PLOS COMPUTATIONAL BIOLOGY

REVIEW

Review of machine learning methods for RNA secondary structure prediction

Qi Zhao[1☯], Zheng Zhao[2☯], Xiaoya Fan[3], Zhengwei Yuan[4], Qian Mao[5,6], Yudong Yao[7]*

1 College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, Liaoning, China, **2** School of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China, **3** School of Software, Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology, Dalian, Liaoning, China, **4** Key Laboratory of Health Ministry for Congenital Malformation, Shengjing Hospital of China Medical University, Shenyang, Liaoning, China, **5** College of Light Industry, Liaoning University, Shenyang, Liaoning, China, **6** Key Laboratory of Agroproducts Processing Technology, Changchun University, Changchun, Jilin, China, **7** Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, New Jersey, United States of America

☯ These authors contributed equally to this work.
* yyao@stevens.edu

Citation: Zhao Q, Zhao Z, Fan X, Yuan Z, Mao Q, Yao Y (2021) Review of machine learning methods for RNA secondary structure prediction. PLoS Comput Biol 17(8): e1009291. https://doi.org/10.1371/journal.pcbi.1009291

Editor: Shi-Jie Chen, University of Missouri, UNITED STATES

Published: August 26, 2021

Funding: This work was supported in part by the Fundamental Research Funds of Northeastern University (N181903008 - Q.Z.); the Research Start-up Fund for Talent of Dalian Maritime University (02500348 - Z.Z.); the Doctoral Scientific Research Foundation of Liaoning Province of China (2019-BS-108 - Q.M.); the Youth Seedling Project of Educational Department of Liaoning Province of China (LQN202002- Q.M.); the Fundamental Research Funds for the Central Universities (82232019 - X.Y.F.); and the National Natural Science Foundation of China (62002056 - Q.Z., 31801623 - Q.M., 81871219 - Z.W.Y). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Secondary structure plays an important role in determining the function of noncoding RNAs. Hence, identifying RNA secondary structures is of great value to research. Computational prediction is a mainstream approach for predicting RNA secondary structure. Unfortunately, even though new methods have been proposed over the past 40 years, the performance of computational prediction methods has stagnated in the last decade. Recently, with the increasing availability of RNA structure data, new methods based on machine learning (ML) technologies, especially deep learning, have alleviated the issue. In this review, we provide a comprehensive overview of RNA secondary structure prediction methods based on ML technologies and a tabularized summary of the most important methods in this field. The current pending challenges in the field of RNA secondary structure prediction and future trends are also discussed.

Introduction

Since its discovery, for a long time, RNA was regarded solely as a message carrier between DNA and protein. However, we are now beginning to understand its important roles, as increasing numbers of noncoding RNAs (ncRNA) are being discovered [1]. According to the latest report, less than 2% of the human genome comprises protein-coding regions [2]. The majority of the remaining genome portions encode ncRNAs [3], which are involved in translation, catalysis, RNA stability, RNA modification, gene expression regulation, protein synthesis, and protein degradation [4–9]. The enormous importance of ncRNAs in various human diseases, such as cancer, diabetes, and atherosclerosis [6,10], is also being recognized.

ncRNA molecules often fold into higher-order structures, and functionally important ncRNA structures are typically conserved during evolution. Similar to protein, the ncRNA function is usually closely related to its structure. Currently, a wide variety of ncRNA sequences are publicly available, and their numbers keep dramatically increasing [11]. By

Competing interests: The authors have declared that no competing interests exist.

contrast, most of their structures remain unknown, which hinders the inference of their function. Hence, efficient determination of ncRNA structure is of great value to research.

Unlike the global folding of protein driven by hydrophobic forces, the RNA folding process is hierarchical [12] (Fig 1). Specifically, the RNA secondary structure, composed of base pairs, forms rapidly from linear RNA (primary structure), with a large energy loss, while the formation of a complex tertiary structure (or 3D structure) is usually much slower [13]. The RNA secondary structure is very stable and abundant in the cell, which is important for ncRNA function [14,15]. Even without the knowledge of the higher-order structure, RNA secondary structure is sufficient to infer function and for other practical applications [15].

Computational predictions are mainstream approaches for identifying RNA secondary structure. A number of prediction methods have been developed since the 1970s. Most of these methods attempt to identify a structure with a minimum free energy (MFE), in agreement with the hypothesis that an RNA molecule is likely to exist in an MFE state, just like protein [16]. Many prominent software applications have been developed incorporating these methods [17–19]. However, in the last 10 years, the accuracy of prediction failed to significantly improve, and neither did the calculating speed. An alternative approach, the machine learning (ML)-based methodology, was proposed to improve the predictions of RNA secondary structure. However, such methods did not receive much attention because of the limited accuracy. That was mainly because of the small size of the training datasets and the limitations of simple ML models. As a result of the recent explosion of RNA sequence data and the improvement of ML techniques, especially deep learning (DL) techniques, the latest ML-based methods supersede the current mainstream methods in terms of accuracy and applicability. We believe that these ML-based methods will inspire the next generation of prediction tools in the near future.

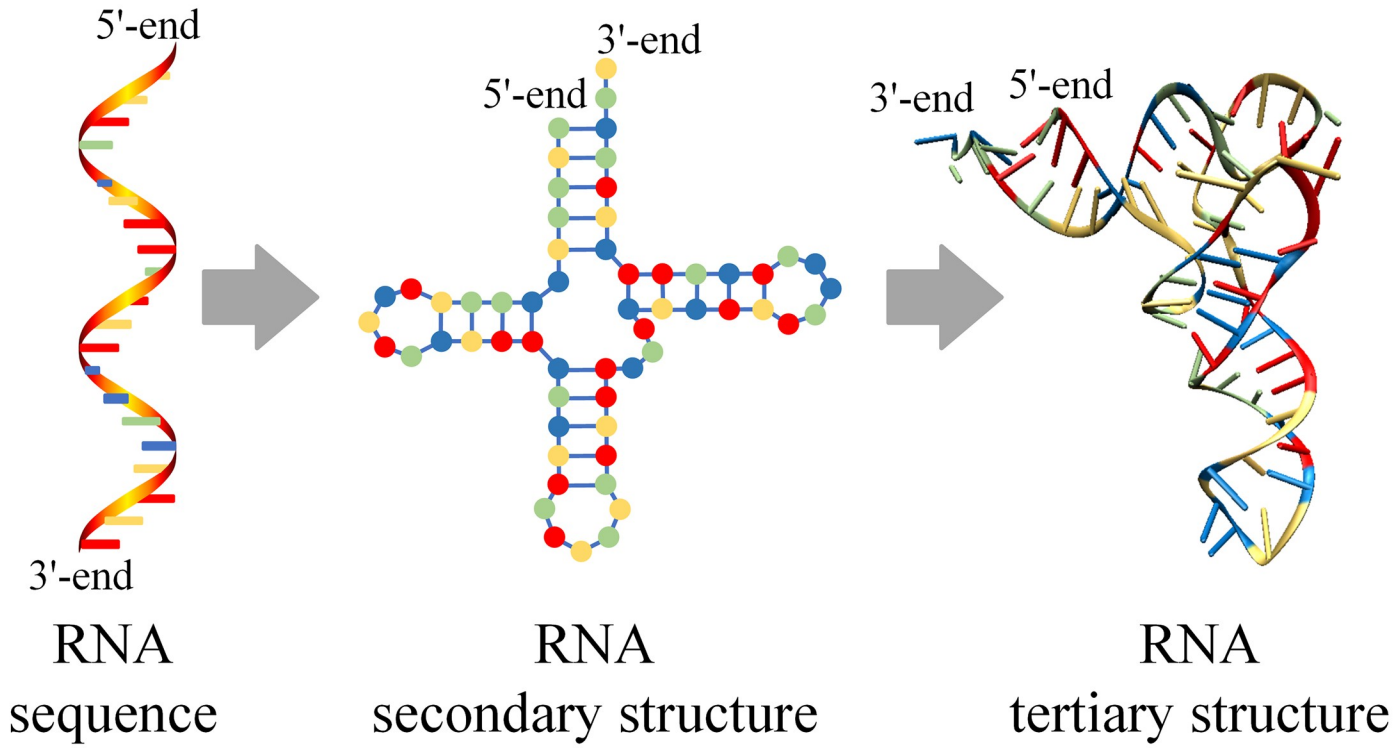

Fig 1. RNA primary (left), secondary (middle), and tertiary structures (right). The RNA folding process is hierarchical, i.e., the RNA secondary structure forms rapidly from linear RNA (primary structure) with a large energy loss, while the formation of a complex tertiary structure is usually much slower.

https://doi.org/10.1371/journal.pcbi.1009291.g001

In this paper, we provide a comprehensive overview of ML-based methods for RNA secondary structure prediction, with a thorough discussion of their advantages and disadvantages. We also provide a tabularized summary (Table 1) of the most important models in the field, and a perspective on the future promising directions, with a special emphasis on DL models. Although several review papers have been published on the topic of RNA secondary structure prediction [20–22], reviews with an emphasis on ML techniques are lacking. We believe that this review will enable researchers to understand the key issues that remain to be solved and facilitate further advances in predicting the RNA secondary structures based on ML.

RNA secondary structure: The basics

The RNA molecule is an ordered sequence of nucleotides that contain 1 of the 4 bases: adenine (A), cytosine (C), guanine (G), and uracil (U), arranged in the 5′ to 3′ direction. Pairing (via hydrogen bonds) of these 4 bases within an RNA molecule gives rise to the secondary structure. Typically, each base pairs with at most one other base. The canonical base pairs include the Watson–Crick base pairs (A–U and G–C) and the wobble base pair (G–U). These base pairs often result in the formation of a nested structure, in which multiple stacked base pairs form a helix, and one or multiple unpaired base pairs form a loop.

It has to be noted that 3 kinds of special base pairs [23] commonly occur in the native RNA secondary structures, i.e., noncanonical base pairs, base triples, and pseudoknots. Noncanonical base pairs are the base pairs other than A–U, G–C, and G–U, and they make up 40% of all base pairs in structured RNAs [24]. Base triples are the cluster of 3 bases interacting, which [25] can stabilize many RNA tertiary interactions [26]. Base triples also occur widely in RNA structures. A pseudoknot [27] occurs when bases in different loops pair with each other, forming a nonnested structure between 2 bases that are located apart from each other. Pseudoknots represent a small fraction of base pairs in known RNA secondary structures but often play an important role in RNA function [28].

Typically, the secondary structure of an RNA molecule with a length n can be regarded as:

1) A set of base pairs $\{(i,j),1\leq i<j\leq n\}$, where (i,j) indicates a base pair formed between the i-th and j-th nucleotide in the RNA sequence; or a set of compatible helixes [28].
2) A contact table (CT table), i.e., a square matrix, with elements in the i-th row and j-th column representing the interaction between the i-th and j-th nucleotides in the RNA sequence.
3) A graph, where nodes represent nucleotides and edges represent base pairing relationships.
4) A labeled sequence with the length n, e.g., "dot-parenthesis" notation, with matching parentheses for paired bases and dots for unpaired bases.
5) A parse tree derived from context-free grammars, of which the leaf nodes comprise the RNA sequence [29].

The above definitions form the basis of both traditional and ML-based RNA secondary structure prediction methods.

Traditional methods of detecting or predicting RNA secondary structure

RNA structure determination is a fast-evolving topic. Many different methods have emerged in the last 20 years. They can be divided into 2 categories, i.e., wet lab experimental approaches and computational predicting approaches.

Table 1. Summary of the ML-based RNA secondary structure prediction methods.

Category		Title	Date	Author	ML Technique	Resource	Reference
Score scheme based on ML model	Free energy parameter-refining approach based on ML	Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs	1998	Xia et al.	Linear regression	Table 1 in the paper (https://pubs.acs.org/doi/10.1021/bi9809425#)	[56]
		Efficient parameter estimation for RNA secondary structure prediction	2007	Andronescu et al.	Constraint generation	http://www.rnasoft.ca/CG/	[73]
		Computational approaches for RNA energy parameter estimation	2010	Andronescu et al.	Loss-augmented max-margin constraint generation model, Boltzmann-likelihood model	http://www.cs.ubc.ca/labs/beta/Projects/RNA-Params	[74]
	Weighted approach based on ML	Rich Parameterization Improves RNA Structure Prediction	2011	Zakov et al.	Discriminative structured-prediction learning framework combined, online learning algorithm	http://www.cs.bgu.ac.il/?negevcb/contextfold	[77]
		A Max-Margin Training of RNA Secondary Structure Prediction Integrated with the Thermodynamic Model	2018	Akiyama et al.	SSVM	https://github.com/keio-bioinformatics/mxfold	[78]
		RNA secondary structure prediction using deep learning with thermodynamic integration	2021	Sato et al.	Deep neural network	http://www.dna.bio.keio.ac.jp/mxfold2/	[79]
	Probabilistic approach based on ML	Stochastic context-free grammars for tRNA modeling	1994	Sakakibara et al.	EM method	-	[29]
		RNA secondary structure prediction using stochastic context-free grammars and evolutionary history	1999	Knudsen and Hein	EM method	-	[82]
		Pfold: RNA secondary structure prediction using stochastic context-free grammars	2003	Knudsen and Hein	EM method		[81]
		CONTRAfold: RNA secondary structure prediction without physics-based models	2006	Do et al.	CLLM	http://contra.stanford.edu/contrafold/	[86]
		A semi-supervised learning approach for RNA secondary structure prediction	2015	Yonemoto et al.	Semi-supervised learning algorithm	-	[87]
Preprocessing and postprocessing based on ML model	Preprocessing based on ML model	A tool preference choice method for RNA secondary structure prediction by SVM with statistical tests	2013	Hor et al.	SVM	-	[88]
		Research on folding diversity in statistical learning methods for RNA secondary structure prediction	2018	Zhu et al.	Statistical context-free grammar model	-	[89]
	Postprocessing based on ML model	Using a neural network to identify secondary RNA structures quantified by graphical invariants	2008	Haynes et al.	MLP	-	[90]
		A predictive model for secondary RNA structure using graph theory and a neural network	2010	Koessler et al.	MLP	-	[91]
Predicting process based on ML model	End-to-end approach	Parallel algorithms for finding a near-maximum independent set of a circle graph	1990	Takefuji et al.	System composed of several interactional neurons	-	[92]
		An Hopfield Neural Network-Based Algorithm for RNA Secondary Structure Prediction	2006	Liu et al.	Hopfield networks	-	[93]
		Secondary Structure Prediction of RNA using Machine Learning Method	2011	Qasim et al.	MLP	-	[96]
		Neural Networks, Adaptive Optimization, and RNA Secondary Structure Prediction	1993	Steeg	MFT network	-	[94]
		RNA secondary structure prediction by MFT neural networks	2003	Apolloni et al.	MFT network with mean field approximation to update network's nodes	-	[139]
		RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning	2019	Singh et al.	Compound deep neural networks, transfer learning	https://sparks-lab.org/server/spot-rna/	[97]
		RNA secondary structure prediction by learning unrolled algorithms	2020	Chen et al.	Compound deep neural networks	https://github.com/ml4bio/e2efold	[99]
		Machine learning a model for RNA structure prediction	2020	Calonaci et al.	CNN, MLP	-	[100]
	Hybrid approach	RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers	2006	Bindewald et al.	Hierarchical network of k-nearest neighbor model	-	[49]
		Developing parallel ant colonies filtered by deep learned constrains for predicting RNA secondary structure with pseudo-knots	2020	Quan et al.	Bi-LSTM	-	[103]
		RNA Secondary Structure Prediction Based on Long Short-Term Memory Model	2018	Wu et al.	Bi-LSTM	-	[102]
		Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter	2019	Lu et al.	Bi-LSTM	-	[101]
		A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming	2019	Zhang et al.	CNN	-	[104]
		DMfold: A Novel Method to Predict RNA Secondary Structure with Pseudoknots Based on Deep Learning and Improved Base Pair Maximization Principle	2019	Wang et al.	Bi-LSTM	https://github.com/linyuwangPHD/RNA-Secondary-Structure-Database.	[105]
		Improving RNA secondary structure prediction via state inference with deep recurrent neural networks	2020	Willmott et al.	Bi-LSTM	https://github.com/dwillmott/rna-state-inf	[107]

"-"indicates "not available."

CLLM, conditional log-linear model; CNN, convolutional neural network; EM, expectation-maximization; MFT, mean field theory; ML, machine learning; MLP, multilayer perceptron; SSVM, structured support vector machine; SVM, support vector machine.

https://doi.org/10.1371/journal.pcbi.1009291.t001

Wet lab experiments

X-ray crystallography [30] and nuclear magnetic resonance (NMR) [31] are the most accurate approaches for determining RNA structures, both of which can offer structural information at

a single base pair resolution. However, both methods are characterized by high experimental cost and low throughput, limiting their wide usage. In addition, RNA molecules are highly unstable and difficult to crystallize. Although many methods have been developed to infer the state of nucleotides (paired or unpaired) in an RNA molecule using enzymatic [32,33] or chemical probes [34,35] coupled with next-generation sequencing [36,37], most of them can only be used to capture the RNA secondary structure in vitro. The obtained structure may differ markedly from the in vivo conformation. In fact, to date, the structure of only a very small percentage (<0.001%) of known ncRNAs has been determined experimentally [38]. Hence, predicting the RNA secondary structure using computational methods is an important alternative to wet lab–based approaches.

Traditional computational methods

Comparative sequence analysis [39,40] is the most accurate computational method for determining the RNA secondary structure. This method is based on the assumption that the RNA secondary structure is evolutionarily conserved to a greater extent than the RNA sequence. This method usually finds the base pairs that covary to maintain Watson–Crick and wobble base pairs (compensatory mutations) [41] of a given sequence using a set of homologous sequences. Han and Kim [42] designed the first comparative sequence analysis algorithm based on the phylogenetic comparative analysis. This algorithm predicts a common secondary structure conserved in the given homologous sequence set with a high time complexity ($O(n^3)$, n being the RNA sequence length). To reduce the running time, Tahi and colleagues [43] implemented another algorithm DCfold with time complexity $O(n^2 \log n)$. DCfold searches for helices based on their lengths and mutation rates using a "divide and conquer" approach. Comparative sequence analysis can also predict the structures with pseudoknots [44–46]; however, the accuracy is very limited. In addition, comparative sequence analysis can be combined with score-based methods [47–50], e.g., RNAalifold [48], KnetFold [49], and ILM [47]. One great limitation of this method is that it requires a large set of homologous sequences. However, only thousands of RNA families are currently known [51], which makes it impossible to obtain homologous sequences for all RNAs. Therefore, most methods for RNA secondary structure prediction are score based, where only a single RNA sequence is required as the input.

Score-based methods are the most widely used methods and have dominated the field of RNA secondary structure prediction in the last 4 decades. These methods assume that the native RNA structure is a structure with a minimum/maximum total score, depending on the hypothesis of RNA folding mechanism or its simplification. Hence, the problem of RNA secondary structure prediction is transformed into an optimization problem. Since the RNA secondary structure can be recursively broken down into smaller elements with independent score contributions, the dynamic programming (DP) algorithm is often employed to identify the optimal structure. Evaluation of the score for structure elements requires a score scheme of many parameters. Nussinov and Jacobson [52] proposed the first, and also the simplest, DP algorithm for finding the maximum-matching structure. The authors proposed to assign one point to each matched base pair and assumed that the native structure is the structure with the maximum score among all the possible conformations. Zuker and Stiegler [53] proposed a more realistic scoring scheme based on free energy, the nearest neighbor model (NN model) [54–57]. It is based on Tinoco's hypothesis (see Section 4.1) [58]. The NN model can be used for the calculation of energy changes of any structure of a given RNA molecule, and the DP algorithm can be also employed to efficiently find the MFE structure. A number of slightly different variations of this method were also proposed [59–62]. For predicting the structure with

noncanonical base pairs, some other score schemes were employed as scoring functions, such as nucleotide cyclic motifs score system [63–65] or equilibrium partition function [66]. In addition, several score-based methods were developed to predict RNA secondary structures with pseudoknots [67–71], where the structure search scope or input RNA length is limited or the types of pseudoknots are restricted to lower the time complexity in general.

However, the folding mechanism hypotheses of score-based methods do not always hold, e.g., the RNA molecule often folds into locally stable structural domains. Furthermore, almost all score-based methods use virtually the same DP algorithm to find the best-scoring structures. However, the running time of the DP algorithm is usually $O(n^3)$ (where n is the RNA sequence length), neglecting the special base pairs and weak interactions. Hence, the computational cost is not acceptable, especially when analyzing an RNA molecule longer than 1,000 nucleotides. Moreover, predicting the special base pairs in RNA structures is still a difficult task. Since an RNA structure with special bases pairs is not a nested structure in general, score-based methods have to employ sophisticated algorithms to capture these special base pairs at the cost of higher time complexity. However, the performance of these methods needs to be further improved.

In fact, it is extremely difficult to fully understand the RNA folding mechanism. ML methods, in contrast, are data driven and requiring no knowledge of such mechanism. These methods can learn the underlying folding patterns from large amount of training data. In the last few decades, ML methods have been used for many aspects of RNA secondary structure prediction methods to improve the prediction performance (see Section 4). However, they did not replace the mainstream score-based methods with respect to accuracy and generalization. This situation has been changing in the last 2 years because of the development of ML techniques, especially DL.

ML-based methods

The ML-based methods for RNA secondary structure prediction can generally be divided into 3 categories (S1 Fig) according to the subprocess that ML participates in, i.e., score scheme based on ML, preprocessing and postprocessing based on ML, and prediction process based on ML. All the ML-based methods in these 3 categories trained their models in a supervised way [72]. These models learn functions that map inputs (features) to outputs by adjusting model parameters based on the known input–output pairs. Many of them employ free energy parameters, encoded RNA sequences, sequence patterns, or evolutionary information as key features, and their outputs can be classification labels (such as paired or unpaired) or continuous values (such as free energy). When a new input is fed to the trained model, the model can classify a corresponding label or predict a corresponding value [72].

Score scheme based on ML

Early ML-based methods usually train an ML model that can generate a new score scheme (Fig 2) and replace the score scheme used in the traditional methods. According to the meaning of the score, ML-based score schemes can be further divided into 3 categories (S1 Fig), i.e., the free energy parameter-refining approach, weighted approach, and probabilistic approach. Although ML-based methods are used for parameter estimation in the score schemes to improve the prediction accuracy, the structure prediction is still formulated as an optimization problem, where the estimated parameters are used for the evaluation of the scores of possible conformations.

Free energy parameter refining based on ML. Considering the score schemes, the free energy–focused approach is the most popular approach. Ever since Tinoco and colleagues [58]

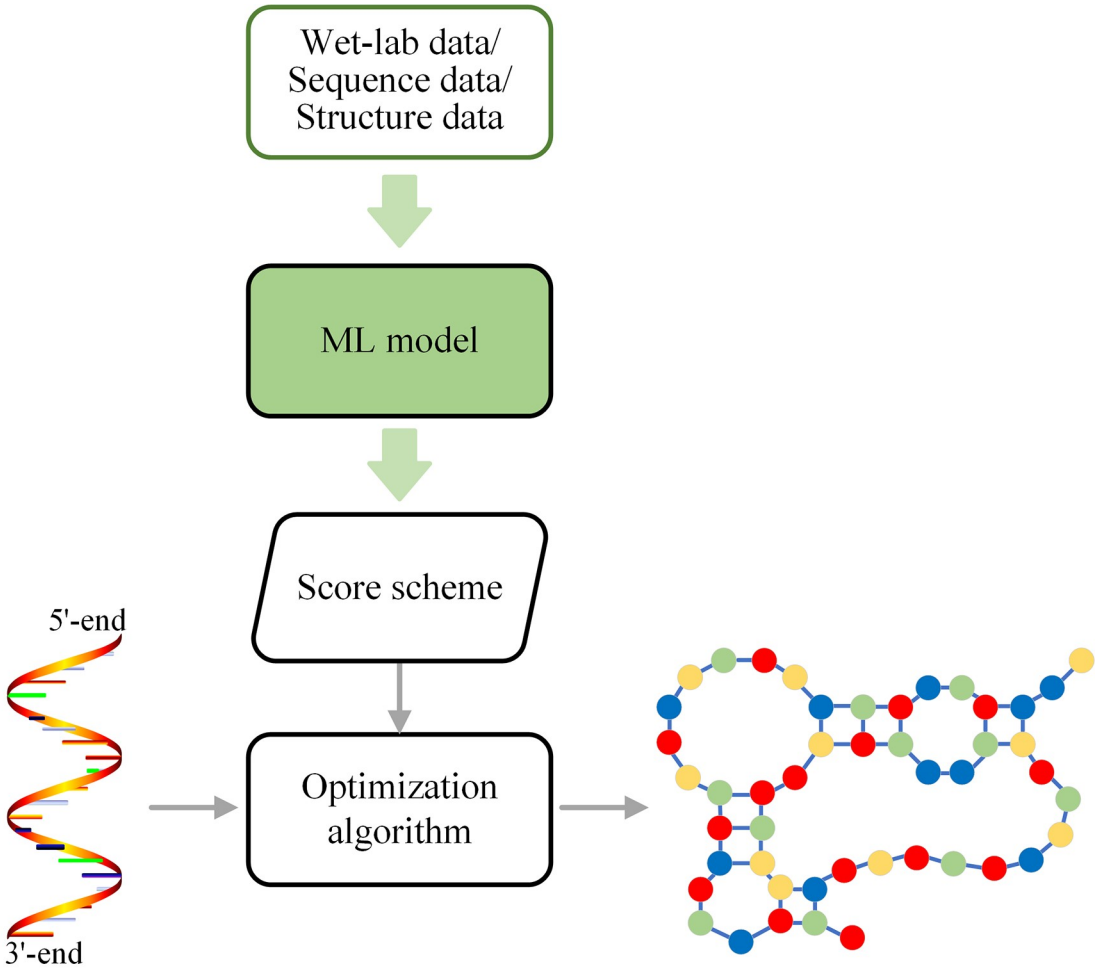

Fig 2. Framework for RNA secondary structure prediction methods with ML-based score schemes. Wet lab data, RNA sequence data, or RNA structure data can be employed to train an ML model to obtain a score scheme. Using this score scheme, an RNA secondary structure can be predicted using a traditional score-based approach from a single RNA sequence.

https://doi.org/10.1371/journal.pcbi.1009291.g002

put forward their hypothesis for free energy calculation (that the free energy of a secondary structure is the sum of the free energy values of its elements), many studies have been devoting their efforts to the problem of assigning free energy values to the elements of RNA molecules. Turner's NN model [57] is the most popular approach and provides a considerably accurate approximation of the RNA free energy. However, the multiple thermodynamic parameters of the NN model have to be based on a large number of optimal melting experiments. These experiments are time and labor consuming [17,19], however, and not all free energy changes in structural elements can be measured because of the associated technical difficulties.

Some ML techniques were adopted to refine the parameters in the energy model. These techniques can employ subtle models to estimate the scores for a richer and more accurate feature representation using known thermodynamic data or RNA secondary structure data. Xia and colleagues [56] first trained a linear regression model using known thermodynamic data to infer some of the thermodynamic parameters and expanded the NN model into a more accurate model, i.e., the INN-HB model. This model provides an improved fit for the known experimental data. A disadvantage of this approach, however, is that the parameters for some

structural elements are fixed before other parameters are calculated, which limits the range of possibilities considered for the overall parameter set. To overcome this problem, Andronescu and colleagues [73] proposed a constraint generation approach to estimate free energy parameters. This method uses different types of constraints to ensure that the energies of reference structures are low relative to the alternatives for the same sequence. Trained on large sets of structural and thermodynamic data, this method achieves 7% higher F-measure than the standard Turner parameters. Subsequently, the authors further modified the method and proposed a loss-augmented max-margin constraint generation model and Boltzmann-likelihood model using a larger dataset [74]. The constraints imposed on parameters ensure that the more inaccurate the structure, the greater the margin between its free energy and that of the reference structure in the training set.

Of note, the parameters determined by the above free energy parameter-refining approaches are thermodynamic and can be used directly in the algorithms embedded by the same energy model, such as miRNA target prediction [75] and RNA folding kinetics simulation [76].

Weighted approaches based on ML. While ML-based free energy parameter approaches successfully improved the accuracy of the RNA secondary structure prediction, the energy model is still far from ideal. Actually, the above methods for the estimation of ML-based parameters can only substitute for some wet lab experiments geared toward obtaining the energy parameters. However, it is entirely possible to obtain an improved score scheme independent of free energy based on ML techniques. Several weighted approaches were proposed that consider the parameters of RNA structure elements as weights instead of free energy changes. By removing the thermodynamic meaning, the weighted approach can utilize ML models to determine thousands of weights for more comprehensive RNA structure elements instead of obtaining a few energy parameters from a large number of wet lab experiments.

By combining a discriminative structured-prediction learning framework with an online learning algorithm, Zakov and colleagues [77] greatly increased the number of weights to approximately 70,000 by examining more types of structural elements with more numerous sequential contexts, using thousands of training datasets. Based on these weights, the ContextFold tool was proposed, marking a significant improvement in the prediction accuracy [77]. Akiyama and colleagues [78] integrated the thermodynamic approach with a structured support vector machine (SSVM) to obtain a large number of weights for detailed structure elements, of which $l1$ regularization was used to relieve overfitting. Then, MXfold was built by combining ML-based weights with experimentally determined thermodynamic parameters, achieving better performance than a model based on thermodynamic parameters or ML-based weights alone. Most recently, MXfold2 [79] was proposed by Sato and colleagues They trained a fairly deep neural network using the max-margin framework with thermodynamic regularization, which made the folding scores predicted by MXfold2 and the free energy calculated by the thermodynamic parameters were as close as possible. This method showed a robust prediction on both sequence-wise and family-wise cross-validation. These studies suggest that ML-based weights can complement the gaps in the thermodynamic parameter approach.

An advantage of the weighted approach is that it decouples structure prediction from energy estimation, which is potentially beneficial for both tasks. However, learned weights are not explainable, partly because of the black-box attribute of ML algorithms. Hence, the obtained scores cannot be used to compute the partition function, base pair binding probabilities, or centroid structures, etc.

Probabilistic approaches based on ML. Stochastic context-free grammars (SCFGs) are an important alternative for predicting RNA structures [29,80–84]. SCFGs specify formal grammar rules and induce a joint probability distribution over possible RNA structures for a

given sequence. In particular, an SCFG model specifies a probability parameter for each production rule in the grammar and thus assigns a probability to each sequence it derives. The probability parameters are learned from datasets of RNA sequences annotated using known secondary structures, without the need for external laboratory experiments [83].

Sakakibara and colleagues [29] first applied SCFGs to tRNA secondary structure prediction. The probability parameters in their SCFG model were learned using an expectation–maximization (EM) method. Knudsen and Hein [82] improved the SCFG model by combining the evolutionary information, and, subsequently, the robust and practical tool Pfold [81] was developed. Sato and colleagues [85] proposed a nonparametric Bayesian extension of SCFGs with the hierarchical Dirichlet process to find an optimal RNA grammar from the training dataset. Using another ML model, the conditional log-linear model (CLLM), Do and colleagues [86] identified probability parameters that are most likely to discriminate correct structures from incorrect ones. CLLM is a flexible probabilistic ML model that generalizes upon SCFGs; the parameters are easily estimated, and arbitrary features can be incorporated in the model. CONTRAfold has achieved the highest single-sequence prediction accuracy to date, compared with the currently available probabilistic models. However, CLLM is very slow, which prevents its application to large training sets, and the estimated parameters have no intrinsic biological meaning. Finally, to take full advantage of the substantial numbers of RNA sequences with unknown structures, Yonemoto and colleagues [87] proposed a semi-supervised learning algorithm to obtain probability parameters in a probabilistic model that combines SCFG and a conditional random field.

However, the probabilistic approach cannot replace MFE methods for secondary structure prediction, as the accuracy of the currently best SCFG has yet to match those of the best free energy–based models. In addition, SCFG cannot describe all RNA structures, e.g., a structure containing special base pairs.

Preprocessing and postprocessing based on ML

ML can be also used in pretreatment, for selecting an appropriate prediction method or a group of appropriate parameters (Figs 3 and S1). A tool based on a support vector machine (SVM) was

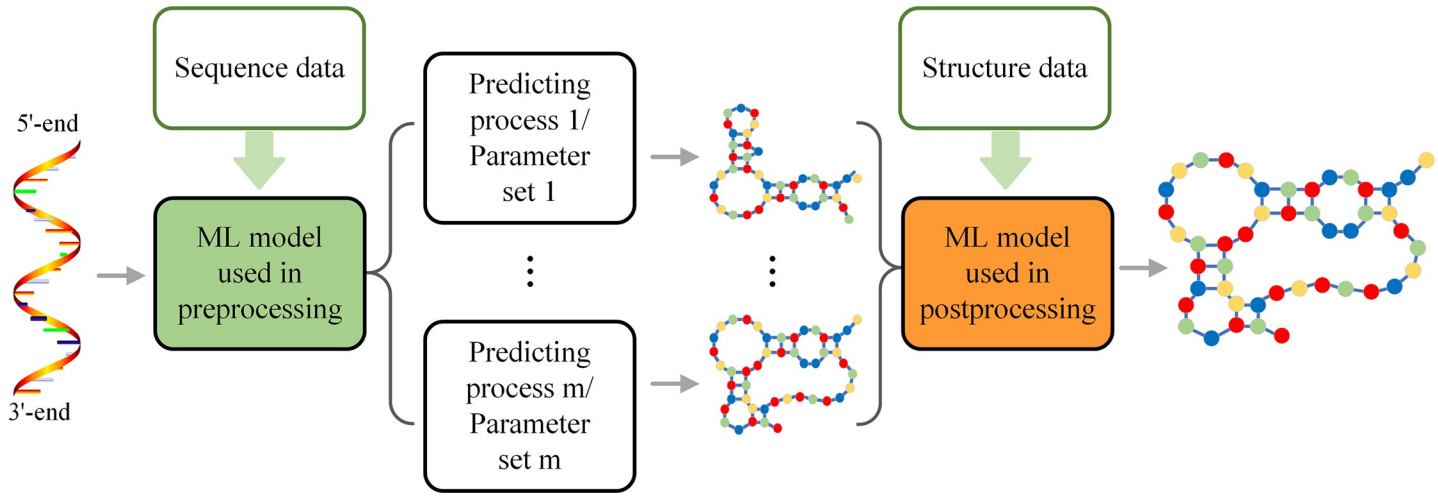

Fig 3. Framework for RNA secondary structure prediction methods with ML-based preprocessing or postprocessing. In RNA secondary structure prediction, ML models (trained by sequence data, in green) can be also used in pretreatment for selecting an appropriate prediction method or a group of appropriate parameters; ML models (trained by structure data, in brown) also can provide a means of determining the most likely structures among the outcomes.

https://doi.org/10.1371/journal.pcbi.1009291.g003

proposed by Hor and colleagues [88] for selecting the prediction method, based on the notion that different RNA sequences have different features and different prediction methods work best with specific RNA species. In another study, Zhu and colleagues [89] assumed that different RNA sequences follow different folding rules. The authors consequently proposed an SCFG model to identify the most probable folding rules before RNA secondary structure prediction.

Since different prediction methods return several different structures, the ML model can provide a means of determining the most likely structures among the outcomes (Figs 3 and S1). Combined with the graph theory, Haynes and colleagues [90] used trees to represent RNA graphical structures (edges as helices, and verticals as loops or bulges). They then trained a multilayer perceptron (MLP) model to distinguish whether a structure is RNA-like or not, using graphical invariants as input features. Assuming that a larger secondary structure is formed upon bonding of 2 smaller secondary RNA structures, Koessler and colleagues [91] also used an MLP model to predict the RNA-like probability of a structure using a special feature vector extracted from the merged trees.

Predicting process based on ML

ML techniques are also directly used to predict RNA secondary structure in an end-to-end fashion or combined with other algorithms as constraints, base state detector, or structure selector. The general framework is shown in Figs 4 and S1.

End-to-end approach. To the best of our knowledge, the ML technique was first introduced into the RNA secondary structure predicting process by Takefuji and colleagues [92]. The authors built on Nussinov and Jacobson's hypothesis (see Section 3.2) [52] and attempted to obtain a near-maximum independent set (MIS) from an adjacent graph (where the vertices represent base pairs, and the edges connect the incompatible vertices) using a system composed of m interactional neurons (m is the number of edges). Liu and colleagues [93] enhanced Takefuji's work by considering the energy contribution of possible base pairs, and a Hopfield neural network (HNN) was employed to obtain MIS. However, HNN is easily trapped in local minima, limiting the accuracy of this method. To avoid this problem, Steeg and Evan [94]

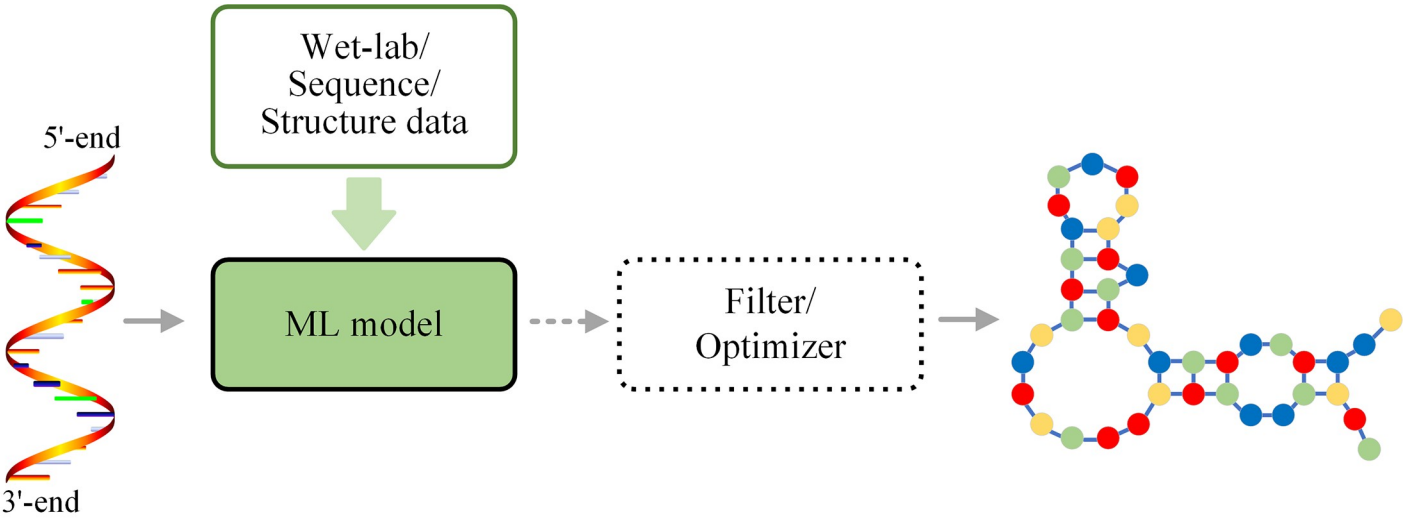

Fig 4. Framework for the RNA secondary structure prediction methods with ML-based prediction process. ML models (trained by wet lab, RNA sequence, or RNA structure data) are directly used to predict RNA secondary structures in an end-to-end way or followed by a filter or optimizer to obtain the optimal RNA secondary structure.

https://doi.org/10.1371/journal.pcbi.1009291.g004

made use of the mean field theory (MFT) networks to identify the optimal structure, which was coupled with a sophisticated objective function with additional biological constraints. The inputs into the MFT networks are the 4 types of bases in an RNA sequence encoded in a one-hot fashion, and the output is in a format similar to CT table. Subsequently, Apolloni and colleagues [95] further developed Steeg's method, especially with respect to the computation speed, so that it could be applied to slightly longer RNA sequences. In addition, this model uses mean field approximation to update the node in both the learning phase and the instant resolution phase. In another study, Qasim and colleagues [96] modified Takefuji's work by building a novel MLP model to obtain MIS. This model contains h neurons in the hidden layer, whose activation function is based on the Kolgomorov's theorem (h is the number of possible base pairs in an RNA sequence).

However, because of the relatively poor performance of the above ML models and a small amount of the available data, ML-based RNA secondary structure prediction models can only process tRNAs, with relatively low accuracy. Currently, the use of DL techniques is rising rapidly, and they are dramatically changing these circumstances. Singh and colleagues [97] proposed the first end-to-end DL model, SPOT-RNA, to predict RNA secondary structure. SPOT-RNA treats the RNA secondary structure as a CT table and employs an ensemble of ultradeep hybrid networks of ResNets and 2D-BLSTMs for the prediction. Of these, the former captures the contextual information from the whole sequence, and the latter is effective for the propagation of long-range sequence dependencies in RNA structure. Transfer learning is used to train SPOT-RNA to effectively utilize limited sample numbers. SPOT-RNA showed superior performance with several RNA benchmark datasets, greatly outperforming the best score-based methods and SCFG-based methods. Recently, the SPOT-RNA2 model [98] was proposed by the same research group. This model employed evolution-derived sequence profiles and mutational coupling as inputs and outperformed SPOT-RNA for all types of base pairs using the same transfer learning approach. E2Efold is another DL model for RNA secondary structure prediction, proposed by Chen and colleagues [99]. It integrates 2 coupled parts, i.e., a transformer-based deep model that encodes sequence information, and a multilayer network based on an unrolled algorithm that gradually enforces the constraints and restricts the output space.

In addition to the encoded RNA sequences being used as the input, other information can also be incorporated into the DL model. Calonaci and colleagues [100] trained an ensemble model based on a combination of SHAPE data, co-evolutionary data (DCA), and RNA sequence data. Their model consists of a convolutional neural network (CNN) subnetwork and an MLP subnetwork to predict penalties based on SHAPE and DCA data, respectively, with an RNAfold [17] module to generate structures using RNA sequences and penalties.

Hybrid approach. Alternatively, ML can be combined with other methods for a hybrid approach for RNA secondary structure prediction. Consequently, the ML model is usually considered as a scoring machine, mapping a score to each (pair of) base(s) in an RNA sequence, whose output is then passed to an independent filter to identify a reasonable structure.

Bindewald and Shapiro [49] combined an ML model and a filter to predict the consensus structure for a group of aligned RNAs. The authors chose a hierarchical network of k-nearest neighbor model to predict the possibility score for each pair of alignment columns and defined the filter by a set of rules derived from native RNA structures. Considering structure prediction as a sequence-labeling question, Lu and colleagues [101] and Wu and colleagues [102] employed a more powerful DL model, Bi-LSTM, to predict the state of each base in an RNA sequence, using a similar rule-based filter to deal with conflicting pairing. Differently from the above studies, Bi-LSTM was used as a structure filter in DpacoRNA [103], and a parallel ant

colony optimization method was used to predict the most probable structures. Another type of an ML-based hybrid approach combines ML models and optimization methods. Liu's group [104] used a CNN model to predict the status distribution of each base in an RNA sequence, and a DP algorithm was employed to find the maximum probability structure. The same group [105] also used the Bi-LSTM model instead and another optimization algorithm, similar to that used in [106]. Instead of developing a new optimizer, Willmott and colleagues [107] utilized an existing SHAPE-directed method (SDM) [108] as the optimizer, which can predict optimal structure from SHAPE data, and trained a Bi-LSTM model to generate SHAPE-like data (i.e., determine the state of each nucleotide) of an RNA sequence as the inputs of SDM.

Compared with the end-to-end approach, the performance of the hybrid approach is relatively poor, perhaps because of a bias between the training objective of the ML part and the overall system objective. Most methods in the hybrid approach are trained and tested using small-scale datasets. Hence, generalization of their abilities requires further verification.

Discussions

It is well known that transcript abundance helps to identify transcripts of interest under different conditions, while the RNA structure helps to explain how these transcripts function. An excellent RNA structure prediction method is not only important for inferring RNA function, but also relates to many downstream studies, including ncRNA detection [109–111], folding dynamics simulations [112], hybridization stability assessment [113], and oligonucleotide [114,115] or drug design [116–120]. It is worth noting that RNA secondary structure prediction is also a useful tool for studying viruses, such as the SARS-CoV-2 virus responsible for the current pandemic [121,122].

The advantages of ML-based methods

Compared with comparative sequence analysis and traditional score-based methods, ML-based methods have some advantages. First, ML-based methods do not necessarily rely on the biological mechanism, which is usually difficult to thoroughly understand. Instead, they can utilize the information contained in various types of data, and, therefore, performance limitation caused by the mechanism hypothesis can be circumvented. ML-based methods can also be easily coupled with known biological mechanisms. Further, in terms of prediction performance, where a large amount of data is available, models with no or little knowledge of biological mechanisms usually perform better than mechanism-dependent ones. This also suggests that the assumed mechanism of RNA folding may be incomplete or not accurate. Second, in contrast to traditional score-based methods, the end-to-end DL methods do not need to consider the difficulties caused by base matching rules. Traditional score-based methods employ sophisticated algorithms to satisfy base matching rules at the cost of high time complexity. However, without the constraint of these rules, end-to-end models [97] can train and predict all the base pairs in RNA structures, regardless of whether the base pairs associate with secondary or tertiary interactions. Third, compared with traditional methods, the ML-based methods can be considerably flexible. The inputs of ML-based models can be either one-dimensional or multidimensional, extracted features or encoded bases, and homogeneous data or heterogeneous data, and the outputs can be CT tables, labeled sequences, nucleotide states, or free energy values. In addition, the construction of the ML models is diverse, from simple Hopfield networks to complex ensemble deep neural networks. Fourth, once the model training is completed, the ML-based end-to-end prediction methods run very fast. Unlike DP algorithm, the time complexity of ML models is independent of the input scale, which is advantageous when dealing with long RNAs.

Datasets and their impacts on ML-based methods

Today, many public RNA structure databases and other related datasets are available online, which provide abundant data for model training. Generally, these databases can be classified into 2 types, i.e., comprehensive databases and dedicated databases. A comprehensive database often consists of RNA structures with different conformations and in different RNA species, for example, RNA Strand (4,666 RNAs available) [123], RCSB Protein Data Bank (PDB, 4,962 RNAs available) [124], and bpRNA-1m (102,348 RNAs available) [125]. Some of these databases (e.g., PDB) collect tertiary structures obtained by wet lab experiments, while others obtained data using comparative sequence analysis method (less accurate than those obtained by wet lab experiments, e.g., pbRNA-1m). Dedicated databases generally involve only a single RNA species (tRNA [126], rRNA [127], or tmRNA databases [128]) or a single type of RNA structure (such as loop [129], pseudoknot [130], or noncanonical base pair [131]) generally. Based on these dedicated databases, some public benchmark datasets were established, such as ArchiveII [132] and RNAStralign [133]. These datasets are generally composed of tens of thousands of RNAs in different RNA species (rRNA, tRNA, SRP, tmRNA, etc.). In addition, other databases used in ML-based methods are Rfam [51] and NNDB [57], which provide RNA family information and thermodynamic parameters, respectively.

Data are extremely important for building ML-based RNA secondary prediction models, especially DL-based models with a large number of parameters. One of the reasons that the recent DL-based methods [79,97,99] outperform the traditional ML-based models is the improvement of the quality and quantity of the training sets. It is worth noting that the performance of DL-based methods may be overestimated due to the data similarity between the training and test set. Most of studies only ensured that the RNAs in test sets of these methods were not so similar (80% similarity [134] as a cutoff typically) to those in the training sets, but RNAs from the same families were not explicitly excluded from the testing set. The sequences and structures in the same RNA family are similar, resulting that the model performance obtained on testing sets is better than reality [79,97].

Another issue that may affect the model performance is the imbalanced RNA families in training sets, e.g., thousands of 16S rRNAs but only a small number of telomerases occur in one dataset. When the length of the input RNA is comparable, trained models tend to perform better on the RNA species that are more prevalence in the training set [99]. How to deal with unbalanced data is an active topic in the ML community. Study [99] adopted an up-sampling strategy to balance the RNAs in different families, and their model performance was further improved.

Generally, the enhancement of predictive ability is associated with the relatively large scale of the ML model, which requires large amounts of data for parameter training. Although a large number of RNA structure data in various formats is available, these are insufficient in terms of training large-scale DL models, especially with respect to the availability of high-accuracy data. Hence, questions on how to effectively utilize the limited data and cope with overfitting of a large-scale DL model are also important issues that remain to be resolved.

Current pending challenges

Enormous progress has been made toward predicting RNA secondary structure by using ML-based methods. These methods are state of the art when considering most indices of prediction performance. However, some issues still require resolving.

First, the accuracy of prediction should be improved further. Sato and colleagues [79] used the RNAs in the newly discovered RNA families to form an independent test set (not used in all the tested methods), and based on this dataset, a rigorous test was performed among 6 most accurate RNA secondary structure prediction methods. The test results showed that, among

these methods, the highest positive predictive value (PPV) is 0.636 (achieved by TORNADO) [84], the highest sensitivity is 0.720 (achieved by RNAfold) [17], and the highest F value is 0.632 (achieved by MXfold2) [79]. Using another independent dataset collected from PDB, Singh and colleagues [98] performed a comprehensive comparison among 27 kinds of well-known RNA secondary structure prediction methods. Their results showed when homologous sequences were available, the highest F value and sensitivity achieved were 0.774 and 0.727, respectively (both by SPOT-RNA2). These results objectively show that there is still much room for improvement in RNA secondary structure prediction. Moreover, many traditional methods neglect special base pairs to avoid a large number of false positives or to limit computational complexity [71,135]. While some methods can predict RNA secondary structures containing pseudoknots [46] or noncanonical base pairs [63], none of them can predict both. Although the recently proposed ML-based methods can predict all kinds of special base pairs, the special base pair prediction accuracy is still limited.

The RNA sequence length limitation is another intractable issue, which becomes quite problematic with the recently discovered long (1,000 to 10,000 nt) ncRNA [136]. Although ML-based methods do not suffer from high time complexity as most score-based methods do, they are unable to effectively capture such long-range interactions within an RNA sequence. On the other hand, training an ML model with such a large-scale input consumes a huge amount of computational resources and is often unrealistic.

For ML-based RNA secondary structure prediction models, overfitting is a very important issue [84], especially for DL-based models with a large number of parameters. The overfitted models perform well on the test RNAs similar to that in the training data but generalize poorly on dissimilar ones. It seems that they only memorize the secondary structure of RNAs in the training data, rather than actually learn the folding mechanism from them. A result in paper [79] showed that E2Efold [99] outperformed many traditional methods on the dataset ArchiveII but performed poorly on the RNAs from newly discovered RNA families. This suggested that E2Efold might suffer from a heavy overfitting. Similarly, another paper [137] reported that the F score of ContextFold also lowered by 24% when testing on a set of structurally dissimilar RNAs to the training set. Although most DL-based methods take many precautions to alleviate overfitting by many techniques (such as using regularization [100], enlarging dataset [97], adding constraints [99], or combining Turner's nearest neighbor free energy parameters), the concerns about overfitting remain.

At last, the folding mechanisms need further exploration. Traditional RNA secondary structure prediction is based on different RNA folding mechanism hypotheses (S1 Table), while data-driven ML-based methods can learn such mechanism implicitly from known data based on different RNA sequences or sequence features. However, to the best of our knowledge, few folding mechanisms have been revealed from the established ML-based models, although great advances have been made in terms of prediction accuracy. Part of the reason is that the interpretability [138] of DL models is still a challenge today.

Future trends of development

Currently, RNA secondary structure prediction is successfully shifting toward ML-based approaches, away from traditional score-based methods, and DL will surely continue to improve the prediction performance. The subtle structure of the DL model is a prerequisite to this end. Since the DL model is being rapidly developed in the natural language processing and image processing fields, using mature DL blocks from these fields, or combining them in such fields constitutes a feasible way to generate an excellent DL model for RNA secondary structure prediction.

Further, using a DL model to predict the free energy parameter is an inevitable trend for more accurate energy estimations, when additional wet lab experimental data become available. However, these parameters may not improve RNA secondary prediction accuracy because they have to be combined with traditional score-based methods. On the other hand, combing an ML-based method and an optimization method is a promising approach for improving prediction performance.

Conclusions

RNA structure is one of the central pieces of information for understanding biological processes, and determining RNA secondary structure will continue to be a hot topic in the computation and biology fields. In this review, we focused on ML-based methods, which involve many aspects of RNA secondary structure prediction. ML techniques have greatly improved the performance of prediction methods, including accuracy, applicability, and running speed. However, to thoroughly resolve the RNA secondary structure prediction problem, a more subtle ML model is still needed. At the moment, ML-based methods cannot be used as substitutes for wet lab experiments for obtaining high-resolution structures. Nonetheless, the advent of DL technologies and high-performance hardware will foster a new generation of RNA secondary prediction tools with an improved accuracy and running speed.

Supporting information

S1 Fig. Classification of ML-based RNA secondary structure prediction methods. According to the subprocess that ML participates in, the ML-based RNA secondary structure prediction methods were classified into 3 categories, i.e., score scheme based on ML (containing 3 subcategories: free energy–refining approach, weighted approach, and probabilistic approach), preprocessing and postprocessing based on ML (containing 2 subcategories: preprocessing and postprocessing), and prediction process based on ML (containing 2 subcategories: end-to-end approach and hybrid approach).
(TIF)

S1 Table. Comparison of RNA secondary structure prediction methods.
(DOCX)

References

1. Fu Y, Xu ZZ, Lu ZJ, Zhao S, Mathews DH. Discovery of Novel ncRNA Sequences in Multiple Genome Alignments on the Basis of Conserved and Stable Secondary Structures. PLoS ONE. 2015; 10(6): e0130200. https://doi.org/10.1371/journal.pone.0130200 PMID: 26075601.
2. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. https://doi.org/10.1038/nature11247 PMID: 22955616.
3. Consortium TF. The transcriptional landscape of the mammalian genome. Science. 2006; 311 (5768):1713. https://doi.org/10.1126/science.1121522 PMID: 16556825.
4. Doudna JA, Cech TR. The chemical repertoire of natural ribozymes. Nature. 2002; 418(6894):222–8. https://doi.org/10.1038/418222a PMID: 12110898.
5. Higgs PG, Lehman N. The RNA World: molecular cooperation at the origins of life. Nat Rev Genet. 2015; 16(1):7–17. https://doi.org/10.1038/nrg3841 PMID: 25385129.
6. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat Rev Genet. 2014; 15(7):469–79. https://doi.org/10.1038/nrg3681 PMID: 24821474.
7. Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. Nature. 2004; 431 (7006):343–9. https://doi.org/10.1038/nature02873 PMID: 15372041.
8. Serganov A, Nudler E. A Decade of Riboswitches. Cell. 2013; 152(1–2):17–24. https://doi.org/10.1016/j.cell.2012.12.024 PMID: 23332744.

9. Wu L, Belasco JG. Let me count the ways: Mechanisms of gene regulation by miRNAs and siRNAs. Mol Cell. 2008; 29(1):1–7. https://doi.org/10.1016/j.molcel.2007.12.010 PMID: 18206964.
10. Zou Q, Li J, Hong Q, Lin Z, Wu Y, Shi H, et al. Prediction of MicroRNA-Disease Associations Based on Social Network Analysis Methods. Biomed Res Int. 2015; 2015:810514. https://doi.org/10.1155/2015/810514 PMID: 26273645.
11. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015; 13(7):e1002195. https://doi.org/10.1371/journal.pbio.1002195 PMID: 26151137.
12. Tinoco I, Bustamante C. How RNA folds. J Mol Biol. 1999; 293(2):271–81. https://doi.org/10.1006/jmbi.1999.3001 PMID: 10550208.
13. Celander DW, Cech TR. Visualizing the higher order folding of a catalytic RNA molecule. Science. 1991; 251(4992):401–7. https://doi.org/10.1126/science.1989074 PMID: 1989074.
14. Zarrinkar PP, Williamson JR. Kinetic Intermediates in RNA Folding. Science. 1994; 265(5174):918–24. https://doi.org/10.1126/science.8052848 PMID: 8052848.
15. Chen SJ, Tan ZJ, Cao S, Zhang WB. The Statistical Mechanics of RNA Folding. Phys Ther. 2006; 35 (3):106–17. https://doi.org/10.3321/j.issn:0379–4148.2006.03.010
16. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973; 181(4096):223–30. https://doi.org/10.1126/science.181.4096.223 PMID: 4124164.
17. Lorenz R, Bernhart SH, Siederdissen CHZ, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011; 6:26. https://doi.org/10.1186/1748-7188-6-26 PMID: 22115189.
18. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31(13):3406–15. https://doi.org/10.1093/nar/gkg595 PMID: 12824337.
19. Bellaousov S, Reuter JS, Seetin MG, Mathews DH. RNAstructure: web servers for RNA secondary structure prediction and analysis. Nucleic Acids Res. 2013; 41(W1):W471–W4. https://doi.org/10.1093/nar/gkt290 PMID: 23620284.
20. Condon A, editor Problems on RNA Secondary Structure Prediction and Design. 30th International Colloquium on Automata, Languages and Programming (ICALP 2003); 2003; Berlin, Heidelberg: Springer Berlin Heidelberg.
21. Fallmann J, Will S, Engelhardt J, Grüning B, Backofen R, Stadler PF. Recent advances in RNA folding. J Biotechnol. 2017; 261:97–104. https://doi.org/10.1016/j.jbiotec.2017.07.007 PMID: 28690134.
22. Seetin MG, Mathews DH. RNA structure prediction: an overview of methods. Methods Mol Biol. 2012; 905:99–122. https://doi.org/10.1007/978-1-61779-949-5_8 PMID: 22736001.
23. Zhao Y, Wang J, Zeng C, Xiao Y. Evaluation of RNA secondary structure prediction for both base-pairing and topology. Biophysics Reports. 2018; 4(3):123–32. https://doi.org/10.1007/s41048-018-0058-y PMID: 20699301.
24. Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. RNA. 2001; 7 (4):499–512. https://doi.org/10.1017/s1355838201002515 PMID: 11345429.
25. Abu Almakarem AS, Petrov AI, Stombaugh J, Zirbel CL, Leontis NB. Comprehensive survey and geometric classification of base triples in RNA structures. Nucleic Acids Res. 2012; 40(4):1407–23. https://doi.org/10.1093/nar/gkr810 PMID: 22053086.
26. Doherty EA, Batey RT, Masquida B, Doudna JA. A universal mode of helix packing in RNA. Nat Struct Biol. 2001; 8(4):339–43. https://doi.org/10.1038/86221 PMID: 11276255.
27. van Batenburg FHD, Gultyaev AP, Pleij CWA. PseudoBase: structural information on RNA pseudoknots. Nucleic Acids Res. 2001; 29(1):194–5. https://doi.org/10.1093/nar/29.1.194 PMID: 11125088.
28. Staple DW, Butcher SE. Pseudoknots: RNA structures with diverse functions. PLoS Biol. 2005; 3(6): e213. https://doi.org/10.1371/journal.pbio.0030213 PMID: 15941360.
29. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, Underwood RC, et al. Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res. 1994; 22(23):5112–20. https://doi.org/10.1093/nar/22.23.5112 PMID: 7800507.
30. Westhof E. Twenty years of RNA crystallography. RNA. 2015; 21(4):486–7. https://doi.org/10.1261/rna.049726.115 PMID: 25780106.
31. Fürtig B, Richter C, Wöhnert J, Schwalbe H. NMR Spectroscopy of RNA. ChemBioChem. 2003; 4 (10):936–62. https://doi.org/10.1002/cbic.200300700 PMID: 14523911.
32. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010; 467(7311):103–7. https://doi.org/10.1038/nature09322 PMID: 20811459.

33. Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods. 2010; 7 (12):995–1001. https://doi.org/10.1038/nmeth.1529 PMID: 21057495.

34. Tijerina P, Mohr S, Russell R. DMS footprinting of structured RNAs and RNA-protein complexes. Nat Protoc. 2007; 2(10):2608–23. https://doi.org/10.1038/nprot.2007.380 PMID: 17948004.

35. Wilkinson KA, Merino EJ, Weeks KM. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc. 2006; 1 (3):1610–6. https://doi.org/10.1038/nprot.2006.249 PMID: 17406453.

36. Bevilacqua PC, Ritchey LE, Su Z, Assmann SM. Genome-Wide Analysis of RNA Secondary Structure. Annu Rev Genet. 2016; 50:235–66. https://doi.org/10.1146/annurev-genet-120215-035034 PMID: 27648642.

37. Tian S, Das R. RNA structure through multidimensional chemical mapping. Q Rev Biophys. 2016; 49: e7. https://doi.org/10.1017/S0033583516000020 PMID: 27266715.

38. Consortium TR. RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res. 2017; 45(D1):D128–D34. https://doi.org/10.1093/nar/gkw1008 PMID: 27794554.

39. Gutell RR, Lee JC, Cannone JJ. The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol. 2002; 12(3):301–10. https://doi.org/10.1016/s0959-440x(02)00339-1 PMID: 12127448.

40. Madison JT, Everett GA, Kung H. Nucleotide Sequence of a Yeast Tyrosine Transfer RNA. Science. 1966; 153(3735):531–4. https://doi.org/10.1126/science.153.3735.531 PMID: 5938777.

41. Gutell RR, Weiser B, Woese CR, Noller HF. Comparative anatomy of 16-S-like ribosomal RNA. Prog Nucleic Acid Res Mol Biol. 1985; 32:155–216. https://doi.org/10.1016/s0079-6603(08)60348-7 PMID: 3911275.

42. Han K, Kim HJ. Prediction of common folding structures of homologous RNAs. Nucleic Acids Res. 1993; 21(5):1251–7. https://doi.org/10.1093/nar/21.5.1251 PMID: 7681944.

43. Tahi F, Gouy M, Regnier M. Automatic RNA secondary structure prediction with a comparative approach. Comput Chem. 2002; 26(5):521–30. https://doi.org/10.1016/s0097-8485(02)00012-8 PMID: 12144180.

44. Tahi F, Engelen S, Regnier M. A fast algorithm for RNA secondary structure prediction including pseudoknots. Third IEEE Symposium on Bioinformatics and Bioengineering. 2003:11–7.

45. Engelen S, Tahi F. Tfold: efficient in silico prediction of non-coding RNA secondary structures. Nucleic Acids Res. 2010; 38(7):2453–66. https://doi.org/10.1093/nar/gkp1067 PMID: 20047957.

46. Bellaousov S, Mathews DH. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA. 2010; 16(10):1870–80. https://doi.org/10.1261/rna.2125310 PMID: 20699301.

47. Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics. 2004; 20(1):58–66. https://doi.org/10.1093/bioinformatics/btg373 PMID: 14693809.

48. Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, et al. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res. 1998; 26 (16):3825–36. https://doi.org/10.1093/nar/26.16.3825 PMID: 9685502.

49. Bindewald E, Shapiro BA. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA. 2006; 12(3):342–52. https://doi.org/10.1261/rna.2164906 PMID: 16495232.

50. Legendre A, Angel E, Tahi F. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. BMC Bioinformatics. 2018; 19(1):13. https://doi.org/10.1186/s12859-018-2007-7 PMID: 29334887.

51. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 2013; 41(D1):D226–D32. https://doi.org/10.1093/nar/gks1005 WOS:000312893300031. PMID: 23125362

52. Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A. 1980; 77(11):6309–13. https://doi.org/10.1073/pnas.77.11.6309 PMID: 6161375.

53. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981; 9(1):133–48. https://doi.org/10.1093/nar/9.1.133 PMID: 6163133.

54. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999; 288(5):911–40. https://doi.org/10.1006/jmbi.1999.2700 PMID: 10329189.

55. Andronescu M, Condon A, Turner DH, Mathews DH. The determination of RNA folding nearest neighbor parameters. Methods Mol Biol. 2014; 1097:45–70. https://doi.org/10.1007/978-1-62703-709-9_3_3. PMID: 24639154.

56. Xia TB, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao XQ, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998; 37(42):14719–35. https://doi.org/10.1021/bi9809425 PMID: 9778347.

57. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 2010; 38(Database issue):D280–2. https://doi.org/10.1093/nar/gkp892 PMID: 19880381.

58. Tinoco I Jr., Uhlenbeck OC, Levine MD. Estimation of secondary structure in ribonucleic acids. Nature. 1971; 230(5293):362–7. https://doi.org/10.1038/230362a0 PMID: 4927725.

59. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999; 49(2):145–65. https://doi.org/10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G PMID: 10070264.

60. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010; 11:129. https://doi.org/10.1186/1471-2105-11-129 PMID: 20230624.

61. Gultyaev AP, van Batenburg FH, Pleij CW. The computer simulation of RNA folding pathways using a genetic algorithm. J Mol Biol. 1995; 250(1):37–51. https://doi.org/10.1006/jmbi.1995.0356 PMID: 7541471.

62. Huang L, Zhang H, Deng D, Zhao K, Liu K, Hendrix DA, et al. LinearFold: linear-time approximate RNA folding by 5′-to-3′ dynamic programming and beam search. Bioinformatics. 2019; 35(14):i295–i304. https://doi.org/10.1093/bioinformatics/btz375 PMID: 31510672.

63. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 2008; 452(7183):51–5. https://doi.org/10.1038/nature06684 PMID: 18322526.

64. Honer zu Siederdissen C, Bernhart SH, Stadler PF, Hofacker IL. A folding algorithm for extended RNA secondary structures. Bioinformatics. 2011; 27(13):i129–36. https://doi.org/10.1093/bioinformatics/btr220 PMID: 21685061.

65. Dallaire P, Major F. Exploring Alternative RNA Structure Sets Using MC-Flashfold and db2cm. Methods Mol Biol. 2016; 1490:237–51. https://doi.org/10.1007/978-1-4939-6433-8_15 PMID: 27665603.

66. Sloma MF, Mathews DH. Base pair probability estimates improve the prediction accuracy of RNA noncanonical base pairs. PLoS Comput Biol. 2017; 13(11):e1005827. https://doi.org/10.1371/journal.pcbi.1005827 PMID: 29107980.

67. Poolsap U, Kato Y, Akutsu T. Prediction of RNA secondary structure with pseudoknots using integer programming. BMC Bioinformatics. 2009;10. https://doi.org/10.1186/1471-2105-10-10 PMID: 19133123.

68. Bon M, Micheletti C, Orland H. McGenus: a Monte Carlo algorithm to predict RNA secondary structures with pseudoknots. Nucleic Acids Res. 2013; 41(3):1895–900. https://doi.org/10.1093/nar/gks1204 PMID: 23248008.

69. Reeder J, Giegerich R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics. 2004;5. https://doi.org/10.1186/1471-2105-5-5 PMID: 14718068.

70. Dirks RM, Pierce NA. A partition function algorithm for nucleic acid secondary structure including pseudoknots. J Comput Chem. 2003; 24(13):1664–77. https://doi.org/10.1002/jcc.10296 PMID: 12926009.

71. Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol. 1999; 285(5):2053–68. https://doi.org/10.1006/jmbi.1998.2436 PMID: 9925784.

72. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015; 349(6245):255–60. https://doi.org/10.1126/science.aaa8415 PMID: 26185243.

73. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics. 2007; 23(13):i19–i28. https://doi.org/10.1093/bioinformatics/btm223 PMID: 17646296.

74. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Computational approaches for RNA energy parameter estimation. RNA. 2010; 16(12):2304–18. https://doi.org/10.1261/rna.1950510 PMID: 20940338.

75. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA. 2004; 10(10):1507–17. https://doi.org/10.1261/rna.5248604 PMID: 15383676.

76. Tang X, Thomas S, Tapia L, Giedroc DP, Amato NM. Simulating RNA folding kinetics on approximated energy landscapes. J Mol Biol. 2008; 381(4):1055–67. https://doi.org/10.1016/j.jmb.2008.02.007 PMID: 18639245.

77. Zakov S, Goldberg Y, Elhadad M, Ziv-Ukelson M. Rich parameterization improves RNA structure prediction. J Comput Biol. 2011; 18(11):1525–42. https://doi.org/10.1089/cmb.2011.0184 PMID: 22035327.

78. 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(6):1840025. https://doi.org/10.1142/S0219720018400255 PMID: 30616476.

79. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun. 2021; 12(1):941. https://doi.org/10.1038/s41467-021-21194-4 PMID: 33574226.

80. Woodson SA. Recent insights on RNA folding mechanisms from catalytic RNA. Cell Mol Life Sci. 2000; 57(5):796–808. https://doi.org/10.1007/s000180050042 PMID: 10892344.

81. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 2003; 31(13):3423–8. https://doi.org/10.1093/nar/gkg614 PMID: 12824339.

82. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics. 1999; 15(6):446–54. https://doi.org/10.1093/bioinformatics/15.6.446 PMID: 10383470.

83. Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics. 2004; 5:71. https://doi.org/10.1186/1471-2105-5-71 PMID: 15180907.

84. Rivas E, Lang R, Eddy SR. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. RNA. 2012; 18(2):193–212. https://doi.org/10.1261/rna.030049.111 PMID: 22194308.

85. 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(4):727–42. https://doi.org/10.1142/S0219720010004926 WOS:000271458900024.

86. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006; 22(14):e90–e8. https://doi.org/10.1093/bioinformatics/btl246 PMID: 16873527.

87. Yonemoto H, Asai K, Hamada M. A semi-supervised learning approach for RNA secondary structure prediction. Comput Biol Chem. 2015; 57:72–9. https://doi.org/10.1016/j.compbiolchem.2015.02.002 PMID: 25748534.

88. 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 Bioinformatics Online. 2013; 9:163–84. https://doi.org/10.4137/EBO.S10580 PMID: 23641141.

89. Zhu Y, Xie ZY, Li YZ, Zhu M, Chen YPP. Research on folding diversity in statistical learning methods for RNA secondary structure prediction. Int J Biol Sci. 2018; 14(8):872–82. https://doi.org/10.7150/ijbs.24595 PMID: 29989089.

90. Haynes T, Knisley D, Knisley J. Using a neural network to identify secondary RNA structures quantified by graphical invariants. Match Commun Math Comput Chem. 2008; 60(2):277–90. https://doi.org/10.1111/j.1467-9892.2007.00552.x WOS:000259765200002.

91. Koessler DR, Knisley DJ, Knisley J, Haynes T. A predictive model for secondary RNA structure using graph theory and a neural network. BMC Bioinformatics. 2010; 11(Suppl 6):S21. https://doi.org/10.1186/1471-2105-11-S6-S21 PMID: 20946605.

92. Takefuji Y, Chen LL, Lee KC, Huffman J. Parallel algorithms for finding a near-maximum independent set of a circle graph. IEEE Trans Neural Netw. 1990; 1(3):263–7. https://doi.org/10.1109/72.80251 PMID: 18282845.

93. Liu Q, Ye X, Zhang Y. A Hopfield Neural Network based algorithm for RNA secondary structure prediction. 1st International Multi Symposium on Computer and Computational Sciences; Hangzhou, China: IEEE; 2006.

94. Steeg EW. Neural networks, adaptive optimization, and RNA secondary structure prediction. Artificial intelligence and molecular biology. 1993:121–60.

95. Apolloni B, Torto LL, Morpurgo A, Zanaboni AM. RNA Secondary Structure Prediction by MFT Neural Networks 2003.

96. Qasim R, Kauser N, Jilani T. Secondary Structure Prediction of RNA using Machine Learning Method. Int J Comput Appl. 2011; 10(6):0975–8887. https://doi.org/10.5120/1486-2003

97. Singh J, Hanson J, Paliwal K, Zhou YQ. SPOT-RNA: RNA Secondary Structure Prediction using an Ensemble of Two-dimensional Deep Neural Networks and Transfer Learning. Nat Commun. 2019; 10(1):1–13. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773.

98. 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. Epub 2021/03/12. https://doi.org/10.1093/bioinformatics/btab165 PMID: 33704363.

99. Chen X, Li Y, Umarov R, Gao X, Song L. RNA Secondary Structure Prediction By Learning Unrolled Algorithms. International Conference on Learning Representations. 2020.

100. Calonaci N, Jones A, Cuturello F, Sattler M, Bussi G. Machine learning a model for RNA structure prediction. 2020; 2(4):lqaa090. https://doi.org/10.1093/nargab/lqaa090 PMID: 33575634.

101. Lu W, Tang Y, Wu H, Huang H, Fu Q, Qiu J, et al. Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter. BMC Bioinformatics. 2019; 20(Suppl 25):684. https://doi.org/10.1186/s12859-019-3258-7 PMID: 31874602.

102. Wu H, Tang Y, Lu W, Chen C, Huang H, Fu Q, editors. RNA Secondary Structure Prediction Based on Long Short-Term Memory Model. 14th International Conference on Intelligent Computing (ICIC); 2018; Wuhan, China.

103. 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; 384:104–14. https://doi.org/10.1016/j.neucom.2019.12.041 WOS:000513853600009.

104. Zhang H, Zhang C, Li Z, Li C, Wei X, Zhang B, et al. A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming. Front Genet. 2019; 10:467. https://doi.org/10.3389/fgene.2019.00467 PMID: 31191603.

105. Wang L, Liu Y, Zhong X, Liu H, Lu C, Li C, et al. 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. https://doi.org/10.3389/fgene.2019.00143 PMID: 30886627.

106. Liu Y, Zhao Q, Zhang H, Xu R, Li Y, Wei L. A New Method to Predict RNA Secondary Structure Based on RNA Folding Simulation. IEEE/ACM Trans Comput Biol Bioinform. 2016; 13(5):990–5. https://doi.org/10.1109/TCBB.2015.2496347 PMID: 26552091.

107. 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. https://doi.org/10.1515/cmb-2020-0002

108. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A. 2009; 106(1):97–102. https://doi.org/10.1073/pnas.0806929106 PMID: 19109441.

109. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. RNAZ 2.0: Improved Noncoding RNA Detection. Biocomputing. 2010; 15:69–79. https://doi.org/10.1142/9789814295291_0009. PMID: 19908359.

110. Washietl S, Will S, Hendrix DA, Goff LA, Rinn JL, Berger B, et al. Computational analysis of noncoding RNAs. Wiley Interdiscip Rev RNA. 2012; 3(6):759–78. https://doi.org/10.1002/wrna.1134 PMID: 22991327.

111. Moulton V. Tracking down noncoding RNAs. Proc Natl Acad Sci U S A. 2005; 102(7):2269–70. https://doi.org/10.1073/pnas.0500129102 PMID: 15703286.

112. Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF. Efficient computation of RNA folding dynamics. J Phys A Math Gen. 2004; 37(17):4731–41. https://doi.org/10.1088/0305-4470/37/17/005 WOS:000221482800006.

113. Rouillard JM, Zuker M, Gulari E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. 2003; 31(12):3057–62. https://doi.org/10.1093/nar/gkg426 PMID: 12799432.

114. Lu ZJ, Mathews DH. Efficient siRNA selection using hybridization thermodynamics. Nucleic Acids Res. 2008; 36(2):640–7. https://doi.org/10.1093/nar/gkm920 PMID: 18073195.

115. Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, et al. The impact of target site accessibility on the design of effective siRNAs. Nat Biotechnol. 2008; 26(5):578–83. https://doi.org/10.1038/nbt1404 PMID: 18438400.

116. Sazani P, Gemignani F, Kang SH, Maier MA, Manoharan M, Persmark M, et al. Systemically delivered antisense oligomers upregulate gene expression in mouse tissues. Nat Biotechnol. 2002; 20(12):1228–33. https://doi.org/10.1038/nbt759 PMID: 12426578.

117. Childs-Disney JL, Wu M, Pushechnikov A, Aminova O, Disney MD. A small molecule microarray platform to select RNA internal loop-ligand interactions. ACS Chem Biol. 2007; 2(11):745–54. https://doi.org/10.1021/cb700174r PMID: 17975888.

118. Palde PB, Ofori LO, Gareiss PC, Lerea J, Miller BL. Strategies for Recognition of Stem-Loop RNA Structures by Synthetic Ligands: Application to the HIV-1 Frameshift Stimulatory Sequence. J Med Chem. 2010; 53(16):6018–27. https://doi.org/10.1021/jm100231t PMID: 20672840.

119. Castanotto D, Rossi JJ. The promises and pitfalls of RNA-interference-based therapeutics. Nature. 2009; 457(7228):426–33. https://doi.org/10.1038/nature07758 PMID: 19158789.

120. Gareiss PC, Sobczak K, McNaughton BR, Palde PB, Thornton CA, Miller BL. 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 (48):16254–61. https://doi.org/10.1021/ja804398y PMID: 18998634.

121. Tavares RdCA, Mahadeshwar G, Wan H, Huston NC, Pyle AM. The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. J Virol. 2020; 95(6):e02190–20. https://doi.org/10.1128/JVI.02190-20 PMID: 33268519.

122. Vandelli A, Monti M, Milanetti E, Armaos A, Rupert J, Zacco E, et al. Structural analysis of SARS-CoV-2 and predictions of the human interactome. Nucleic Acids Res. 2020; 48(20):11270–77283. https://doi.org/10.1093/nar/gkaa864 PMID: 33068416.

123. Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinformatics. 2008; 9:340. https://doi.org/10.1186/1471-2105-9-340 PMID: 18700982.

124. Burley SK, Bhikadiya C, Bi CX, Bittrich S, Chen L, Crichlow GV, 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. 2021; 49(D1):D437–D51. https://doi.org/10.1093/nar/gkaa1038 PMID: 33211854.

125. 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(11):5381–94. https://doi.org/10.1093/nar/gky285 PMID: 29746666.

126. Juhling F, Morl M, Hartmann RK, Sprinzl M, Stadler PF, Putz J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009; 37:D159–D62. https://doi.org/10.1093/nar/gkn772 PMID: 18957446.

127. Gutell RR. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. Nucleic Acids Res. 1993; 21(13):3051–4. https://doi.org/10.1093/nar/21.13.3051 PMID: 8332526; PubMed Central PMCID: PMC7524024.

128. Zwieb C, Gorodkin J, Knudsen B, Burks J, Wower J. tmRDB (tmRNA database). Nucleic Acids Res. 2003; 31(1):446–7. https://doi.org/10.1093/nar/gkg019 PMID: 12520048.

129. Richardson KE, Kirkpatrick CC, Znosko BM. RNA CoSSMos 2.0: an improved searchable database of secondary structure motifs in RNA three-dimensional structures. Database-Oxford. 2020:baz153. https://doi.org/10.1093/database/baz153 PMID: 31950189.

130. Korunes KL, Myers RB, Hardy R, Noor MAF. PseudoBase: a genomic visualization and exploration resource for the Drosophila pseudoobscura subgroup. Fly. 2021; 15(1):38–44. https://doi.org/10.1080/19336934.2020.1864201 PMID: 33319644.

131. Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang ZD, Zhao Q, et al. NCIR: a database of non-canonical interactions in known RNA structures. Nucleic Acids Res. 2002; 30(1):395–7. https://doi.org/10.1093/nar/30.1.395 10.1093/nar/30.1.395. PMID: 11752347.

132. Sloma MF, Mathews DH. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. RNA. 2016; 22(12):1808–18. https://doi.org/10.1261/rna.053694.115 PMID: 27852924.

133. Tan Z, Fu YH, Sharma G, Mathews DH. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. Nucleic Acids Res. 2017; 45(20):11570–81. https://doi.org/10.1093/nar/gkx815 PMID: 29036420.

134. Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28(23):3150–2. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610.

135. Lyngso RB, Pedersen CN. RNA pseudoknot prediction in energy-based models. J Comput Biol. 2000; 7(3–4):409–27. https://doi.org/10.1089/106652700750050862 PMID: 11108471.

136. Johnsson P, Lipovich L, Grander D, Morris KV. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta. 2014; 1840(3):1063–71. https://doi.org/10.1016/j.bbagen.2013.10.035 PMID: 24184936.

137. Rivas E. The four ingredients of single-sequence RNA secondary structure prediction. A unifying perspective. RNA Biol. 2013; 10(7):1185–96. https://doi.org/10.4161/rna.24971 PMID: 23695796.

138. Carvalho DV, Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics-Switz. 2019; 8(8). https://doi.org/10.3390/electronics8080832 WOS:000483554300063.

139. Apolloni B, Lotorto L, Morpurgo A, Zanaboni A. RNA Secondary Structure Prediction by MFT Neural Networks. Psychol Forsch. 2003:143–8.