermodynamic parameters in the NN model requires many optimal melting experiments, which are labor-intensive and time-consuming [61, 62]. Then ML techniques have been employed to refine parameters in the energy model, which utilize models to score and provide more abundant features based on known RNA secondary structure data or thermodynamic data. Xia et al. [50] used known thermodynamic data to train a linear regression model for inferring thermodynamic parameters. However, certain structural element parameters are predetermined prior to the computation of other parameters, thereby constraining the potential options for the entire parameter set. To address this limitation, Andronescu et al. [63] put forward a constraint generation approach that employs various constraints to ensure that the energy of reference structures is the lowest among other alternative structures. The team trained this model on a mass of thermodynamic and structural data to infer free energy parameters, achieving a 7% higher F-measure than standard Turner parameters. Further, the research team proposed a Boltzmann-likelihood model and loss-augmented max-margin constraint generation model using a larger set of data to impose constraints on parameters [64]. In addition, it is worth noting that the parameters derived from the above approaches are thermodynamic, so they can be directly applied in algorithms embedded within the same energy model, examples are, RNA folding kinetics simulation [65] and miRNA target prediction [66].

3.1.2. ML-Based Weighted Methods

ML-based free energy parameter refining approaches successfully improve the accuracy of prediction, however, those methods can only be alternatives for wet lab experiments aimed at obtaining energy parameters. Thus, weighting methods were proposed, of which the scoring scheme is independent of the free energy assumption, treating the parameters of the RNA structure as weights rather than free energy changes. Zakov et al. [67] utilized a discriminative structured-prediction learning framework along with an online learning algorithm and significantly expanded the number of weights to around 70,000. The authors achieved this by investigating a wider range of structural elements with more extensive sequential contexts and employing thousands of training datasets. Then ContextFold as a substantial accuracy enhancement model is introduced based on these resulting weights [67]. Akiyama et al. [68] improved a structured support vector machine (SSVM) by the thermodynamic approach to obtain a large set of weights for detailed structural elements. To mitigate overfitting, they applied L1 regularization. Subsequently, they developed MXfold by merging the ML-based weights with experimentally determined thermodynamic parameters, yielding better performance than models solely based on thermodynamic parameters or ML-based weights. Sato et al. improved the model as MXfold2 [69], which uses CNN to calculate the folding scores of RNA sequences, and applies dynamic programming (DP) and the max-margin framework to predict the structure. The max-margin framework includes structural hinge loss function, thermodynamic regularization, and L1 regularization, ensuring that the folding scores are closely aligned with the free energy calculated using the thermodynamic parameters. MXfold2 displayed robust predictions in both sequence-wise and family-wise cross-validation. These studies indicate that weighted approaches based on ML can break through the limitations of the thermodynamic parameter approach. They separate structure prediction from energy estimation, making it advantageous for both tasks. In this case, weighted approaches can achieve more satisfying results. However, a drawback is that the learned weights lack explainability due to the inherent black-box nature of ML algorithms. Therefore, the obtained scores cannot be directly used for computations such as the partition function, centroid structures, or base pair binding probabilities, among others.

3.1.3. ML-Based Probabilistic Methods

As ML technologies improve, stochastic context-free grammars (SCFGs) appear as a significant method for probabilistic approaches for RNA structure prediction [70–74]. SCFGs extend traditional context-free grammars (CFGs) by assigning probabilities to production rules, allowing them to generate structures with different probabilities. It provides a framework for generating diverse possible structures and estimating their probabilities. In an SCFG model, each production rule in the grammar is associated with a probability parameter that assigns a probability to each derived sequence. It typically estimates probability parameters by learning RNA sequence datasets with known secondary structures, eliminating the requirement for external laboratory experiments [73]. The application of SCFGs for tRNA secondary structure prediction was first introduced by Sakakibara et al. [70]. They used an expectation-maximization (EM) method to learn the probability parameters. Knudsen and Hein [72] further enhanced the SCFG model by incorporating evolutionary information, leading to the development of the robust and practical tool Pfold [72].

Sato et al. [75] proposed a nonparametric Bayesian extension of SCFGs using the hierarchical Dirichlet process (HDP). Traditional SCFGs are required to define a fixed number of generation rules and parameters in advance, whereas non-parametric Bayesian extension allows SCFGs to adaptively learn the complexity and structure of the model. Thus, it is flexible to different data and identifies an optimal RNA grammar from the training dataset expressively and adaptably. To leverage the abundance of RNA sequences with unknown structures, Yonemoto et al. [76] proposed a semi-supervised learning algorithm to improve prediction accuracy. This algorithm determines probability parameters in a probabilistic model that combines SCFG and a conditional random field, enabling the incorporation of both labeled and unlabeled data. Even though, the probabilistic approach, such as SCFG, cannot fully take the place of Minimum Free Energy (MFE) methods, since the accuracy of the best SCFG models still falls short of the top-performing free energy-based models. Additionally, SCFGs have limitations in describing certain RNA structures, such as those containing special base pairs that deviate from the conventional Watson-Crick base pairing. These constraints highlight the importance of considering both probabilistic and free energy-based approaches to achieve more accurate and comprehensive RNA structure predictions.

Do et al. [77] proposed a new method CONTRAfold without physics-based models. Novelly, CONTRAfold applies the conditional log-linear model (CLLM) to determine probability parameters that effectively differentiate correct RNA structures from incorrect ones. CLLM is a flexible probabilistic ML model that allows easy parameter estimation and incorporation of any chosen features into the model, providing a framework for capturing complex relationships between input features and target variables. Compared to previously available probabilistic models, CONTRAfold reaches the highest accuracy in single-sequence RNA structure prediction. However, CLLM is computationally slower than SCFGs, limiting its application to large-scale training datasets. Since CLLM imposes fewer structural constraints on the output sequence, when encountering sequences with specific base pairs, it potentially leads to the possibility of generating false RNA structures. In addition, the estimated parameters lack explicit biological interpretation due to its black-box feature.

3.2. ML-Based Preprocessing and Postprocessing

ML can be applied in preprocessing to simplify the prediction process (Figure 4). Hor et al. [78] introduced a tool based on support vector machine (SVM) that aims to choose the most effective prediction method. They believe that different RNA sequences possess distinct features so that specific prediction methods perform better for one RNA sequence. By utilizing SVM, the tool can identify the prediction method that is likely to yield the best results for a given RNA sequence. Similarly, based on the assumption that folding rules differ from RNA sequences, Zhu et al. [79] put forth an SCFG model to identify the most probable folding rules for an RNA sequence ahead of the prediction process. By doing so, the accuracy of the prediction can be improved.

Additionally, processing long RNA sequences can be costly, time-consuming, and complicated. To solve this problem, Zhao and colleagues [80] designed a DL-based model RNA-par using transfer learning. RNA-par splits RNA sequence into several independent fragments (i-fragments) to improve prediction performance. RNA-par consists of a 4-layer 1D-CNN block for extracting sequence features, a Bi-LSTM block capturing information from both sequences, and a 2-layer ResNet block acting as prediction head to generate the outputs i-fragments. Since i-fragments are shorter sequences, RNA-par makes it convenient for the following prediction process.

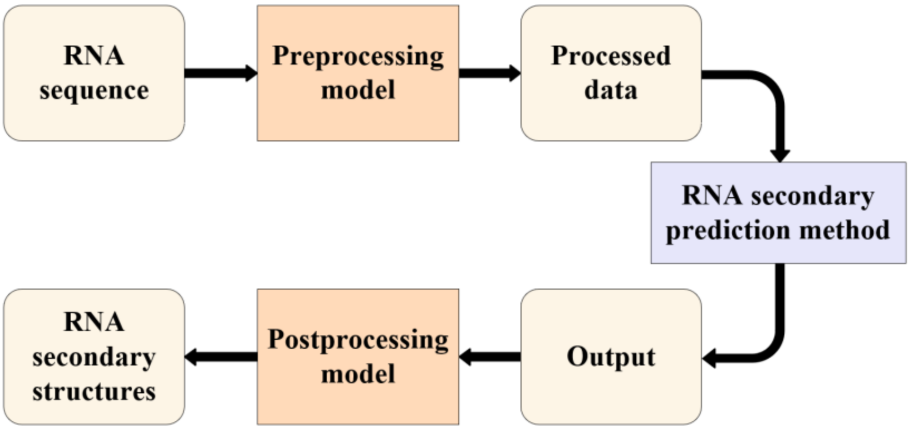

Figure 4. Framework of preprocessing and postprocessing. To simplify the prediction process in the ML-based prediction model, the preprocessing model processes RNA sequences into other forms of data, and the postprocessing model transcript outputs into RNA secondary structures.

ML is also used in postprocessing to reach a better result. Since various methods yield multiple structures for an RNA sequence, postprocessing models can be utilized to determine the most probable structures among the outcomes (Figure 4). Haynes et al. [81] combined ML with graph theory to represent RNA graphical structures using trees, where edges represent helices and vertices represent bulges or loops. Using graphical invariants as input features, a multilayer perceptron (MLP) model is trained to identify whether the result is an RNA structure. This approach enables the ML model to distinguish between structures that are likely to represent RNA structures and those that are not. Additionally, an assumption from Koessler et al. [82] indicates that a larger one is formed when two smaller RNA secondary structures bond together. They extract a feature vector from the merged trees and apply it to an MLP model to predict the probability of an RNA-like structure. By leveraging this MLP model, they were able to estimate the likelihood of a given structure resembling an RNA-like structure. Details of above models are summarized in Supplementary Table S2.

3.3. ML-Based Predicting Process

ML techniques can be utilized as end-to-end prediction approaches or integrated with other algorithms as filters or optimizers. The models based on both approaches are detailed in the Supplementary Table S3.

3.3.1. End-To-End Approaches

End-to-end approaches usually directly predict the secondary structure from the RNA sequence without intermediate steps or external information (Figure 5). They aim to capture the inherent structure-sequence relationship in training sample, and learn the mapping between the sequence and its secondary structure in a single integrated model.

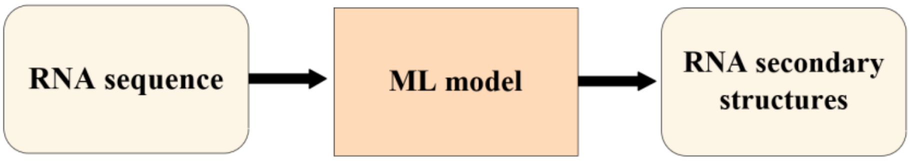

Figure 5. Framework of end-to-end approaches. End-to-end approaches directly predict secondary structures from RNA sequences without intermediate steps or external information.

Built upon Nussinov and Jacobson's hypothesis [46], ML techniques were first introduced to RNA secondary structure prediction by Takefuji et al. [83]. They used a system of interactional neurons to obtain a near-maximum independent set (MIS) from an adjacent graph representing base pairs. To improve this work, Qasim et al. [84] built a new MLP model with h neurons in the hidden layer to obtain MIS, and its activation function is based on Kolgomorov's theorem (h representing the number of possible base pairs in an RNA sequence). In other aspects, Liu et al. [85] considered the energy contribution of base pairs and employed a Hopfield neural network (HNN) to get the MIS. Apolloni et al. [86] enhanced computational speed and applied mean-field approximation in both the instant resolution and learning phases, slightly enlarging the input RNA length for this approach. Unfortunately, HNN was limited by its susceptibility to local minima, so Steeg and Evan [87] utilized mean field theory (MFT) networks coupled with an objective function and biological constraints to identify the optimal structure. In this method, MFT receives four types of RNA bases that are encoded in a one-hot fashion and outputs a CT-like table.

However, since ML-based models are limited to processing tRNAs only due to the lack of data, DL techniques are thriving to break through the challenges. Singh et al. [80] proposed SPOT-RNA, the first end-to-end DL model for RNA secondary structure prediction. SPOT-RNA turns sequences into CT tables and employs ultra-deep hybrid networks consisting of ResNets and Bi-BLSTMs. ResNets obtain the contextual information while Bi-BLSTMs capture dependencies between distant nucleotides in the RNA sequence. SPOT-RNA has an outstanding performance on benchmark datasets compared to score-based methods and SCFG-based methods. The same team later introduced the SPOT-RNA2 model [88], which incorporated evolution-derived sequence profiles and mutational coupling, outperforming the model SPOT-RNA. Furtherly, Fu and colleagues [89] introduced a special model UFold which converts the sequence into an image of all possible base pairs, and processes through U-net and a 1D-convolution to generate contact scores between bases. Unlike other models, it innovatively abandons raw sequences but adopts a 3D vector analogous to an image as input, making the model fully convolutional to achieve higher efficiency and ability to process pseudoknots. Another DL model, E2Efold, put forward by Chen et al. [90], consists of a transformer-based deep model and a multilayer network based on an unrolled algorithm. E2Efold takes the RNA sequence as input, employs the deep model to encode the sequence information, and then the multilayer network to filter the output. One of the advantages of E2Efold is its ability to process longer RNA sequences, including those large molecules with complex structures. It is also able to capture non-local interactions in the sequence and take these relationships into account when generating secondary structures. However, E2Efold suffers from a severe overfitting, and generalized on unseen RNAs.

Besides primary sequences, DL models can be combined with other information. Calonaci et al. [91] trained an ensemble model that combines co-evolutionary data (DCA), SHAPE data, and RNA sequence data. It has an MLP subnetwork based on DCA data and a CNN subnetwork based on SHAPE data for predicting penalties, then its folding module generates structures using penalties and RNA sequences.

3.3.2. Hybrid Approaches

Hybrid approaches that combine ML models with other methods have been explored for predicting RNA secondary structure prediction. One of these approaches combined ML models with filters to predict a possible structure and another is to hybrid ML models with optimization methods. The framework is shown in Figure 6.

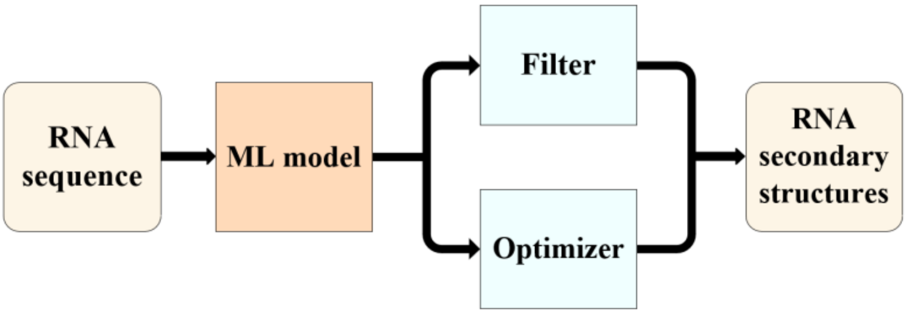

Figure 6. Framework of hybrid approaches. One of the hybrid approaches combines ML models with filters to predict a possible structure, the other combines ML models with optimization methods.

ML Filter Combined Methods

For these methods, they typically include an ML model and a filter in the process to achieve the output. Bindewald and Shapiro [42] integrated the ML model with a filter to reach the consensus structure of a set of aligned RNAs. The model gets the possibility score for each pair of alignment columns by employing a hierarchical network of k-nearest neighbor models. Filters with rules of native RNA structures constrain the result of the ML model to get outputs. Wu et al. [92] and Lu et al. [93] regarded predicting structure as a sequence-labeling problem, and they predicted the state of bases by Bi-LSTM and applying a rule-based filter to cope with controversial pairings. Another innovative model DpacoRNA [94] used Bi-LSTM as a structured filter and employed a parallel ant colony optimization method to hunt for the maximum probable structures. A recent study [95] constructed an composite network that integrates Bi-LSTM [96], Transformer [97], and U-Net [98] to calculate pairwise scores between bases. Utilizing four established rules, the network constructs a filter to discern potential RNA structures.

ML Optimization Combined Methods

In these approaches, the ML model finds the relationship between each base or each pair of bases, and the optimization method picks out the optimal structure. CNNFold model, published by Booy et.al [99], consists of multiple residual blocks and a readout layer post-processing to predict a score matrix for all possible pairings. They also developed an algorithm called Argmax post-processing converting the score matrix into the best secondary structure. It is worth noting that CNNFold and its variants can predict structures with pseudoknots well. Similar to CNNFold, Liu's group [100] proposed a model combined with DL and DP, that predicts the status distribution of each base by the CNN model and finds the most probable structure using the DP algorithm. To improve the result, they replaced the CNN with the Bi-LSTM model integrated with another optimization algorithm [101]. Willmott et al. [102] adopted the SHAPE-directed method (SDM) to predict optimal structure rather than developing a new optimization model, meanwhile, they trained a Bi-LSTM model that generates SHAPE-like data of an RNA sequence as the inputs for SDM. Recently, Chen and Chan [103] proposed a DL-based model, REDFold, which utilizes the UFold [89] architecture. Its encoder accepts a 2D contact matrix as input, while the decoder yields a score map. They employ constrained optimization instead of DP to identify the optimal structure, thereby enabling their model to predict non-nested folding patterns.

4. Databases

In ML-based RNA structure prediction research, access to comprehensive and reliable structural data is essential for model training and performance evaluation. Generally, the quantity and quality of training data directly influence the learning ability and prediction accuracy of an ML-based model. A rich set of training samples helps models capture a more comprehensive range of RNA structural features and enhances its robustness when faced with unseen data. In addition, the representativeness of the database is vital. A sample containing various types of RNA secondary structures enables the model to better understand the complexity of RNA structures.

Several databases have been developed to provide researchers with extensive resources for studying RNA sequences and structures. Among these, some databases offer a broad spectrum of RNA data (Such as Comprehensive Databases), including diverse species and structural conformations, while others focus on specific aspects or types of RNA (Such as Specialized Databases).

4.1. Comprehensive Databases

Comprehensive databases are large, general-purpose collections of RNA structures that include a wide variety of RNA species and structural conformations. These databases often contain a large number of experimentally obtained RNA structures or computational predictions, making them useful for ML-based RNA structure prediction. RNA STRAND [104] is a database that provides a diverse collection of RNA sequences, containing 4,666 RNA samples. It is designed to offer structural and sequence information for RNA research. RCSB Protein Data Bank (PDB) [105] is an authoritative database for biomolecular structures, providing 4,962 RNA structures. It primarily includes tertiary structures obtained through experimental methods, such as X-ray crystallography and nuclear magnetic resonance, which offer a solid foundation for analyzing the conformation of RNA. bpRNA-1m [106] is a large-scale database, offering 102,348 RNA structures, which are mainly constructed using a novel annotation tool called bpRNA. While the accuracy of secondary structures provided by bpRNA-1m is relatively lower, its vast data volume makes it valuable for ML-based RNA secondary structure prediction models. RNAcentral [107] is the largest RNA secondary structure database, containing secondary structures obtained by computational methods R2DT [108].

4.2. Specialized Databases

The tRNAdb 2009 database [109] is one of the earliest specialized databases, which focuses on the structures and functions of tRNA. It provides detailed tRNA sequences and their corresponding structural information. The rRNA database [110] is dedicated to structural data of ribosomal RNA (rRNA). The tmRDB database [111] focuses on post-transcriptionally modified RNA (tmRNA), which plays an important role in bacteria, participating in protein synthesis and quality control. In addition, there are also some specialized structure databases. These databases typically focus on one specific type of RNA structure, such as loop [112], pseudoknot [113], or non-canonical base pair [114]. In addition, based on these specialized databases, benchmark datasets such as ArchiveII [115] and RNAStralign [116] have been established. They contain various types of RNA sequences with high sequence diversity, making them especially suitable for training and evaluating the performance of RNA structure prediction models.

4.3. Other Important Databases

In addition, there are other important databases such as Rfam [117] and NNDB [53]. Rfam is a widely used RNA family database, which provides classification information including consensus secondary structures and a covariance model for each RNA family. NNDB provides crucial thermodynamic parameters for modeling the stability of RNA secondary structures, especially when calculating RNA folding energies. NNDB provides foundational data for machine learning models, helping improve the accuracy of RNA secondary structure prediction.

5. Discussions

As it is known to all, the abundance of transcripts is widely recognized as a valuable indicator for identifying transcripts of interest in different conditions, while understanding RNA structure is crucial for unraveling their functional mechanisms. A highly accurate RNA structure prediction method also has implications for various downstream investigations, including but not limited to, simulations of folding dynamics [118], the detection of ncRNAs [64, 119, 120], applications in oligonucleotide [121, 122] or drug design [123–127], and assessment of hybridization stability [128]. Even more, RNA secondary structure prediction also serves as a valuable tool in studying viruses, an example is, the SARS-CoV-2 virus [129, 130].

5.1. Pros of ML-Based Methods

ML-based methods offer several advantages over comparative sequence analysis and traditional score-based methods. Firstly, rather than rely on intricate biological mechanisms, ML-based methods tend to leverage information from diverse data types, bypassing performance limitations imposed by specific mechanism hypotheses. ML-based methods are also easy to integrate with known biological mechanisms, providing a flexible framework for analysis. Having approaches to the mass of available datasets, models knowing less knowledge of biological mechanisms usually outperform those models dependent on biological mechanisms. This infers the guess that current biological mechanisms of RNA folding might be faulty. Secondly, ML-based methods, particularly end-to-end DL models, eliminate the need to consider base matching rules. In traditional score-based methods, they utilize complex algorithms to meet base matching rules, which causes high time complexity, leaving difficulties for them to improve. In contrast, end-to-end DL models [80] can be trained to predict all base pairs in RNA structures without these rules, despite whether these base pairs are associated with secondary or tertiary interactions. Thirdly, ML-based methods offer considerable flexibility compared to traditional methods. The input data for ML models are various, no matter whether they are one-dimensional or multidimensional, homogeneous or heterogeneous, features extracted from the data or encoded bases, matrixes, or diagrams. Similarly, the outputs of ML models can also vary, including CT tables, nucleotide states, labeled sequences, or free energy values. ML models can be constructed using a diverse array of techniques, ranging from simple Hopfield networks to complex ensemble DNNs. Additionally, similarly to tasks in Natural Language Processing, RNA secondary structure prediction can also benefit as a downstream task by utilizing representations obtained from pre-trained foundation models [131–133] as inputs, thereby enhancing the accuracy of the predictive model's structure predictions. Lastly, end-to-end prediction methods exhibit fast runtime once ML models are trained. Outperformed the DP algorithm, the time complexity of ML models remains independent of the input scale, providing potential capacities for processing long RNA sequences. In summary, ML-based methods offer advantages such as flexibility, independence from base matching rules, and fast runtime, making them a promising approach for RNA structure prediction.

5.2. Remaining Challenges and Prospects

Though RNA secondary structure prediction methods using ML techniques are considered state-of-the-art in terms of prediction performance across various measures, there are still some issues needed to be addressed. To begin with, there is a need to further enhance the accuracy of predictions. Surveys [69, 88] show that there is a long way to go in improving the accuracy of RNA secondary structure prediction methods. On one hand, since RNA structures unpredictably vary in different cellular environments [134], multiple structure options instead of the most possible one should be considered when analyzing input sequences to gain predictions is worthy of consideration. On the other hand, combining an ML-based method with an optimization approach shows promise in enhancing prediction performance. ML-based methods can leverage their ability to learn complex patterns from data and make accurate predictions while optimization methods can refine and optimize the predicted structures to align with known structural constraints and principles. This combination offers a synergistic approach that combines the strengths of both paradigms, showing its potential for future development. ML-based prediction of RNA secondary structures relies on capturing the interactions between distant nucleotides, however, when dealing with long RNA sequences, getting these long-range interactions within RNA sequences can be a challenge. Meanwhile, training a large-scale inputs ML model demands impractical computational resources. To face this sequence length limitation, striking a balance between computational efficiency and capturing long-range interactions is considerable. Innovative approaches such as hierarchical modeling, integration of experimental data, and leveraging parallel and distributed computing resources are expected solutions to develop. Furthermore, numerous traditional approaches disregard special base pairs to minimize the occurrence of false positives and reduce computational complexity [57, 135]. Though certain methods can process structures with non-canonical base pairs [136] or pseudoknots [48] none of them can accurately predict both simultaneously, even ML-based methods suffer limited accuracy. Therefore, finding solutions for predicting special base pairs is an inevitable future trend. Overfitting is another critical concern for ML-based RNA secondary structure prediction models, particularly for DL models [74]. Overfitted models tend to perform well on RNAs that are structurally similar to training data but poor on structure-dissimilar ones. Instead of truly learning the folding mechanism, these models often end up memorizing the secondary structure patterns present in the training data. Although DL-based methods employ various techniques to mitigate overfitting, such as regularization [91], constraint addition [90], dataset enlargement [80], or integration of Turner's nearest neighbor free energy parameters, concerns regarding overfitting persist.

6. Conclusion

Understanding RNA structure is crucial for comprehending biological processes, and the prediction of RNA secondary structure remains a prominent topic in the fields of computation and biology. Though ML techniques have significantly enhanced the accuracy, applicability, and computational speed of the prediction process, more sophisticated ML models and DL technologies are needed to facilitate the development of a new generation of RNA secondary structure prediction tools with improved accuracy and computational efficiency.

Supplementary Materials

The following supporting information can be downloaded at: https://www.sciltp.com/journals/aim/2024/1/363/s1. Table S1: ML-based scorescheme, Table S2: ML-based preprocessing and postprocessing, Table S3: ML-based predicting process.

Author Contributions

Q.Z.: Project administration, writing–review, editing; J.C.: Writing–original draft, visualization; Z.Z.: Writing–review, editing; Q.M.: Writing–review, editing; H.S.: Writing–review, editing, visualization; X.F.: Project administration, supervision, Writing–review, editing.

Funding

This research was funded by grants from General Project of Science and Technology Foundation of Liaoning Province of China, grant number 2023-MS-091.

Data Availability Statement

Not applicable.

Conflicts of Interest The authors declare no conflict of interest.

References

1. Wang, D.; Farhana, A. Biochemistry, RNA Structure. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2024.
2. Zhao, Y.; Wang, J.; Zeng, C.; Xiao, Y. Evaluation of rna secondary structure prediction for both base-pairing and topology. *Biophys. Rep.* **2018**, *4*, 123–132.
3. Leontis, N.B.; Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **2001**, *7*, 499–512.
4. Almakarem, A.S.A.; Petrov, A.I.; Stombaugh, J.; Zirbel, C.L.; Leontis, N.B. Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Res.* **2012**, *40*, 1407–1423.
5. Doherty, E.A.; Batey, R.T.; Masquida, B.; Doudna, J.A. A universal mode of helix packing in RNA. *Nat.*

Struct. Biol. **2001**, *8*, 339–343.

6. Van Batenburg, F.H.D.; Gultyaev, A.P.; Pleij, C.W.A. PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Res.* **2001**, *29*, 194–195.
7. Staple, D.W.; Butcher, S.E. Pseudoknots: RNA Structures with Diverse Functions. *PLoS Biol.* **2005**, *3*, e213.
8. ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **2012**, *489*, 57–74.
9. Kovalchuk, I. Chapter 24-Non-coding RNAs in genome integrity. In *Genome Stability*, 2nd ed.; Kovalchuk, I., Kovalchuk, O., Eds.; Volume 26 of Translational Epigenetics; Academic Press: Boston, MA, USA 2021; pp. 453–475.
10. Kasprzyk, M.E.; Kazimierska, M.; Sura, W.; Dzikiewicz-Krawczyk, A.; Podralska, M. Chapter 3-Non-coding RNAs: Mechanisms of action. In *Navigating Non-Coding RNA*; Sztuba-Solinska, J., Ed.; Academic Press: Cambridge, MA, USA, 2023; pp. 89–138.
11. Doudna, J.A.; Cech, T.R. The chemical repertoire of natural ribozymes. *Nature* **2002**, *418*, 222–228.
12. Higgs, P.G.; Lehman, N. The RNA World: Molecular cooperation at the origins of life. *Nat. Rev. Genet.* **2015**, *16*, 7–17.
13. Mortimer, S.A.; Kidwell, M.A.; Doudna, J.A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **2014**, *15*, 469–479.
14. Meister, G.; Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **2004**, *431*, 343–349.
15. Serganov, A.; Nudler, E. A Decade of Riboswitches. *Cell* **2013**, *152*, 17–24.
16. Wu, L.; Belasco, J.G. Let me count the ways: Mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell* **2008**, *29*, 1–7.
17. Zou, Q.; Li, J.; Hong, Q.; Lin, Z.; Wu, Y.; Shi, H.; Ju, Y. Prediction of MicroRNA-Disease Associations Based on Social Network Analysis Methods. *BioMed Res. Int.* **2015**, *2015*, 810514, .
18. Tinoco, I.; Bustamante, C. How RNA folds. *J. Mol. Biol.* **1999**, *293*, 271–281.
19. Georgakopoulos-Soares, I.; Parada, G.E.; Hemberg, M. Secondary structures in RNA synthesis, splicing and translation. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2871–2884.
20. Celander, D.W.; Cech, T.R. Visualizing the higher order folding of a catalytic RNA molecule. *Science* **1991**, *251*, 401–407.
21. Stephens, Z.D.; Lee, S.Y.; Faghri, F.; Campbell, R.H.; Zhai, C.; Efron, M.J.; Iyer, R.; Schatz, M.C.; Sinha, S.; Robinson, G.E. Big Data: Astronomical or Genomical? *PLoS Biol.* **2015**, *13*, e1002195.
22. Zarrinkar, P.P.; Williamson, J.R. Kinetic intermediates in RNA folding. *Science* **1994**, *265*, 918–924.
23. The statistical mechanics of RNA folding. *Physics* **2006**, *35*, 218–229.
24. Zhao, Q.; Zhao, Z.; Fan, X.; Yuan, Z.; Mao, Q.; Yao, Y. Review of machine learning methods for RNA secondary structure prediction. *PLoS Comput. Biol.* **2021**, *17*, e1009291.
25. Condon, A. Problems on RNA Secondary Structure Prediction and Design. In *Automata, Languages and Programming*; (Baeten, J.C.M., Lenstra, J.K., Parrow, J., Woeginge, G., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2003; pp. 22–32.
26. Fallmann, J.; Will, S.; Engelhardt, J.; Grüning, B.; Backofen, R.; Stadler, P.F. Recent advances in RNA folding. *J. Biotechnol.* **2017**, *261*, 97–104.
27. Seetin, M.G.; Mathews, D.H. RNA structure prediction: An overview of methods. *Methods Mol. Biol.* **2012**, *905*, 99–122.
28. Fürtig, B.; Richter, C.; Wöhnert, J.; Schwalbe, H. NMR spectroscopy of RNA. *ChemBioChem* **2003**, *4*, 936–962.
29. Westhof, E. Twenty years of RNA crystallography. *RNA* **2015**, *21*, 486–487.
30. Tijerina, P.; Mohr, S.; Russell, R. DMS Footprinting of Structured RNAs and RNA-Protein Complexes. *Nat. Protoc.* **2007**, *2*, 2608–2623.
31. Wilkinson, K.A.; Merino, E.J.; Weeks, K.M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.* **2006**, *1*, 1610–1616.
32. Kertesz, M.; Wan, Y.; Mazor, E.; Rinn, J.L.; Nutter, R.C.; Chang, H.Y.; Segal, E. Genome-wide Measurement of RNA Secondary Structure in Yeast. *Nature* **2010**, *467*, 9322.
33. Underwood, J.G.; Uzilov, A.V.; Katzman, S.; Onodera, C.S.; Mainzer, J.E.; Mathews, D.H.; Lowe, T.M.; Salama, S.R.; Haussler, D. FragSeq: transcriptome-wide RNA structure probing using high-throughput

sequencing. *Nat. Methods* **2010**, *7*, 995–1001.

34. Bevilacqua, P.C.; Ritchey, L.E.; Su, Z.; Assmann, S.M. Genome-Wide Analysis of RNA Secondary Structure. *Annu. Rev. Genet.* **2016**, *50*, 235–266.
35. Tian, S.; Das, R. RNA structure through multidimensional chemical mapping. *Q. Rev. Biophys.* **2016**, *49*, e7.
36. RNAcentral: A comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.* **2017**, *45*, D128–D134.
37. Gutell, R.R.; Lee, J.C.; Cannone, J.J. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* **2002**, *12*, 301–310.
38. Madison, J.T.; Everett, G.A.; Kung, H. Nucleotide sequence of a yeast tyrosine transfer RNA. *Science* **1966**, *153*, 531–534.
39. Gutell, R.R.; Weiser, B.; Woese, C.R.; Noller, H.F. Comparative anatomy of 16-S-like ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **1985**, *32*, 155–216.
40. Ruan, J.; Stormo, G.D.; Zhang, W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* **2004**, *20*, 58–66.
41. Hofacker, I.L.; Fekete, M.; Flamm, C.; Huynen, M.A.; Rauscher, S.; Stolorz, P.E.; Stadler, P.F. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **1998**, *26*, 3825–3836.
42. Bindewald, E.; Shapiro, B.A. Rna secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* **2006**, *12*, 342–352.
43. Legendre, A.; Angel, E.; Tahi, F. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. *BMC Bioinformatics* **2018**, *19*, 1–15.
44. Han, K.; Kim, H.J. Prediction of common folding structures of homologous RNAs. *Nucleic Acids Res.* **1993**, *21*, 1251–1257.
45. Tahi, F.; Gouy, M.; Régnier, M. Automatic RNA secondary structure prediction with a comparative approach. *Comput. Chem.* **2002**, *26*, 521–530.
46. Nussinov, R.; Jacobson, A.B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 6309–6313.
47. Engelen, S.; Tahi, F. Tfold: Efficient in silico prediction of non-coding RNA secondary structures. *Nucleic Acids Res.* **2010**, *38*, 2453–2466.
48. Bellaousov, S.; Mathews, D.H. ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA* **2010**, *16*, 1870–1880.
49. Burge, S.W.; Daub, J.; Eberhardt, R.; Tate, J.; Barquist, L.; Nawrocki, E.P.; Eddy, S.R.; Gardner, P.P.; Bateman, A. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **2013**, *41*, D226–D232.
50. Xia, T.; SantaLucia, J.; Burkard, M.E.; Kierzek, R.; Schroeder, S.J.; Jiao, X.; Cox, C.; Turner, D.H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **1998**, *37*, 14719–14735.
51. Mathews, D.H.; Sabina, J.; Zuker, M.; Turner, D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **1999**, *288*, 911–940.
52. Andronescu, M.; Condon, A.; Turner, D.H.; Mathews, D.H. The determination of RNA folding nearest neighbor parameters. *Methods Mol. Biol.* **2014**, *1097*, 45–70.
53. Turner, D.H.; Mathews, D.H. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* **2010**, *38*, D280–D282.
54. Bon, M.; Micheletti, C.; Orland, H. McGenus: A Monte Carlo algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res.* **2013**, *41*, 1895–1900.
55. Reeder, J.; Giegerich, R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinform.* **2004**, *5*, 104.
56. Dirks, R.M.; Pierce, N.A. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* **2003**, *24*, 1664–1677.
57. Rivas, E.; Eddy, S.R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **1999**, *285*, 2053–2068.
58. Sato, K.; Kato, Y. Prediction of RNA secondary structure including pseudoknots for long sequences. *Brief. Bioinform.* **2021**, *23*, bbab395.
59. Poolsap, U.; Kato, Y.; Akutsu, T. Prediction of RNA secondary structure with pseudoknots using integer

programming. *BMC Bioinformatics* **2009**, *10*, 1–11.

60. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260.
61. Lorenz, R.; Bernhart, S.H.; Siederdissen, C.H.Z.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26.
62. Bellaousov, S.; Reuter, J.S.; Seetin, M.G.; Mathews, D.H. RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.* **2013**, *41*, W471–W474.
63. Andronescu, M.; Condon, A.; Hoos, H.H.; Mathews, D.H.; Murphy, K.P. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* **2007**, *23*, i19–i28.
64. Washietl, S.; Will, S.; Hendrix, D.A.; Goff, L.A.; Rinn, J.L.; Berger, B.; Kellis, M. Computational analysis of noncoding RNAs. *Wiley Interdiscip. Rev. RNA* **2012**, *3*, 759–778.
65. Tang, X.; Thomas, S.; Tapia, L.; Giedroc, D.P.; Amato, N.M. Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.* **2008**, *381*, 1055–1067.
66. Rehmsmeier, M.; Steffen, P.; Höchsmann, M.; Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **2004**, *10*, 1507–1517.
67. Zakov, S.; Goldberg, Y.; Elhadad, M.; Ziv-Ukelson, M. Rich parameterization improves RNA structure prediction. *J. Comput.Biol. A J. Comput. Mol. Cell Biol.* **2011**, *18*, 1525–1542.
68. Akiyama, M.; Sato, K.; Sakakibara, Y. A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. *J. Bioinform. Comput. Biol.* **2018**, *16*, 1840025.
69. Sato, K.; Akiyama, M.; Sakakibara, Y. Rna secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **2021**, *12*, 941.
70. akakibara, Y.; Brown, M.; Hughey, R.; Mian, I.S.; Sjölander, K.; Underwood, R.C.; Haussler, D. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* **1994**, *22*, 5112–5120.
71. Woodson, S.A. Recent insights on RNA folding mechanisms from catalytic RNA. *Cell. Mol. Life Sci.* **2000**, *57*, 796–808.
72. Knudsen, B.; Hein, J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **1999**, *15*, 446–454.
73. Dowell, R.D.; Eddy, S.R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinform.* **2004**, *5*, 71.
74. Rivas, E.; Lang, R.; Eddy, S.R. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **2012**, *18*, 193–212.
75. Sato, K.; Hamada, M.; Mituyama, T.; Asai, K.; Sakakibara, Y. A non-parametric bayesian approach for predicting RNA secondary structures. *J. Bioinform. Comput. Biol.* **2010**, *8*, 727–742.
76. Yonemoto, H.; Asai, K.; Hamada, M. A semi-supervised learning approach for RNA secondary structure prediction. *Comput. Biol. Chem.* **2015**, *57*, 72–79.
77. Do, C.B.; Woods, D.A.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22*, e90–e98.
78. Hor, C.-Y.; Yang, C.-B.; Chang, C.-H.; Tseng, C.-T.; Chen, H.-H. A tool preference choice method for RNA secondary structure prediction by SVM with statistical tests. *Evol. Bioinform.* **2013**, *9*, EBO–S10580.
79. Zhu, Y.; Xie, Z.; Li, Y.; Zhu, M.; Chen, Y.-P.P. Research on folding diversity in statistical learning methods for RNA secondary structure prediction. *Int. J. Biol. Sci.* **2018**, *14*, 872–882.
80. Singh, J.; Hanson, J.; Paliwal, K.; Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **2019**, *10*, 5407.
81. Haynes, T.; Knisley, D.; Knisley, J. Using a neural network to identify secondary RNA structures quantified by graphical invariants. *Commun. Math. Comput. Chem.* **2008**, *60*, 277–290.
82. Koessler, D.R.; Knisley, D.J.; Knisley, J.; Haynes, T. A predictive model for secondary RNA structure using graph theory and a neural network. *BMC Bioinform.* **2010**, *11*, S21.
83. Takefuji, Y.; Chen, L.L.; Lee, K.C.; Huffman, J. Parallel algorithms for finding a near-maximum independent set of a circle graph. *IEEE Trans. Neural Netw.* **1990**, *1*, 263–267.
84. Qasim, R.; Kauser, N.; Jilani, T. Secondary Structure Prediction of RNA using Machine Learning Method. *Int. J. Comput. Appl.* **2010**, *10*, 15–22.
85. Liu, Q.; Ye, X.; Zhang, Y. A Hopfield Neural Network Based Algorithm for RNA Secondary Structure Prediction. In Proceedings of the First International Multi-Symposiums on Computer and Computational

Sciences (IMSCCS'06), Hangzhou, China, 20–24 June 2006; Volume 1, pp. 10–16.

86. Apolloni, B.; Lotorto, L.; Morpurgo, A.; Zanaboni, A.M. RNA Secondary Structure Prediction by MFT Neural Networks. *Psychol. Forsch.* **2003**, *2003*, 143–148.
87. Steeg, E.W. *Neural Networks, Adaptive Optimization, and RNA Secondary Structure Prediction*; American Association for Artificial Intelligence: Palo Alto, CA, USA, 1993; pp. 121–160.
88. Singh, J.; Paliwal, K.; Zhang, T.; Singh, J.; Litfin, T.; Zhou, Y. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics* **2021**, *37*, 2589–2600.
89. Fu, L.; Cao, Y.; Wu, J.; Peng, Q.; Nie, Q.; Xie, X. UFold: Fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **2022**, *50*, e14.
90. Chen, X.; Li, Y.; Umarov, R.; Gao, X.; Song, L. RNA Secondary Structure Prediction By Learning Unrolled Algorithms. *arXiv* **2020**, arXiv:2002.05810.
91. Calonaci, N.; Jones, A.; Cuturello, F.; Sattler, M.; Bussi, G. Machine learning a model for RNA structure prediction. *NAR Genom. Bioinform.* **2020**, *2*, lqaa090.
92. Wu, H.; Tang, Y.; Lu, W.; Chen, C.; Huang, H.; Fu, Q. RNA Secondary Structure Prediction Based on Long Short-Term Memory Model. In *Intelligent Computing Theories and Application*; Huang, D.-S., Bevilacqua, V., Premaratne, P., Gupta, P., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; pp. 595–599.
93. Lu, W.; Tang, Y.; Wu, H.; Huang, H.; Fu, Q.; Qiu, J.; Li, H. Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter. *BMC Bioinform.* **2019**, *20*, 684.
94. Quan, L.; Cai, L.; Chen, Y.; Mei, J.; Sun, X.; Lyu, Q. Developing parallel ant colonies filtered by deep learned constrains for predicting RNA secondary structure with pseudo-knots. *Neurocomputing* **2020**, *384*, 104–114.
95. Fei, Y.; Zhang, H.; Wang, Y.; Liu, Z.; Liu, Y. LTPConstraint: a transfer learning based end-to-end method for RNA secondary structure prediction. *BMC Bioinformatics* **2022**, *23*, 354.
96. Hochreiter, S. *Long Short-Term Memory*; Neural Computation MIT-Press: Cambridge, MA, USA, 1997.
97. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
98. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
99. Booy, M.; Ilin, A.; Orponen, P. RNA secondary structure prediction with convolutional neural networks. *BMC Bioinformatics* **2022**, *23*, 58.
100. Zhang, H.; Zhang, C.; Li, Z.; Li, C.; Wei, X.; Zhang, B.; Liu, Y. A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming. *Front. Genet.* **2019**, *10*, 467.
101. Wang, L.; Liu, Y.; Zhong, X.; Liu, H.; Lu, C.; Li, C.; Zhang, H. Dmfold: A novel method to predict rna secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. *Front. Genet.* **2019**, *10*, 143.
102. Willmott, D.; Murrugarra, D.; Ye, Q. Improving RNA secondary structure prediction via state inference with deep recurrent neural networks. *Comput. Math. Biophys.* **2020**, *8*, 36–50.
103. Chen, C.C.; Chan, Y.M. REDfold: Accurate RNA secondary structure prediction using residual encoder-decoder network. *BMC Bioinform.* **2023**, *24*, 122.
104. Andronescu, M.; Bereg, V.; Hoos, H.H.; Condon, A. RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. *BMC Bioinform.* **2008**, *9*, 340.
105. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G.V.; Christie, C.H.; Dalenberg, K.; Costanzo, L.D.; Duarte, J.M.; et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2020**, *49*, D437–D451.
106. Danaee, P.; Rouches, M.; Wiley, M.; Deng, D.; Huang, L.; Hendrix, D. bprna: large-scale automated annotation and analysis of rna secondary structure. *Nucleic Acids Res.* **2018**, *46*, 5381–5394.
107. Rnacentral 2021: Secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **2021**, *49*, D212–D220.
108. Sweeney, B.A.; Hoksza, D.; Nawrocki, E.P.; Ribas, C.E.; Madeira, F.; Cannone, J.J.; Gutell, R.; Maddala, A.; Meade, C.D.; Williams, L.D.; et al. R2DT is a framework for predicting and visualising RNA secondary structure using templates. *Nat. Commun.* **2021**, *12*, 3494.

109. Jühling, F.; Mörl, M.; Hartmann, R.K.; Sprinzl, M.; Stadler, P.F.; Pütz, J. tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **2009**, *37*, D159–D162.

110. Gutell, R.R. Collection of small subunit (16S-and 16S-like) ribosomal RNA structures. *Nucleic Acids Res.* **1994**, *22*(17), 3502–3507.

111. Zwieb, C.; Gorodkin, J.; Knudsen, B.; Burks, J.; Wower, J. tmRDB (tmRNA database). *Nucleic Acids Res.* **2003**, *31*, 446–447.

112. Richardson, K.E.; Kirkpatrick, C.C.; Znosko, B.M. RNA CoSSMos 2.0: An improved searchable database of secondary structure motifs in RNA three-dimensional structures. *Database J. Biol. Databases Curation* **2020**, *2020*, baz153.

113. Korunes, K.L.; Myers, R.B.; Hardy, R.; Noor, M.A.F. PseudoBase: a genomic visualization and exploration resource for the Drosophila pseudoobscura subgroup. *Fly* **2021**, *15*, 38–44.

114. Nagaswamy, U.; Larios-Sanz, M.; Hury, J.; Collins, S.; Zhang, Z.; Zhao, Q.; Fox, G.E. NCIR: A database of non-canonical interactions in known RNA structures. *Nucleic Acids Res.* **2002**, *30*, 395–397.

115. Sloma, M.F.; Mathews, D.H. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA* **2016**, *22*, 1808–1818.

116. Tan, Z.; Fu, Y.; Sharma, G.; Mathews, D.H. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.* **2017**, *45*, 11570–11581.

117. Kalvari, I.; Nawrocki, E.P.; Ontiveros-Palacios, N.; Argasinska, J.; Lamkiewicz, K.; Marz, M.; Griffiths-Jones, S.; Toffano-Nioche, C.; Gautheret, D.; Weinberg, Z.; et al. Rfam 14: Expanded coverage of metagenomic, viral and microrna families. *Nucleic Acids Res.* **2021**, *49*, D192–D200.

118. Wolfinger, M.T.; Svrcek-Seiler, W.A.; Flamm, C.; Hofacker, I.L.; Stadler, P.F. Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.* **2004**, *37*, 4731.

119. Gruber, A.R.; Findeiß, S.; Washietl, S.; Hofacker, I.L.; Stadler, P.F. RNAz 2.0: Improved noncoding RNA detection. *Pac. Symp. Biocomputing.* **2010**, *2010*, 69–79.

120. Moulton, V. Tracking down noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2269–2270.

121. Lu, Z.J.; Mathews, D.H. Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res.* **2008**, *36*, 640–647.

122. Tafer, H.; Ameres, S.L.; Obernosterer, G.; Gebeshuber, C.A.; Schroeder, R.; Martinez, J.; Hofacker, I.L. The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.* **2008**, *26*, 578–583.

123. Sazani, P.; Gemignani, F.; Kang, S.-H.; Maier, M.A.; Manoharan, M.; Persmark, M.; Bortner, D.; Kole, R. Systemically delivered antisense oligomers upregulate gene expression in mouse tissues. *Nat. Biotechnol.* **2002**, *20*, 1228–1233.

124. Childs-Disney, J.L.; Wu, M.; Pushechnikov, A.; Aminova, O.; Disney, M.D. A small molecule microarray platform to select RNA internal loop-ligand interactions. *ACS Chem. Biol.* **2007**, *2*, 745–754.

125. Palde, P.B.; Ofori, L.O.; Gareiss, P.C.; Lerea, J.; Miller, B.L. Strategies for Recognition of Stem-loop RNA Structures by Synthetic Ligands: Application to the HIV-1 Frameshift Stimulatory Sequence. *J. Med. Chem.* **2010**, *53*, 6018–6027.

126. Castanotto, D.; Rossi, J.J. The promises and pitfalls of RNA-interference-based therapeutics. *Nature* **2009**, *457*, 426–433.

127. Gareiss, P.C.; Sobczak, K.; McNaughton, B.R.; Palde, P.B.; Thornton, C.A.; Miller, B.L. Dynamic Combinatorial Selection of Molecules Capable of Inhibiting the (CUG) Repeat RNA – MBNL1 Interaction in vitro: Discovery of Lead Compounds Targeting Myotonic Dystrophy (DM1). *J. Am. Chem. Soc.* **2008**, *130*, 16254–16261.

128. Rouillard, J.M.; Zuker, M.; Gulari, E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.* **2003**, *31*, 3057–3062.

129. Tavares, R.D.C.A.; Mahadeshwar, G.; Wan, H.; Huston, N.C.; Pyle, A.M. The Global and Local Distribution of RNA Structure throughout the SARS-CoV-2 Genome. *J. Virol.* **2021**, *95*, e02190-20.

130. Vandelli, A.; Monti, M.; Milanetti, E.; Armaos, A.; Rupert, J.; Zacco, E.; Bechara, E.; Ponti, R.D.; Tartaglia, G.G. Structural analysis of SARS-CoV-2 genome and predictions of the human interactome. *Nucleic Acids Res.* **2020**, *48*, 11270–11283.

131. Wang, X.; Gu, R.; Chen, Z.; Li, Y.; Ji, X.; Ke, G.; Wen, H. Uni-Rna: Universal Pre-Trained Models Revolutionize Rna Research. *bioRxiv* **2023**, *2023*, 548588.

132. Chen, J.; Hu, Z.; Sun, S.; Tan, Q.; Wang, Y.; Yu, Q.; Zong, L.; Hong, L.; Xiao, J.; Shen, T.; et al. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions.

arXiv **2022**, arXiv:2204.00300.

133. Akiyama, M.; Sakakibara, Y. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR Genom. Bioinform.* **2022**, *4*, lqac012.
134. Zhang, J.; Fei, Y.; Sun, L.; ; Zhang, Q.C. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat. Methods* **2022**, *19*, 1193–1207.
135. Lyngsø, R.B.; Pedersen, C.N. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* **2000**, *7*, 409–427.
136. Parisien, M.; Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **2008**, *452*, 51–55.