Advances in RNA secondary structure prediction and RNA modifications: Methods, data, and applications

Shu Yang[1,*]**, Nhat Truong Pham**[2,*]**, Ziyang Li**[3,*]**, Jae Young Baik**[1,*]**, Joseph Lee**[1]**, Tianhua Zhai**[1]**, Weicheng Yu**[1]**, Bojian Hou**[1]**, Tianqi Shang**[1]**, Weiqing He**[1]**, Duy Duong-Tran**[1,4,†]**, Mayur Naik**[3,†]**, Li Shen**[1,†]

[1] Dept. of Biostatistics, Epidemiology and Informatics, University of Pennvylvania, Philadelphia, USA
[2] Dept. of Integrative Biotechnology, Sungkyunkwan University, Suwon, Republic of Korea
[3] Dept. of Computer and Information Science, University of Pennvylvania, Philadelphia, USA
[4] Dept. of Mathematics, United States Naval Academy, Annapolis, USA

Abstract

Due to the hierarchical organization of RNA structures and their pivotal roles in fulfilling RNA functions, the formation of RNA secondary structure critically influences many biological processes and has thus been a crucial research topic. This review sets out to explore the computational prediction of RNA secondary structure and its connections to RNA modifications, which have emerged as an active domain in recent years. We first examine the progression of RNA secondary structure prediction methodology, focusing on a set of representative works categorized into thermodynamic, comparative, machine learning, and hybrid approaches. Next, we survey the advances in RNA modifications and computational methods for identifying RNA modifications, focusing on the prominent modification types. Subsequently, we highlight the interplay between RNA modifications and secondary structures, emphasizing how modifications such as m6A dynamically affect RNA folding and vice versa. In addition, we also review relevant data sources and provide a discussion of current challenges and opportunities in the field. Ultimately, we hope our review will be able to serve as a cornerstone to aid in the development of innovative methods for this emerging topic and foster therapeutic applications in the future.

Keywords: RNA secondary structures, RNA modifications, bioinformatics, machine learning, deep learning, RNA language models

1 Introduction

Ribonucleic acids (RNAs) are crucial for fundamental biological processes, from catalyzing biochemical reactions to regulating gene expression in all organisms. Beyond their well-known role as intermediaries in the central dogma, RNAs exhibit extraordinary functional versatility. Ribozymes catalyze essential biochemical reactions, such as peptide bond formation in the ribosome, while riboswitches sense metabolites and regulate gene expression in response to environmental changes. Regulatory RNAs, including microRNAs and long non-coding RNAs, orchestrate complex cellular processes, ranging from development and differentiation to stress response and disease progression. Emerging classes, such as circular RNAs and RNA aptamers, continue to expand our understanding of RNA's regulatory and structural repertoire. This functional diversity is underpinned by the structure of RNA, where its precise shape and structure enable it to carry out catalytic functions, bind to diverse molecules, and regulate complex cellular processes [1,2,3,4,5]. As a result, understanding RNA structure is not only fundamental to biological mechanisms but also holds promise for therapeutic innovation, including the development of RNA-based drugs and vaccines [6].

The structure of RNA exhibits a hierarchical and sequential organization: the primary structure consists of a linear sequence of nucleotides, the secondary structure such as stems and loops is dictated by base pair interactions, the tertiary structure builds upon these shapes with three-dimensional folding, and this determines quaternary interactions with other molecules [7,8,9]. Among these, the secondary structure holds particular significance as it serves as a critical scaffold for higher-order folding and governs key aspects of RNA function. RNA secondary structure (RSS) formation is driven by the ability of nucleotides to engage in both canonical Watson–Crick basepairings and noncanonical ones,

---

such as GU pairs. These pairings give rise to distinct secondary structural motifs (see Figure 1 for an example), including stems, bulges, different types of loops, and pseudoknots, which consist of various configurations of paired and unpaired nucleotides. These secondary elements further interact with each other through mechanisms like coaxial stacking and kissing loop interactions, ultimately assembling into the more complex tertiary or quaternary structures of RNA. Notably, RNA secondary structure is often more conserved than its primary sequence across homologous RNAs, emphasizing its central role in defining RNA behavior and function [10].

Determining RNA structures through experimental techniques (e.g., X-ray crystallography) remains a significant challenge due to their low throughput, high resource demands, and technical limitations. These methods have resolved only a fraction of known RNA molecules, leaving significant gaps in our understanding of RNA structure. To address these limitations, computational prediction has emerged as an essential tool for scaling RNA structural analysis. However, while protein folding has seen transformative advancements, such as AlphaFold's ability to predict 3D structures with remarkable accuracy [11], RNA folding remains a difficult challenge. Firstly, one fundamental obstacle in RNA structure prediction is the inadequacy and bias of existing datasets. While protein structure prediction has benefited from large, high-quality databases, RNA data in repositories like the Protein Data Bank (PDB) is both limited and heavily skewed towards simpler structures, such as tRNAs and rRNA subunits. For example, there are 25 times more proteins structure data entries than RNAs in PDB. There are >19,600 protein families in Pfam, but only >4,100 RNA families in Rfam. Moreover, the lack of structural diversity also hinders the development of models that can accurately predict the wide variety of RNA structural motifs, e.g. long-range interactions in lncRNAs. In addition, unlike proteins, whose folding is driven by well-characterized forces like hydrophobic interactions and standardized motifs, RNA folding is more dynamic and involves intricate secondary structures that form the scaffold for complex tertiary interactions. This makes secondary structure prediction an essential focus in RNA research.

Over the years, three major computational prediction strategies have emerged for RSS folding: thermodynamic, comparative, and machine learning (including deep learning). Thermodynamic models predict structures by minimizing free energy, using experimentally derived parameters to estimate the stability of base-pair interactions and loops. While effective for canonical structures, these methods are limited by incomplete energy models and challenges in handling non-canonical pairs and pseudoknots. Comparative approaches leverage evolutionary conservation to predict RNA secondary structures, assuming that functionally important structures are preserved through compensatory mutations. These methods typically rely on a multiple sequence alignment (MSA) of homologous RNA sequences and use probabilistic models, such as stochastic context-free grammars (SCFGs), to identify co-varying base pairs. Instead of minimizing energy, they predict the structure that best explains the observed evolutionary covariation, selecting the one with the highest likelihood. Machine learning and deep learning methods, in contrast, learn structural patterns in a data-driven way, capturing complex interactions without explicit reliance on energy models or sequence homology. These models, exemplified by SPOT-RNA [12] and Ufold [13], offer state-of-the-art accuracy for diverse RNA structures, including those with intricate motifs and pseudoknots, marking a paradigm shift in RNA structure prediction. Notably, foundational models like RNA-FM [14] have further advanced the field, leveraging vast amounts of RNA sequences with unsupervised learning.

Accurate prediction of RSS from the RNA sequence is fundamental to understanding RNA function and interactions; however, there are many other factors affecting RNA folding in the cell. Among them, chemical modifications of RNA, such as methylation and pseudouridylation, play pivotal roles in influencing RNA thermodynamic stability by altering the free energy landscape. These modifications can impact base pairing, reshape the secondary structure, and modulate RNA interactions with proteins and other biomolecules, thereby influencing key biological processes such as translation, splicing, and RNA stability. Indeed, the so-called epitranscriptome of RNA modifications is suggested to have broad regulation of RNA structuredness, supported by observations of transcriptome wide RNA modifications [15,16]. Notably, RNA modifications are remarkably diverse and widely distributed across various RNA species. While covalent nucleotide modifications were traditionally recognized as abundant in transfer RNAs (tRNAs), advances in high-throughput sequencing technologies have revealed their widespread occurrence in messenger RNAs (mRNAs) and non-coding RNAs. As of this review, the Modomics database catalogs over 335 natural RNA modifications [17], though most remain incompletely characterized. Emerging evidence underscores the critical roles of these modifications in shaping RNA secondary and tertiary structures [16,18], facilitating RNA-protein interactions [19,20], and regulating splicing [21]. A comprehensive understanding of the interplay between RNA modifications and structural dynamics is crucial for

elucidating RNA biology, with far-reaching implications for regulatory mechanisms and therapeutic development.

Specifically, post-transcriptional RNA modifications, which involve chemical alterations to RNA molecules, have recently emerged as a major focus of research[22,23,24]. Over the past decade, several prominent modifications have been extensively explored, including N6-methyladenosine (m6A), N1-methyladenosine (m1A), 5-methylcytosine (m5C), 5-methyluridine (m5U), N6,2'-O-dimethyladenosine (m6Am), N7-methylguanosine (m7G), N4-acetylcytosine (ac4C), pseudouridine (Ψ), 2'-O-methyladenosine (Am), 2'-O-methylcytidine (Cm), 2'-O-methylguanosine (Gm), 2'-O-methyluridine (Um), uridylation, and adenosine-to-inosine (A-to-I) RNA editing[25,26,27]. Among these, m6A is recognized as the most prevalent and abundant mRNA modification in eukaryotes, accounting for approximately 0.1–0.6% of all adenosines. This modification involves the addition of a methyl group to the nitrogen atom at the sixth position of adenosine and is conserved across a wide range of organisms, from bacteria to mammals. Another well-studied modification, m5C, is characterized by the addition of a methyl group to the carbon-5 position of the cytosine base. This modification is commonly observed across various RNA species, including tRNAs, ribosomal RNAs (rRNAs), mRNAs, enhancer RNAs (eRNAs), and non-coding RNAs. Additionally, 2'-O-methylation (Nm or 2OM) is an RNA modification that occurs co-transcriptionally or post-transcriptionally, involving the addition of a methyl group to the 2' hydroxyl group of the ribose sugar in the RNA backbone. These modifications collectively contribute to the structural and functional diversity of RNA molecules, enabling their involvement in a broad spectrum of biological processes.

Subsequently, several experimental methods have been developed to accurately identify modifications in RNA. However, these methods tend to be labor-intensive, require specialized tools, can damage RNA samples, and present challenges when working with minimal RNA quantities. To address these limitations, high-throughput techniques based on deep sequencing have been introduced to identify RNA modifications at the transcriptome level. Nevertheless, these techniques remain costly, time-consuming, and necessitate specialized expertise. Consequently, computational methods have been developed to complement the experimental techniques. Recently, several databases have been established, serving as foundational resources for developing computational methods to identify RNA modifications, such as RMBase V3.0[28], MODOMICS[17], m7GHub V2.0[29], and the dataset by Song *et al.*[25]. Furthermore, a variety of computational tools have emerged for several common RNA modifications, including 2OM (H2Opred[30], Meta-2OM[31], and Nmix[32]), ac4C (ac4C-AFL[33], Voting-ac4C[34], iRNA-ac4C[35], and TransAC4C[36]), m5C (Deepm5C[37], MLm5C[38], and m5C-pred[39]), m6A (MST-m6A[40], CLSM6A, and BLAM6A-Merge[41]), m7G (Moss-m7G[42] and THRONE[43]), and multiple transcriptome modifications (MultiRM[25], TransRNAm[27], and class incremental learning for RNA modifications (CIL-RNA)[26]). Detailed information on these tools will be discussed in later sections.

Here, this review aims to survey the advances in RNA secondary structure prediction and RNA modification, with an emphasis on the connections between these two, and to outline future directions for this emerging field. There are a number of existing reviews on related topics recently.[44,45,46,47,48] Our review sets itself apart by: 1. focusing on a representative set of diverse RSS prediction methods to offer more in-depth descriptions for each, rather than giving a broad but brief overview, and to show the methodology progression in this field, rather than concentrating on specific types of methodology (e.g., deep learning), so that the review could be hopefully beneficial for researchers or practitioners relatively familiar with the topic and also useful for those with non-expert backgrounds; and 2. dedicating substantial efforts to discussing the potential applications of RNA secondary structure prediction in the context of RNA modifications, a cutting-edge topic with immense potentials in human disease study. In the following sections of the paper, we will first review the progression of RNA secondary structure prediction methods by concentrating on a set of selected works representing different strategies in the field. Then, we will shift our focus to RNA modification prediction methods with another set of representative works corresponding to different modification types. After that, we will delve into the interplay between RNA secondary structure and RNA modification, building on top of the earlier sections to highlight the close relationship between the two. Additionally, we will provide a summary of available data that are commonly used in this area. Lastly, we will discuss the challenges and opportunities for future work to conclude the review.

2 Computational tools for RNA secondary structure prediction

As mentioned above, RNA structure is formed hierarchically, and the secondary structure formation is key to study the functions of the RNA. Thus, the in silico RNA secondary structure prediction has long been a cornerstone in bioinformatics. The first fast algorithm for RSS prediction was published in 1980[50] by Nussinov and Jacobson, a

```
GAUGAGACGGACAUUCGUAUACCGGACAUUAAGACGUAAAAGACCAGUGCUCAAGUAGGGAUUCCUAAAUCAUAACACGGAUGCCAUGUCAUCCUUUA
(((((....(((((.(((((.....((.((((((((............)).)))))))))...(((((....))))...))))).)))))...))))).....
```

Figure 1: **An example of RNA secondary structures.** As shown in this example, an RSS can be decomposed into a set of stem-loop structural motifs (indicated with different colors). The RNA sequence and the corresponding dot-bracket notation are shown at the bottom. This plot is generated with the help of the Forna visualization tool from the ViennaRNA package[49]. The example assumes pseudoknot-free.

foundational contribution that continues to influence the field even now. After that, the field has evolved significantly over the past decades, especially in recent years due to the success in deep learning and large foundation models. Therefore, we select a number of representative methods to review here, reflecting the progression of methodology developments and hopefully shedding lights on potential future directions.

RNA secondary structure prediction methods can be classified into different categories based on different but subtly related criteria. Conventional, the most commonly seen classifications being the energy-based model versus the probability model (e.g., stochastic context-free grammar model), according to the type of parameters used, or single-sequence structure prediction versus comparative structure prediction, according to the type of inputs required. In a loose sense, energy-based method roughly overlaps with single-sequence structure prediction, while the probability model largely overlaps with comparative prediction. Recent advancements of machine learning- and deep learning-based approaches have drastically changed the research paradigm. In addition, the emerging hybrid methods have made the boundaries among different classes become even more vague. So here, in order to show the methodology progression in this field, we follow a straightforward notion to refer to these methods as energy-based, comparative, learning-based, and hybrid methods. A summary of the key ideas of the different classes are shown in Table 1. Besides, technically speaking, SCFG is also learning-based as a probabilistic model itself; but since it has been widely-used in comparative methods, we single it out without putting it together with the other machine learning or deep learning methods.

2.1   Thermodynamic free energy-based methods

The idea of energy-based structure prediction is based on the principle of free energy minimization. Since the secondary structure of RNA molecules is predominantly determined by interactions like hydrogen bonds and base stackings, computing the energy of these interactions provides insights into the structure. For a closed system with fixed entropy, equilibrium corresponds to a state that minimizes the system's free energy. Therefore, the most stable secondary structure of RNA is assumed to be the one with the minimum free energy (MFE). Methods in this category predict the secondary structure that minimizes the overall free energy of an RNA in thermodynamic equilibrium by considering all potential RNA secondary structures and their respective abundance according to the Boltzmann distribution.

The energy-based structure prediction typically takes a single RNA sequence as input and predicts the best structure of that sequence to be the most stable structure, i.e. the one with the minimum free energy. Such prediction uses the

Table 1: **A summary of RNA secondary structure prediction strategies.**

Strategy	Method	Input
Energy-based	Based on thermodynamic free energy, find an optimal structure with minimum free energy	Single sequence
Comparative	Based on co-variation and probabilistic model, find the optimal structure with maximum likelihood	Multiple sequence alignment
Learning-based [1]	Data driven, advanced machine learning or deep learning approaches	Diverse
Hybrid	Combining two or more strategies mentioned above	Diverse

[1] Learning-based strategy here refers to those non-SCFG-based machine learning and deep learning methods since, technically speaking, the classic SCFG model is also learning-based.
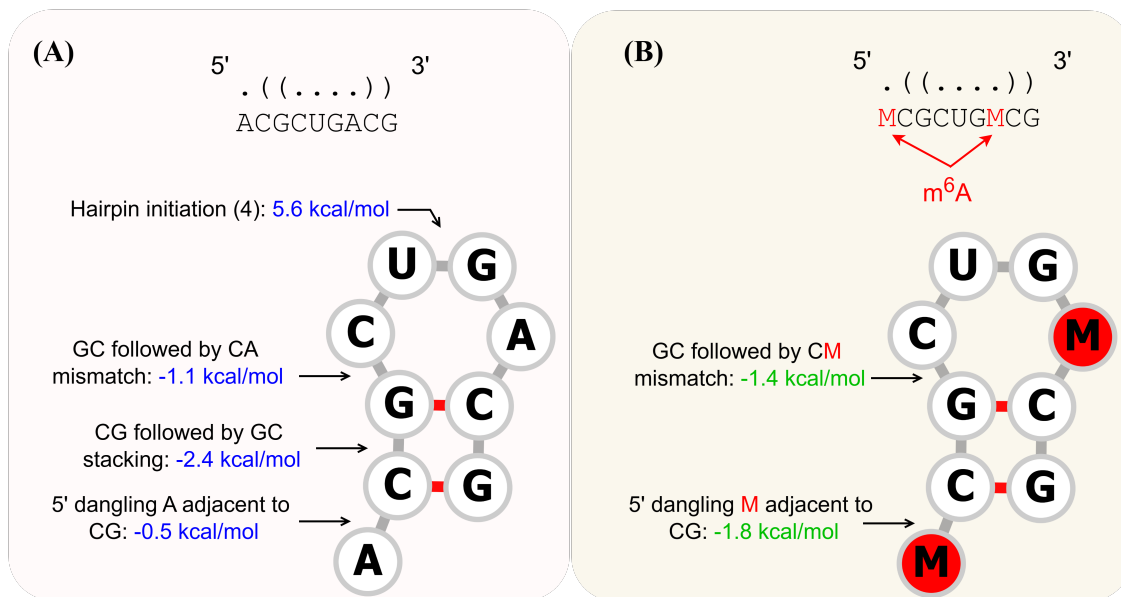
Figure 2: **RSS free energy computation based on nearest neighbor energy model. (A).** An example of the free energy calculation for an RSS using Turner's nearest neighbor parameters[51]. This figure is generated with the help of the Forna visualization tool from the ViennaRNA package[49]. The overall free energy of a given RNA secondary structure can be expressed as the sum of free energies across different structural units. For this small hairpin structure shown here, the overall free energy is the sum of the destabilizing loop and bulge energy (e.g. [+5.6] for hairpin initiation (4) as shown in the figure) and the stabilizing energy contributions from pairs of neighboring basepairs (e.g. [-2.4] for the CG followed by GC stacking interaction). **(B).** A side-by-side comparison of m6A modified RNA with a different nearest neighbor model[18], assuming the same RSS as the unmodified one. As we can see, the energy values are very different for structures involving the normal A from those involving m6A (denoted by letter M in the figure). For example, the 5' dangling A has [-0.5] while m6A has [-1.8].

stacking of base pairs as its basic unit. As shown in Figure 1, an RNA secondary structure can be decomposed into a set of stem-loop structural motifs (indicated with different colors) such as stacking/stem, hairpin loop, internal loop,

bulge, multibranch loop, and external loop/dangling region, etc. The prediction method should have parameters to account for all these motifs, in terms of their energetic contributions.

The nearest neighbor models[52,53] provide the thermodynamic free energy parameters and are the basis for most methods involving energy-based computations. The nearest neighbor models are a set of rules and associated parameters that predict the folding free energy of a secondary structure by decomposing the structure into loop substructures enclosed by the nearest neighboring basepairs. The model rules have two general assumptions: 1. the free energy of a basepair or loop substructure depends only on the sequence of that substructure and the sequence of the directly adjacent basepairs. 2. the total free energy can be calculated by summing the energies predicted for each substructure. The set of model parameters stores the energy values for the smallest unit of secondary structure, corresponding to thermodynamic quantities pre-determined by wet lab experiments such as optical melting experiments. The most famous one is the Turner's nearest neighbor model, which has been widely used by many methods including those energy-based methods in Table 2.

Figure 2**(A)** shows an example of the nearest neighbor model of an RNA structure. The additivity characteristic indicates that globally optimal structures are composed of locally optimal sub-structures, which is ideal for computational approaches like dynamic programming algorithms to deal with. Dynamic programming (DP) algorithm is the first and most widely used approach for energy-based RSS prediction[50,54,55]. It can recursively calculate the minimum energy structure. As the goal in RNA structure prediction is to find the RNA structure that minimizes the overall free energy for a given RNA input sequence, and as the overall free energy can be expressed as the sum of free energy contributions from structural building blocks recursively, efficient DP algorithms exist to calculate the optimal RNA secondary structure in $O(N^3)$ time and $O(N^2)$ memory for an input sequence of length N.

Besides the structural motifs shown in Figure 1, additional structures can be formed when unpaired bases match with distant ones, such as the base A at position 40 and the U at position 6 in Figure 1. This would form the so-called pseudoknot structure, which could make the prediction much harder. For example, the Zuker algorithm[54] underlying many DP algorithms like RNAfold[56] etc. is incapable of predicting pseudoknot; while a later enhanced DP algorithm by Rivas and Eddy to deal with pseudoknots has time complexity of $O(N^6)$ and space complexity $O(N^4)$[57], which is often intractable in practice. Note that since many of the methods reviewed below do not deal with pseudoknots, we will specify this capability if otherwise for clarity.

RNAFOLD One of the pioneers and most widely used MFE-based RNA secondary structure prediction methods is RNAFOLD from the ViennaRNA package[56]. It employs Zuker's dynamic programming algorithm[54] for efficient computation of the optimal global folding with Turner's nearest neighbor free energy parameters as the scores[58,51]. By accepting a single RNA sequence in FASTA format as input, RNAfold systematically evaluates possible base-pairing interactions to identify the thermodynamically most stable secondary structure, simultaneously reporting the predicted MFE values and providing dot-bracket notations for quick visualization. Zuker's algorithm has a time complexity of $O(N^3)$ that scales cubically with the sequence length N. RNAfold also has the functionality to output the MFE probability and base-pairing probability matrix by utilizing McCaskill's partition function approach[59] (also $O(N^3)$ runtime but with a larger constant factor than the Zuker's) to consider the thermodynamic ensemble of all structures following the Boltzmann distribution. RNAfold is based on well-established models in the early days and serves as the main program in the ViennRNA package, making it one of the most classic and reliable MFE-based methods for RNA secondary structure prediction.

RNASTRUCTURE RNASTRUCTURE is a software tool for RNA secondary structure prediction and analysis[60,61]. The algorithms in RNASTRUCTURE employ nearest neighbor parameters to predict the stability of secondary structures. These parameters, based on the Turner group[62,51,63], include both free energy change at 37°C and enthalpy change to facilitate the prediction of conformational stability at various temperatures. RNASTRUCTURE offers a range of algorithms, including methods for secondary structure prediction, base pair probability estimation, bimolecular structure prediction, and identifying common structures between two sequences. These features are complemented by a user-friendly JAVA-based GUI with cross-platform compatibility.

SIMFOLD SIMFOLD is a computational tool designed to predict RNA secondary structures using thermodynamic free-energy models. It employs advanced parameter estimation techniques, such as the Constraint Generation (CG)

and Boltzmann Likelihood (BL) methods, to optimize energy parameters for RNA folding. SIMFOLD supports the Turner energy model and its variants, effectively integrating structural and thermodynamic data to enhance prediction accuracy. Benchmark studies demonstrate that SIMFOLD achieves a significant improvement in F-measure over standard Turner parameters, particularly when optimized with the BL method. The tool accepts input sequences in FASTA format and outputs RNA secondary structures with minimum free energy, offering accurate and reliable predictions for pseudoknot-free RNA configurations. SIMFOLD's robust thermodynamic modeling and energy parameter refinements make it an essential tool for RNA structure prediction tasks.

LINEARFOLD LINEARFOLD is the first linear-time and linear-space prediction algorithm for RNA secondary structures. It uses the thermodynamic free energy model[51] from Vienna RNAfold[56]. Traditional algorithms for predicting RNA structures, such as dynamic programming-based methods, scale with cubic time complexity $O(N^3)$, which limits their use for long RNA sequences. LINEARFOLD overcomes this bottleneck by scanning the sequence in a left-to-right (5'-to-3') direction, following the transcription process, rather than bottom-up, utilizing a beam search heuristic to reduce the complexity to $O(N)$, making it significantly faster without a large sacrifice in accuracy. It accepts input sequences in both FASTA format and pure-sequence format. LINEARFOLD demonstrates superior efficiency and scalability without sacrificing accuracy, making it particularly effective for long sequences and long-range base pair predictions.

2.2 Comparative methods

The comparative strategy usually uses probabilities as parameters and stochastic context free grammars as the underlying model. Although there are exceptions[64], it typically needs a functionally equivalent multiple sequence alignment as input and predicts the best structure of that alignment to be the most likely structure, e.g., the one with the maximum likelihood.

The idea of the comparative structure prediction is based on evolution, as shown in Figure 3. It assumes functionally important RNA structures are conserved through evolution. So, it looks for conserved base pairs, especially those compensatory mutations through evolution. An algorithm in this category needs the alignment of functionally equivalent sequences, and it finds those co-varying base pairs in the alignment columns. Instead of using energy parameters, it uses a probabilistic model like SCFG. The final predicted RNA structure would be the one that best explains those co-variations according to the model.

Table 2: **A summary of representative RNA secondary structure prediction methods included in the review.**

	Method	Input	Other notes	URLs
LINEARFOLD[65]	**(E)**. Incremental Beam Search algorithm with a thermodynamic energy model	Single sequence (FASTA)	Sequence length limits to 100,000	Code, Web server
RNAFOLD[56]	**(E)**. MFE with Zuker's and McCaskill's algorithms	Single sequence (FASTA)	One of the most classic MFE methods and the main RSS prediction tool in ViennaRNA	Code, Web server
RNASTRUCTURE[60,61]	**(E)**. Software with a set of folding algorithms	Single or multi-sequence (FASTA, SEQ)	Predicts MFE structures, base pair probabilities, bimolecular structures, and structures common to two sequences	Code, Web server
SIMFOLD[66,67]	**(E)**. Minimum free energy based on a discriminative framework	Single sequence (FASTA)	Current implementation includes suboptimal folding calculations, as well as partition functions, base pair probabilities, and gradient computations.	Code

Continuation of table

	Method	Input	Other notes	URLs
PPFOLD [68,69]	**(C)**. SCFG with phylogenetics information	Alignment; phylogenetic tree (optional)	Uses probabilities as parameters to measure the co-varying tendency of positions pair	Code
RNADECODER [70,71]	**(C)**. SCFG with several phylogenetic models	Alignment; phylogenetic tree; codon annotation	Designed explicitly to take protein-coding context into account; does not assume global RNA structure	Code, Web server
TORNADO [64]	**(C)**. Generalized super-grammar for SCFG described in TORNADO programming language	Single sequence (FASTA, Stockholm)	General-purpose SCFG tool with flexible adaptability; supports complex structural modeling and customization	Code
CNNFOLD [72]	**(L)**. Deep learning-based model using CNN architecture	Single sequence (FASTA)	Simple implementation of CNN-based method	Code
CONTRAFOLD [73]	**(L)**. Conditional log-linear models (CLLMs), a discriminative generalization of SCFG	Single sequence (FASTA)	Construct the CLLM from the energy model to find the maximum-expected-accuracy structure. Sequence length limits to 1000 on the web server.	Code, Web server
E2EFOLD [74]	**(L)**. End2end learning with a transformer-based Deep Score Network and a multilayer Post-Processing Network with an unrolled algorithm to reduce overfitting	One-hot encoding of single sequence	The unrolled algorithm uses a primal-dual constrained optimization to incorporate base pairing constraints	Code
REDFOLD [75]	**(L)**. Deep learning-based model using ResNet and FC-DenseNet network	Single sequence (FASTA)		Code, Web server
SPOT-RNA [12]	**(L)**. Ensemble of ResNets, 2D-BiLSTM and dilated CNN models; transfer learning	One-hot encoded single sequence (or batch sequences)	Able to predict all base pairs including noncanonical and non-nested (pseudoknot) ones	Code, Web server
SPOT-RNA2 [76]	**(L)**. Ensemble of dilated CNN models; transfer learning	One-hot encoding and predicted basepair probability from single sequence, PSSM and DCA from evolution	Extends SPOT-RNA by incorporating additional evolutionary information	Code, Web server
UFOLD [13]	**(L)**. Deep learning-based model using U-Net architecture	Single sequence (FASTA)	One of the first deep-learning models that converts RNA sequences to an "image" format	Code

Continuation of table

	Method	Input	Other notes	URLs
MxFold [77]	**(H)**. Thermodynamic and structured support vector machines hybrid	Single sequence (FASTA, bpseq)	Integrates thermodynamic parameters with machine learning to improve prediction accuracy; limited scalability for very long sequences	Code, Web server
MxFold2 [78]	**(H)**. Thermodynamic and deep learning hybrid	Single sequence (FASTA, bpseq)	Integrates thermodynamic parameters with CNN, BiLSTM to leverage the power of deep models	Code, Web server
RNAalifold [79]	**(H)**. MFE and covariation	Alignment	In contrast to Pfold, uses free energies as parameters; modifies the scoring scheme of conventional MFE based dynamic programming algorithm	Code, Web server
CentroidFold [80]	**(H)**. γ-centroid estimator	Single sequence (FASTA); multiple alignment (CLUSTAL)	Sequence length limits to 400	Code, Web server
RNAErnie [81]	**(H)**. Pre-trained, foundational model using the transformer architecture	Single sequence (FASTA)	Masks tokens on various semantic levels (e.g. RNA motifs) to encode richer biological information	Code
RNA-FM [14]	**(H)**. Pre-trained, foundational model using the transformer architecture	Single sequence (FASTA)	First to utilize a foundation model or pre-training approach	Code

Note: Methods are sorted by their categories and alphabet order. **E**: energy-based methods. **C**: comparative methods. **L**: learning-based methods. **H**: hybrid methods. MFE: Minimum free energy. SCFG: Stochastic Context-Free Grammar. PSSM: Position Specific Score Matrix. DCA: Direct Coupling Analysis

PFOLD & PPFOLD PFOLD is an RNA secondary structure predicting program that employs a stochastic context-free grammar [82,83]. As mentioned above, an SCFG is a probabilistic model that uses probabilities as parameters to measure the co-varying tendency of position pairs. The co-varying tendency assumes compensatory mutations at paired positions occur in a correlated way. Since the function of RNA sequences largely depends on their structures, evolutionarily related RNAs that exert similar functions are very likely to have similar structures. Thus, those highly co-varying positions across a set of evolutionarily related RNAs would maintain the structure (i.e. base pairing) even though the sequence similarity may be low. PFOLD takes as input a multiple sequence alignment that contains target RNA homologous sequences in fasta format and predicts the consensus secondary structure of the alignment using the so called KH-99 algorithm, named after PFOLD's authors. The KH-99 algorithm essentially couples a phylogenetic model calculated from the alignment using Felsenstain maximum likelihood algorithm [84] with the SCFG, and it finds the most likely RSS using the CYK dynamic programming algorithm derived from natural language processing [85]. An updated implementation of PFOLD called PPFOLD was later developed [68] with Java multithreaded computation to accelerate the SCFG and phylogenetic calculations, which demonstrates much improved scalability. In addition to PFOLD's original output, PPFOLD generates a symmetric base-pairing probability matrix so that for each position, the probability of it being base-paired is computed as the average probability of all pairs that contain it.
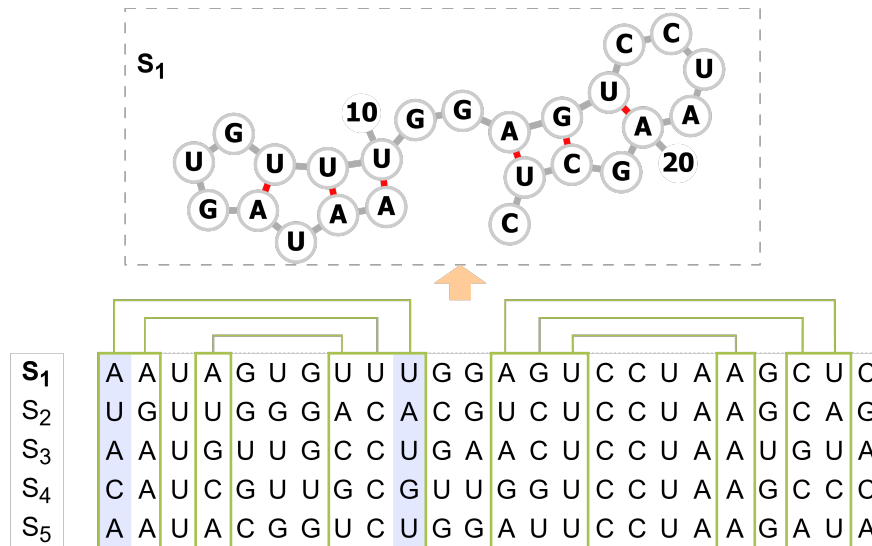
Figure 3: **A schematic illustration of comparative RNA secondary structure prediction strategy.** Assuming the functionally important structures (not necessarily the sequences) are conserved through evolution, the idea is to first align equivalent sequences from related species ($S_1, .., S_5$), then find pairs of co-varying alignment columns (green), finally incorporate the co-varying information into the predicted structure.

RNADECODER RNADECODER[70,71] is a comparative method for making predictions of RNA secondary structure. It also employs an SCFG model but is more complex: it is explicitly designed to take protein-coding context into account. So for the input, besides the sequence alignment, the protein-coding annotation of the input mRNA alignment (specifying codon positions), as well as a phylogenetic tree for those sequences in the alignment, is also required. RNADECODER is able to distinguish between loop/bulge region and un-paired region outside of the RNA structure and does not assume global RNA structure. RNADECODER has two modes: (1) in scanning mode, it scans for posterior probabilities of a position being loop/bulge and unstructured in the alignment; (2) in folding mode, it predicts the RNA structure and explicitly labels stem-pairing, loop/bulge and unstructured positions. In contrast to the other programs, for a given position, RNADECODER computes its probabilities for being loop/bulge and unstructured. So, the sum of these two probabilities is the non-base-pairing probability for that position.

TORNADO TORNADO[64] is a flexible SCFG-based approach and programming language itself designed for RSS prediction, which offers greater adaptability compared to the other SCFG method. It accepts RNA sequences in Stockholm or FASTA formats as input and outputs predicted secondary structures in dot-bracket notation. The method employs dynamic programming algorithms, such as CYK or posterior decoding, to predict structures and uses maximum likelihood optimization to train SCFG parameters from sequence-structure pairs. Its flexibility enables the modeling of complex structural elements, including base-pair stacking and loop dependencies, with SCFG models like the one used by PFOLD[82]. While this adaptability makes TORNADO capable of addressing a wide range of RNA modeling tasks, it still has a cubic time complexity $O(KN^3)$ for a sequence of length N and a design-dependent constant K, which can be a limitation for long sequences or grammars with high complexity.

2.3 Advanced machine learning and deep learning based methods

Since the classic methods such as the Zuker algorithm faced limitations due to the complexity and incompleteness of experimentally determined free energy parameters etc., machine learning and more recently deep learning methods have raised as powerful alternatives by leveraging large datasets of RNA sequences and their structures to provide pure data-driven, supervised-learning solutions. These advanced learning-based methods have introduced sophisticated model parameterization and training frameworks, bypassing explicit thermodynamic or co-variation assumptions, and as a result, they significantly improved prediction accuracy in cases like pseudoknot and long-range basepair

interactions etc.

CONTRAFOLD CONTRAFOLD[73] is a machine learning-based method for RNA secondary structure prediction that employs a discriminative probabilistic model, conditional log-linear model (CLLM), instead of a SCFG-like generative model or relying solely on free energy minimization. It learns model parameters directly from the training data but resembles an MFE model for its CLLM construction. Consequently, CONTRAfold aims to find the maximum-expected accuracy structure, which is the probabilistic counterpart of the MFE structure in energy-based models. The method takes a single RNA sequence as input, and like many other methods, it uses a dynamic programming algorithm and has a time complexity $O(n^3)$. As a result, the input RNA sequence is normally constrained under a practical upper limit (e.g., 1,000 as instructed on its web server) to maintain reasonable run times and memory usage. By incorporating benefits from both probabilistic and thermodynamic models, CONTRAfold demonstrates competitive or superior accuracy compared to purely SCFG-based methods such as Pfold and purely energy-based methods such as ViennaRNA.

SPOT-RNA SPOT-RNA[12] is a deep learning-based RNA secondary structure prediction method that utilizes an ensemble of residual networks (ResNets)[86], two-dimensional bidirectional long short-term memory (2D-BLSTM)[87,88] modules, and dilated convolutional neural networks (CNNs)[89]. It also leverages transfer learning to integrate information from a large dataset containing low-quality structure labels at a single base-pair level with a small but high-resolution training set. The model with transfer learning reduced the risk of overfitting and showed remarkable performance generalization across diverse datasets. Moreover, rather than relying on accurate thermodynamic parameters, SPOT-RNA adopts a pure machine-learning strategy so that all base pairs can be trained and predicted, regardless of whether it is associated with local or nonlocal (tertiary) interactions. As a result, SPOT-RNA can predict base pairs beyond the canonical Watson–Crick interactions, including noncanonical and pseudoknotted (non-nested) configurations, which are often challenging for traditional prediction methods.

SPOT-RNA2 SPOT-RNA2[76] is an extension builds upon SPOT-RNA[12] by incorporating evolutionary information and refined neural architectures. Specifically, it uses an ensemble of dilated CNNs only to simplify the model architecture with faster computations for long-range interactions, and it reuses the transfer learning step to leverage the information in both the low-resolution and high-resolution structure data. In addition to the one-hot encoding and predicted base-pair probabilities from single sequences, SPOT-RNA2 integrates PSSMs (Position-Specific Scoring Matrices) and DCA (Direct Coupling Analysis) features, enabling it to capture evolutionary information as a supplement to improve prediction accuracy. By incorporating these heterogeneous inputs, SPOT-RNA2 outperforms the single-sequence-based SPOT-RNA, especially on highly homologous sequences, and enhances the ability to infer structurally complex patterns such as noncanonical pairs more effectively.

E2EFOLD E2EFOLD[74] is an end-to-end deep learning model for RNA secondary structure prediction, designed to address challenges in dealing with complex RNA structures, particularly pseudoknots. Unlike traditional approaches that rely on energy minimization and dynamic programming, it directly predicts the RNA base-pairing matrix while integrating hard constraints, through a deep architecture combining a transformer-based Deep Score Network for sequence representation and a multilayer Post-Processing Network for constraints on legit base-pairing types. Specifically, the post-processing network employs an unrolled algorithm that utilizes a primal-dual constrained optimization to ensure the base-pairing constraints are enforced to reduce the space of valid structures and mitigate overfitting. As a result, E2Efold achieved state-of-the-art performance at then, on benchmark datasets (we will describe in later section 5) including RNAStralign and ArchiveII, significantly improving prediction accuracy for both nested and pseudoknot structures. A notable related work later introduced E2EFOLD-3D which is an extension of E2EFOLD for the de novo RNA tertiary structure prediction[90].

UFOLD UFOLD[13] uses a deep CNN architecture called U-Net[91] to generate RNA secondary structure prediction. UFold provides both a web server and a local software option. It accepts multiple RNA sequences in FASTA format as input and outputs the predicted secondary structure in dot-bracket notation. Regarding its architecture, UFold converts the original sequence into an image of size $17 \times L \times L$, where L is the length of the RNA sequence, using this image as the input to a U-Net architecture to generate a predicted contact map. The 17 can be thought of as 17 different channels, 16 types of unique Waston-Crick base-pairing, and an extra channel used in CDPFold[92] to deal with

sparsity. UFold's novelty came from converting raw RNA sequences into an "image" representation, which allows for long-range contact predictions, a fully convolutional framework, and highly efficient parallel computability.

CNNFOLD CNNFold[72] utilizes a simple yet effective CNN architecture. The method encodes RNA sequences of length L into a two-dimensional $L \times L$ map with eight channels, where each channel represents specific base-pairing relationships or structural constraints. Six channels capture possible base pairings (e.g., A-U, G-C), one channel indicates unpaired bases along the main diagonal, and another flags invalid pairings due to constraints like short distances or incompatible bases. This representation facilitates the prediction of both local and long-distance pairings, as well as structural motifs like stems. The output of the model is an $L \times L$ base-pair scoring matrix. To ensure the validity of predicted structures, CNNFold modifies the Blossom algorithm to handle self-loops, which is crucial for representing and predicting structures like pseudoknots. While CNNFold doesn't particularly stand out in complexity, it serves as a notable example of a biologically informed architecture and its distinct post-processing techniques.

REDFOLD REDFOLD[75] is another deep learning-based model that is unique in using an encoder-decoder network incorporating ResNet[93] and FC-DenseNet[94] networks. It provides both a web server and the source code for the software. As input, it takes FASTA formatted RNA sequence files, and for output, it outputs in the dot-bracket notation. REDfold, much like UFold, first generates the RNA sequence into two-dimensional contact matrices by trying all possible combinations of Waston-Crick base pairing. These matrices are fed into the network via feature mapping and basic convolution modules (BCMs) consisting of 2-dimensional convolution, batch-normalization, and rectified linear unit (ReLU). They also introduce a dense connected module (DCM), which is made up of a series of BCMs to avoid bottlenecks from the encoding steps. The decoder network also consists of DCMs that ultimately transition up to generate a scoring matrix of size $L \times L$ where L is the length of the RNA sequence. REDfold claims that the incorporation of ResNet and FC-DenseNet networks in their encoder-decoder networks makes the process much more efficient and effective, producing highly accurate predictions.

2.4 Hybrid methods

Despite the recent advances in the class of learning-based methods, its reliance on large and unbiased datasets highlights challenges such as overfitting and generalization to complex structural patterns. To overcome these challenges, efforts from the wet-lab have been dedicated to improving both the quantity and quality of RNA structural data, while computationally, exploring less data-hungry approaches, such as hybrid methods that integrate constraints to the deep model space as well as transfer learning from pre-trained general foundation models, seems to offer a promising path.

RNAALIFOLD RNAALIFOLD[79] is a program in the ViennaRNA package, representing an early work adopting hybrid strategy to combine energy and comparative approaches. As the name suggests, it computes the minimum energy structure for a set of *aligned* input sequences. In contrast to PFOLD, RNAALIFOLD uses free energies rather than probabilities as parameters. It combines the co-varying information from the fixed alignment with the minimum free energy model by modifying the scoring scheme of the dynamic programming algorithm used in conventional thermodynamic methods. It produces the consensus minimum free energy structure in dot-bracket notation and a dot plot of the symmetric base-pairing probability matrix.

CENTROIDFOLD CENTROIDFOLD is based on the γ-centroid estimator[95] for high-dimensional discrete spaces, which is generally more accurate than an maximum expected accuracy estimator (e.g. the one used in CONTRAfold[73]) under the same probability distribution. CENTROIDFOLD supports multiple probabilistic models, including the CONTRAfold model, the McCaskill model (from the ViennaRNA package), the RNAalifold model, and the Pfold model. Benchmarks indicate that CentroidFold with the McCaskill model using Boltzmann likelihood parameters[96] achieves the most accurate predictions. CENTROIDFOLD accepts RNA sequences in FASTA format as input, producing the predicted secondary structure.

MXFOLD In order to deal with overfitting issues in learning-based RSS prediction methods, Akiyama et al. introduced a novel method MXFOLD, which integrates thermodynamic information and machine learning models together. It accepts RNA sequences in FASTA or bpseq format as input and outputs predicted secondary structures through a Zuker-style dynamic programming (DP) algorithm[54]. The DP algorithm predicts an optimal secondary structure that maximizes the sum of scores from a hybrid model. The model combines the Turner's nearest-neighbor parameters[51,58],

which are experimentally determined free-energy parameters, with a fine-grained scoring model based on structured support vector machines (SSVMs). This hybrid model trains weights of SSVMs for complex structural features, such as specific loop configurations and base-pair stacking. By combining thermodynamic parameters with machine-learned scores, the method ensures both accuracy and robustness, particularly for unobserved substructures. For example, given a sequence, MXFOLD decodes the structure using a scoring function:

$$f(x, y) = f_T(x, y) + f_W(x, y) \tag{1}$$

where f_T encapsulates thermodynamic contributions, and f_W reflects the machine learning-based refinements. This hybrid framework demonstrates superior prediction performance while mitigating risks of overfitting, thereby advancing RNA secondary structure prediction.

MXFOLD2 Inspired by the success of deep learning in biological sequence analysis, Sato et al. further extended MXFOLD[77] with deep neural networks replacing the SSVMs in the model part and developed MXFOLD2[78]. MXFOLD2 adopts a combination of CNN and BiLSTM layers for the network architecture to learn four different types of folding parameters: helix stacking, helix opening, helix closing, and unaired region scores. These parameters are then fed into a Zuker-style dynamic programming algorithm[54], just like the Turner nearest neighbor parameters, to calculate the final score of an RNA secondary structure. Similar to MXFOLD, MXFOLD2 integrates the thermodynamic information to reduce overfitting by separately computing minimum free energy scores through the same DP algorithm as well and added to the deep neural network scores as a form of regularization. As a result, MXfold2 aligns deep learning folding scores with free energy folding scores and achieves superior performance compared to several traditional and DNN-based models, including MXfold, across sequence-wise and family-wise testing datasets.

RNA-FM RNA-FM[14] represents one of the first approaches to integrate pre-training. Unlike previous deep learning-based strategies, which rely on labeled data specific to secondary structure, RNA-FM leverages the vast pool of unannotated RNA sequence data through self-supervised learning. Its architecture is based on a transformer model comprising 12 bidirectional encoder layers, pre-trained on 23 million RNA sequences from RNAcentral by reconstructing masked tokens. This approach enables RNA-FM to learn rich, task-agnostic representations of RNA sequences, capturing implicit structural and evolutionary patterns without requiring labels. These embeddings are then fine-tuned for downstream structure-related and function-related applications, offering flexibility across diverse RNA prediction tasks. RNA-FM's advantages are more pronounced in structural tasks, like secondary structure prediction, likely due to differences between its training data and the datasets used for functional tasks, as well as the inherent complexity of RNA structure-function relationships.

RNAERNIE RNAErnie[81] is a recent foundational model, also pre-trained on RNAcentral data, that distinguishes itself through the integration of motif-aware pretraining strategies and a type-guided fine-tuning mechanism. RNAErnie features 12 multilayer transformer blocks with a hidden state dimension of 768. While RNA-FM had base-level masking alone, RNAErnie's pretraining phase incorporates a motif-aware multilevel masking strategy, which includes base-level, subsequence-level, and motif-level masking. This approach enriches RNA representations by capturing both fundamental sequence patterns and biologically significant motifs, such as those derived from databases like ATtRACT and SpliceAid. Additionally, RNAErnie tokenizes coarse-grained RNA types (e.g., miRNA, lnRNA) as special vocabularies, appending them to RNA sequences during pretraining to improve domain adaptation and enhance the model's understanding of RNA-specific features.

For the downstream task, RNAErnie employs a type-guided fine-tuning strategy, which predicts RNA types from sequence embeddings and incorporates these as auxiliary inputs into task-specific modules. Their study examines multiple architectures: FBTH (frozen backbone with trainable task-specific head), TBTH (trainable backbone and head for end-to-end learning), and, novelly, STACK (ensemble learning with type-guided parallel modules). For secondary structure prediction, RNAErnie combines its embeddings with a Zuker-style dynamic programming approach like MXFOLD, predicting RNA secondary structure by maximizing the cumulative scores of adjacent loops. Fine-tuning is performed using a max-margin framework, minimizing structured hinge loss with thermodynamic regularization.

3 Computational tools for RNA modification prediction

This section reviews recent tools for several common RNA modification types, including 2'-O-methylation (Nm or 2OM), N4-acetylcytosine (ac4C), 5-methylcytosine (m5C), N6-methyladenosin (m6A), N7-methylguanosine (m7G), and multiple widely occurring transcriptome modifications. Table 3 summarizes these tools, with detailed descriptions provided below.

Table 3: A summary of representative RNA modification prediction methods included in the review.

	Method	Input	Other notes	URLs
H2Opred [30]	Hybrid deep learning with multi-feature fusion	Multiple sequences or file in FASTA format	No more than 500 sequences per submission	Web server
Meta-2OM [31]	Multi-classifier meta-learning	Multiple sequences or file in FASTA format		Code, Web server
Nmix [32]	Hybrid deep learning with multi-feature fusion and ensemble learning	Multiple sequences or file in FASTA format	Up to 5000 sequences per submission	Code, Web server
ac4C-AFL [33]	Adaptive feature representation learning	Multiple sequences or file in FASTA format	No more than 20 sequences per submission	Web server
Voting-ac4C [34]	Pre-trained large RNA language model and ensemble learning	Multiple sequences or file in FASTA format		Web server
iRNA-ac4C [35]	Machine learning, minimum-Redundancy-Maximum-Relevance combined with incremental feature selection strategies	Multiple sequences in FASTA format	Sequence length must be 201 nt	Web server
TransAC4C [36]	Transformer-based encoder and Bi-LSTM networks combined with 1D CNN	Multiple sequences or file in FASTA format	Select either 415 nt or 21 nt	Code
DeepM5C [37]	Hybrid deep learning	CSV format file		Code
MLm5C [38]	A combination of hybrid machine learning models	Multiple sequences or file in FASTA format		Code, Web server
m5C-pred [39]	XGBoost framework with feature selection	Multiple sequences in FASTA format	Select the species for prediction	Code, Web server
MST-m6A [40]	Multi-scale transformer-based framework	TSV format file	Sequence length must be 201 nt	Code
CLSM6A [97]	Interpretable deep learning-based approach	Multiple sequences or file in FASTA format	Sequence length must be 201 nt	Code, Web server
BLAM6A-Merge [41]	Attention mechanisms with multimodal feature fusion and Blastn tool	FASTA format file	Required Blastn of version 2.14.0	Code
Moss-m7G [42]	Motif-based interpretable deep learning	FASTA format file		Code
THRONE [43]	Three-layer ensemble learning	Multiple sequences or file in FASTA format		Web server
MultiRM [25]	Attention-based multi-label neural networks	Single RNA sequence in string format	Minimum length of 51 nt	Code
TransRNAm [27]	Transformer-based encoder and CNN			Code
CIL-RNA [26]	Transformer-based encoder and Bi-GRU network with class incremental learning	CSV format file		Code

Note: Bi-GRU: Bidirectional gated recurrent unit. Bi-LSTM: Bidirectional long short-term memory. CNN: Convolutional neural networks. nt: nucleotides. These tools are grouped based on the method category.

3.1 2'-O-methylation (Nm or 2OM)

H2OPRED H2OPRED represents the first hybrid deep learning framework developed for the identification of 2OM in human RNA. This framework utilizes a combination of stacked one-dimensional convolutional neural networks (1D CNN) and attention-based bidirectional gated recurrent unit (BiGRU) modules to effectively capture both spatial and temporal information derived from conventional feature descriptors and embeddings based on natural language processing (NLP). The resultant high-level feature representations are subsequently integrated to facilitate the final classification of nucleotide modifications, specifically Am, Cm, Gm, Um, or Nm. H2OPRED thus accommodates both nucleotide-specific and generic 2OM modification types. The associated web server accepts input in FASTA format, providing users with probabilistic scores and class labels for the corresponding sequences. Additionally, users have the option to upload a FASTA file to the web server to execute predictions and retrieve results.

META-2OM META-2OM is a multi-classifier meta-learning approach that integrates eight distinct machine learning classifiers with eighteen different feature encoding algorithms, all coordinated by a meta-learner, to identify 2OM in human RNA. Notably, probabilistic features from 144 baseline models were generated and subsequently utilized to train a logistic regression model for the final classification task. This tool is available as both a web server and a downloadable codebase, accommodating the analysis of multiple sequences or file uploads in FASTA format. Upon completion of the submission process, the system returns probabilistic scores and class labels corresponding to the input sequences.

NMIX NMIX is a hybrid deep learning framework developed for the identification of 2OM sites in human RNA. Initially, one-hot, Z-curve, and RSS encodings were extracted from the RNA sequences. Subsequently, 1D and 2D CNN were designed, incorporating multi-head self-attention and residual connection modules to extract multi-dimensional features from the one-hot and Z-curve encodings as well as the RSS encoding. These feature representations were later fused through average pooling and concatenation for the purpose of final classification. Additionally, a Bayesian optimization-based technique was employed to construct an ensemble learning framework that effectively addresses the challenges presented by imbalanced datasets. Given an RNA sequence, NMIX outputs the nucleotide base, a probabilistic score, and a corresponding class label.

3.2 N4-acetylcytosine (ac4C)

AC4C-AFL AC4C-AFL is an adaptive feature representation learning framework designed for the identification of ac4C in human mRNA. Initially, a pre-analysis was conducted to determine the optimal sequence length for ac4C identification, leading to the conclusion that a length of 201 nucleotides (nt) is optimal. Subsequently, a novel ensemble feature importance scoring function was proposed to identify the optimal feature dimensions from sixteen sequence-derived feature descriptors, employing a sequence forward search strategy. Utilizing these optimal features, 176 single-feature best-performing models were constructed using eleven distinct machine learning algorithms, and their probabilistic features were generated to train the final model for ac4C identification. AC4C-AFL is publicly accessible, allowing users to input sequences in FASTA format or directly upload FASTA files to obtain predicted results, including probabilistic scores and class labels.

VOTING-AC4C VOTING-AC4C is the first framework that harnesses the capabilities of the pre-trained large RNA language model, RNAErnie[81], in conjunction with six conventional feature descriptors: one-hot encoding, encoding nucleic acid composition (ENAC), C2, nucleotide density (ND), trinucleotide composition profile (TPCP), and k-spaced nucleotide pair frequencies (KSNPF). This integration aims to enhance the prediction of RNA ac4C sites. A deep neural network (DNN) model was specifically designed for feature reduction and selection. Subsequently, a soft voting ensemble learning model was constructed by integrating eXtreme Gradient Boosting (XGB), CatBoost (CB), and multilayer perceptron (MLP) for the final prediction. Similar to other tools, VOTING-AC4C accepts multiple sequences in FASTA format for predicting ac4C sites.

TRANSAC4C TRANSAC4C is an interpretable framework that employs a transformer-based encoder to leverage the relationships between words in natural language sequences, translating these relationships into biological contexts for model interpretation. Notably, this study involved reconstructing a previous dataset to generate a new dataset characterized by varying sequence lengths, distinct species, and diverse RNA types, thereby facilitating a comprehensive

analysis of ac4C in RNA. Specifically, RNA sequences were tokenized using 3-mers and subsequently embedded as inputs for a transformer-based bidirectional long short-term memory (Bi-LSTM) module, which extracts contextual information. A 1D convolutional neural network (CNN) module was then designed to capture essential spatial information before processing the features through several fully connected (FC) layers for final classification. This tool is capable of predicting ac4C from multiple RNA sequences in FASTA format. However, it requires the selection of either 415 or 21 nt corresponding to the appropriate models.

IRNA-AC4C IRNA-AC4C is a machine learning-based predictor designed for the identification of ac4C in mRNA. A novel high-quality dataset was constructed to develop this ac4C prediction tool, utilizing the Gradient Boosting Decision Tree (GBDT) classifier along with optimal hybrid features. These optimal features were identified by linearly combining k-mer encoding, nucleotide chemical property encoding, and accumulated nucleotide frequency encoding, followed by the application of minimum redundancy maximum correlation (mRMR) and incremental feature selection (IFS) techniques to select the optimal feature dimensions. IRNA-AC4C offers a publicly accessible web server that supports the input of multiple sequences in FASTA format, specifically with a fixed length of 201 nt.

3.3   5-methylcytosine (m5C)

DEEPM5C DEEPM5C is a deep learning (DL)-based hybrid stacking tool developed for the prediction of m5C in the human genome. Initially, a novel benchmark dataset was constructed, and four distinct feature encodings were employed to extract relevant features, which included three conventional feature descriptors and a natural language processing (NLP)-based embedding known as word2vec. Subsequently, four DL-based classifiers and four machine learning (ML)-based classifiers were utilized to train a total of 32 baseline models. The probabilistic features generated from these models were then stacked to train the final model using a one-dimensional convolutional neural network (1D CNN). While this tool provides the source code, the instructions are brief. It supports input files in CSV format.

MLM5C MLM5C a hybrid ML-based model designed to identify m5C sites. This model combines four ML classifiers with eleven RNA sequence-derived conventional feature descriptors. Subsequently, 44 single-feature baseline models were generated and ranked based on their performance, with the probabilistic features of the top 20 models stacked to train the final predictive model for m5C identification. Although this tool demonstrates superior performance compared to state-of-the-art methods at the time, its approach resembles several publications from the same author group. MLM5C framework provides both a web server and source code for making predictions and facilitating local deployment. It supports the analysis of multiple sequences or the upload of files in FASTA format. Upon completing the submission job, the tool returns probabilistic scores and class labels for the input sequences.

M5C-PRED M5C-PRED is an ML-based framework that incorporates a feature selection strategy to predict m5C sites in five different species: Arabidopsis thaliana, Danio rerio, Drosophila melanogaster, Homo sapiens, and Mus musculus. This tool employs the XGB algorithm and utilizes five conventional feature descriptors, including the composition of k-spaced nucleic acid pairs (CKSNAP), enhanced nucleic acid composition (ENAC), label encoding (LE), nucleotide chemical properties (NCP), and electron-ion interaction pseudopotentials of trinucleotides (PseEIIP). Additionally, SHapley Additive exPlanations (SHAP) analysis is employed to identify the optimal features, which are then used to retrain the XGB for the final model. M5C-PRED framework offers both a web server and source code; however, it necessitates the selection of a specific species for making predictions.

3.4   N6-methyladenosin (m6A)

CLSM6A CLSM6A is an interpretable DL-based architecture developed for the prediction of N6-methyladenosine (m6A) modification sites across various cell lines and tissues in Homo sapiens. Specifically, RNA sequences are transformed into a 2D matrix using an ENAC-based encoding algorithm. A CNN module is then employed to extract and learn the spatial information from these features, which are subsequently input into an MLP for final classification. Additionally, both model-based and propagation-based methods are utilized to interpret the predictions made by the model. Furthermore, CLSM6A offers both a web server and publicly available source code for prediction and local deployment. However, the input sequence length must be precisely 201 nt.

MST-M6A MST-M6A is a multi-scale dual transformer-based architecture designed for the accurate identification of m6A modification sites across eight cell lines and three tissues in Homo sapiens. This framework employs a shared transformer architecture coupled with dual k-mer tokenization to exploit multi-scale feature representations from RNA sequences, thereby capturing global contextual information and enriching feature representations. These feature representations are subsequently fused using a channel feature fusion module and processed through three CNN layers before being input into an MLP module for final classification. Additionally, this tool provides publicly available source code to facilitate local deployment and ensure reproducibility of results.

BLAM6A-MERGE BLAM6A-MERGE is a tool designed for the identification of m6A modification sites across twelve benchmark datasets derived from six cell lines, including CD8T, Hek293_abacm, Hek293_sysy, HeLa, and MOLM13, and operates in two modes: full transcript and mature mRNA. Notably, BLAM6A-MERGE employs various attention-based mechanisms to extract multimodal features from RNA sequences. Subsequently, a stacking ensemble learning framework is utilized to integrate four specific classifiers derived from sequence data with a Blastn-based classifier, establishing a meta-learning approach for final classification. This tool provides source code for local deployment, facilitating the reproducibility of results. However, it is essential to note that using Blastn tool version 2.14.0 is required to generate the final predictions.

3.5 N7-methylguanosine (m7G)

MOSS-M7G MOSS-M7G is an interpretable DL-based method designed for the identification of m7G sites, utilizing word-detect and motif-based embedding within a transformer architecture. The word-detect module employs a 1D CNN to capture the motif probabilistic matrix derived from the one-hot encoding of RNA sequences. This motif probabilistic matrix is then transformed into motif-based embeddings and combined with an additional [CLS] token embedding, serving as input for the transformer architecture to capture high-level contextual information. Finally, the feature representation extracted from the transformer architecture through the [CLS] token is fed into an MLP module for final classification. This tool provides source code for making predictions and accepting input in FASTA format files.

THRONE THRONE is a three-step ensemble learning framework developed for the identification of m7G sites in human RNA. Initially, nine conventional feature descriptors and six machine learning (ML)-based classifiers were utilized to generate 54 single-feature ML-based baseline models. Subsequently, the probabilistic features extracted from these baseline models were concatenated into a 54-dimensional feature vector and trained using six ML-based classifiers. Finally, the probabilistic features from the six ML-based meta-models served as input for a six-dimensional feature vector, which was used to identify the best meta-learner model for final classification. Furthermore, THRONE offers a web server that facilitates predictions of m7G sites by allowing users to upload files or directly enter multiple sequences in FASTA format.

3.6 Multiple widely occurring transcriptome modifications

MULTIRM MULTIRM is an attention-based DL framework designed for the prediction of twelve prevalent transcriptome modifications, including m1A, m5C, m5U, m6A, m6Am, m7G, Ψ, A-to-I, and four types of 2OM modifications. Initially, one-hot encoding is employed to convert the RNA sequence into a matrix format. Subsequently, three embedding strategies are applied to extract features from the one-hot encoded input, specifically utilizing 1D CNN, Word2Vec, and Hidden Markov Models. These feature representations are then fused and processed through an attention-based LSTM network for multi-label classification to predict the twelve RNA modification types. Moreover, the attention weights and integrated gradients are utilized to identify sequence motifs corresponding to those identified by motif sequence-based tools for each RNA modification. Although MULTIRM offers both a web server and source code for predicting RNA modifications and facilitating local deployment, it is important to note that the web server is inactive at the time of this review.

TRANSRNAM TRANSRNAM is an interpretable transformer-based architecture designed for the identification of twelve common RNA modifications. Initially, Word2Vec is employed to convert RNA sequences into matrix representations. Subsequently, a transformer-based encoder is utilized to learn high-level contextual information from the features extracted via Word2Vec. These feature representations are then processed through a CNN block with a skip connection
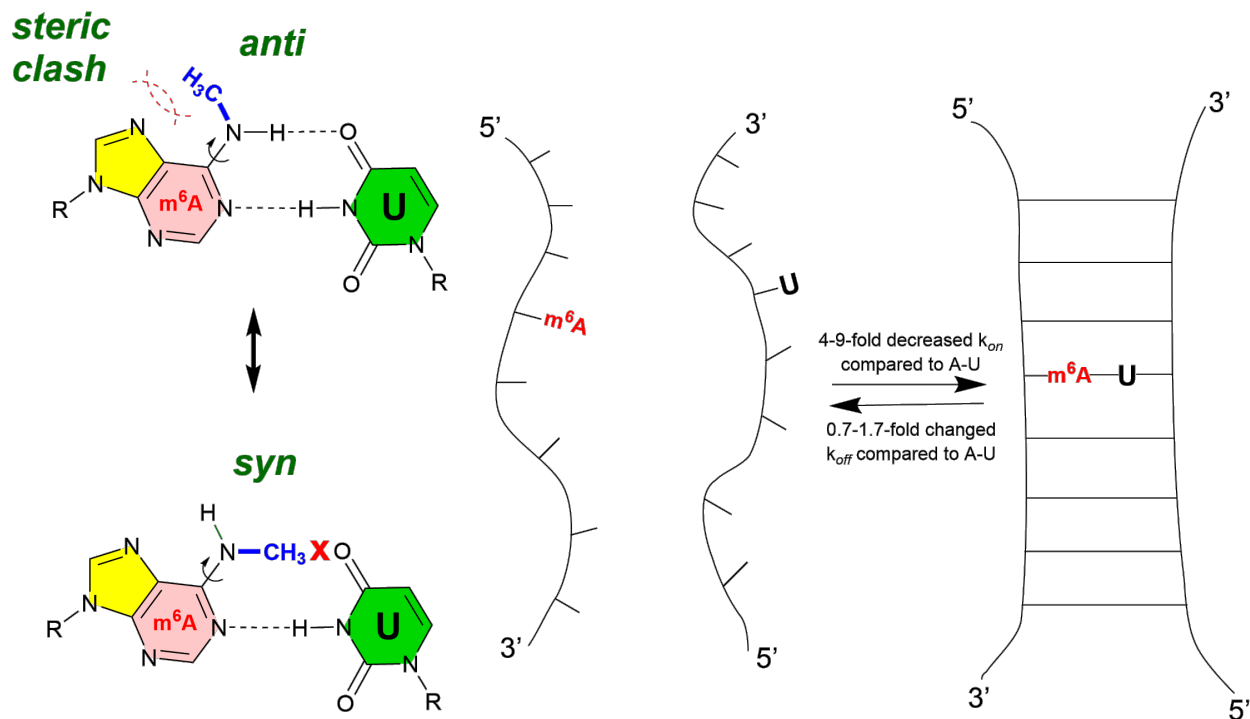
Figure 4: **An example of the interplay between RNA modification and secondary structures.** The m6A modification is used here to illustrate the effects of modifications on RNA secondary structures. The left plot shows that stable m6A:U basepairing is only feasible in the anti-conformation of the m6A base, which is energetically less favorable. The right plot shows the m6A:U base-pairing has a significant change on the annealing rate constant while the unpairing does not, when comparing to the normal A:U case. The figure is remade based on Hofler and Duss (2024)[98].

to capture spatial feature representations effectively. Finally, these spatial features are fed into twelve parallel FC networks to predict the twelve types of RNA modifications simultaneously. Furthermore, the attention weights of TRANSRNAM are extracted for model interpretation purposes. Although TRANSRNAM provides its implementation through a GitHub repository, it is noteworthy that there is no associated web server for online predictions.

CIL-RNA CIL-RNA is a class incremental learning framework designed to predict multiple types of RNA modifications. This framework employs a baseline classifier integrated with a transformer-based encoder, followed by a Bi-GRU and an MLP for final predictions. Specifically, it utilizes four incremental learning strategies to train the baseline classifier: parameter regularization, function regularization, replay, and template-based classification. Notably, CIL-RNA can be extended to predict new categories of RNA modification sites without retraining on previous data, thereby enhancing computational efficiency. To facilitate local deployment and ensure reproducibility, CIL-RNA provides its implementation via a GitHub repository and supports input files in CSV format.

4 The interplay between RNA secondary structure and RNA modification

4.1 RSS motifs aid RNA modifications prediction

RNA structure plays a crucial role in RNA modifications, particularly in the m6A modification[98]. The folding and structural characteristics of RNA influence not only how and where these modifications are applied but also their interactions with proteins. Such RNA binding proteins reply on specific RSS patterns to bind to target RNA and initiate RNA modifications. For example, the m6A writer complex (METTL3/METTL14) is recruited to RNA based on these structural features, enabling site-specific and co-transcriptional installation of the m6A modification. This

recruitment is facilitated by interactions with histone marks and the C-terminus of RNA Polymerase II, ensuring that the modification occurs precisely at the required locations during transcription. Specifically, the METTL3/METTL14 complex preferentially recognizes and modifies RNA sequences that contain specific motifs, particularly the DRACH motif (D = A/G/U; R = A/G; H = A/C/U). This sequence specificity ensures that the complex accurately targets the most relevant RNA substrates for modification.

Recently, researchers have integrated RNA secondary structure-based feature encodings to improve the performance of RNA modifications [99]. For instance, Xiang *et al.* [100] utilized RNAfold (reviewed above) to fold the 101-bp mRNA fragment, yielding an MFE (minimum free energy) value. Then, it was combined with conventional feature descriptors to train a support vector machine (SVM) classifier to identify mRNA m6A sites. Similarly, Geng *et al.* [32] also utilized RNAfold to generate structural expressions for each sequence. Using an ensemble learning approach, they combined it with one-hot encoding and Z-curve theory to enhance the prediction of 2'-O-methylation sites. These findings suggest that RNA secondary structure is crucial for understanding RNA's biological functions and properties, leading to improved performance of RNA modifications.

4.2 RNA modifications affect secondary structure formation

RNA modifications, such as N6-methyladenosine (m6A), have a profound impact on RNA secondary structure prediction by altering the chemical and physical properties of RNA molecules [98]. Methylation of nucleotides modifies their characteristics, influencing RNA structure and interactions with cellular partners. These modifications can either promote or hinder the formation and functionality of protein-RNA complexes, as well as alter the base-pairing kinetics of RNA. For instance, the addition of a methyl group to the nitrogen position of adenosine affects the stability of base pairs, consequently slowing the rate of duplex formation. This alteration may lead to local destabilization of RNA structures, which standard secondary structure prediction algorithms often overlook.

In the study conducted by Liu *et al.* [101], it was demonstrated that m6A plays a crucial role in regulating RNA-protein interactions by modulating the accessibility of RNA-binding motifs (RBMs). This modification can reshape the local structure of mRNA and long non-coding RNA (lncRNA), thereby facilitating the binding of RNA-binding proteins such as heterogeneous nuclear ribonucleoprotein C (HNRNPC). The presence of m6A enhances the binding affinity of these proteins, influencing essential processes such as pre-mRNA processing, gene expression, and RNA maturation. This phenomenon, often referred to as the "m6A-switch," describes how m6A-dependent structural remodeling of RNA regulates interactions between RNA and proteins, allowing for effective access to binding sites that are vital for various biological functions.

Moreover, Lewi *et al.* provided evidence supporting the notion that both RNA structures and RNA modifications collaboratively shape RNA–protein interactions [19]. Specifically, m6A has been shown to destabilize stem structures, which enhances the accessibility of RNA-binding proteins to their binding sites. This dynamic interplay between RNA modifications and structure is crucial for regulating various aspects of gene expression, including mRNA stability, splicing, and translation efficiency. The structural changes induced by RNA modifications not only facilitate the recruitment of specific RNA-binding proteins (RBPs) but also influence the fate of mRNAs, such as their translation or decay, contributing to the overall regulation of gene expression.

Additionally, Tanzer *et al.* asserted that RNA modifications can significantly impact RNA structures by either stabilizing or destabilizing base pairs [16]. For example, A-to-I editing can destabilize double-stranded RNAs (dsRNAs) by converting A-U pairs into less stable I-U pairs, which increases flexibility and potential for refolding. Conversely, this editing can create stabilizing I-C pairs that enhance hybridization stability in certain contexts. Furthermore, m6A modification is known to weaken RNA structures and is particularly abundant in 3' UTRs, influencing mRNA preprocessing events such as splicing and polyadenylation. Overall, RNA modifications dynamically reshape the structural landscape of RNA, enabling diverse functional interactions and responses, while advanced probing techniques like PARS and SHAPEseq continue to elucidate the global impact of these modifications.

Brümmer *et al.* found that A-to-I editing significantly enhances RNA secondary structures by reducing the accessibility of microRNA (miRNA) target sites [102]. This stabilization arises from the incorporation of inosine, which modifies the thermodynamic properties of RNA, leading to a more compact structure. Consequently, edited mRNAs exhibit reduced accessibility to Argonaute 2 (AGO2)-miRNAs, which typically bind to and destabilize unedited mRNAs. Importantly,

A-to-I editing does not substantially alter the sequences of miRNA target sites; rather, it influences their accessibility through structural modifications. Experimental validation has shown that edited transcripts display higher expression levels than their unedited counterparts, underscoring the critical role of A-to-I editing in regulating mRNA stability and abundance through modulation of RNA secondary structures.

Furthermore, Boo and Kim recently emphasized the emerging role of RNA modifications in the regulation of mRNA stability, including m6A, N6,2'-O-dimethyladenosine (m6Am), 8-oxo-7,8-dihydroguanosine (8-oxoG), pseudouridine (Ψ), 5-methylcytidine (m5C), and N4-acetylcytidine (ac4C)[20]. Modifications such as m6A and its derivatives can either enhance or diminish mRNA stability, thereby affecting translation efficiency and degradation rates. These modifications can potentially alter the secondary and tertiary structures of mRNA, impacting the accessibility of RNA-binding proteins involved in mRNA surveillance and decay pathways. Consequently, specific RNA modifications can lead to either the stabilization or destabilization of mRNA, ultimately influencing protein synthesis levels and the overall expression of genes.

4.3 RSS prediction with energy parameters for modified nucleotides

As introduced earlier, the change in free energy associated with RSS folding is estimated using a set of empirical parameters from the nearest neighbor model, which is derived from optical melting experiments conducted on model systems in the wet lab. Given the impact of RNA modifications on folding stability and the prevalence of m6A, Kierzek et al.[18] constructed a dedicated dataset that consists of a complete set of all nearest-neighbor parameters incorporating m6A in order to make modified RNA secondary structure prediction (Figure 2(**B**) shows an example usage of this set). Similar to the Turner nearest neighbor parameters[51,58], Kierzek et al. developed a full set of thermodynamic parameters for m6A as well as normal ACUG under the nearest neighbor free energy model. Specifically, they estimated corresponding free energy changes at 37°C with optical melting experiments of synthesized oligonucleotides. After obtaining these new nearest neighbor parameters for m6A, they extended RNASTRUCTURE[60] that we discussed earlier to incorporate m6A into the alphabet and use the new parameters to make RSS predictions. This work showed promise in the accurate modeling of m6A-modified RNAs. Notably, the authors also reported transcriptome-wide predictions with m6A, showing that methylation reduces the probability of adenosine being buried in helices (i.e., 21% for A and 13% for m6A), potentially driving widespread structural changes that influence RNA-protein interactions. Besides, the NNDB database[103,53] (more details in the datasets section next) has also collected this set of nearest-neighbor parameters for RNAs with m6A modifications.

In a follow-up study, Szabat et al. conducted around 100 optical melting experiments in addition, focusing on m6A versus normal adenine to test and refine the corresponding free energy nearest neighbor parameters[104]. Specifically, they utilized the RRACH motif, which is a known consensus motif of N^6-methylation in mammalian cells, with R representing a purine and H representing one of {A, C, U}. They experimented with the central site of RRACH, with or without methylation, in various secondary structure contexts, including helices, bulges, internal loops, dangling ends, and terminal mismatches. With the m6A-expanded-nearest-neighbor parameters, the authors estimated the folding free energy changes, i.e., the folding stability, and compared them to the measured values from melting experiments. As a result, the overall root mean squared deviation (RMSD) between experimental and predicted free energy changes across all experiments was 0.67 kcal/mol, indicating robust accuracy for the m6A-expanded parameters. Moreover, the agreements between experimental and predicted folding free energy change with m6A were similar to those with normal A for most structural contexts. The authors also revised the original parameters[18] under each structural context to further calibrate. This work validates the potential RSS prediction capability for modified RNAs and provides a foundation to expand our understanding of m6A's roles in epitranscriptomic and gene regulations. A later version update of RNAstructure also includes these revised parameters to its command line release to expand its functionality[61].

Furthermore, as the most widely used RSS prediction tool, the ViennaRNA package also recently included support for modified RNA bases, starting from its version 2.6.0 update[105]. Instead of concentrating on one type of modification or experiments from one study, a comprehensive search was performed to identify available energy parameters from existing literature, which covers parameter sets like the m6A one discussed above[18,104] and many more. In total, the authors collected six different types of modifications from a number of experiments, including 7-deaza-adenonsine (7DA)[106], inosine[107,108], pseudouridine[109], non-standard purine nucleotide nebularine[110], dihydrouridine[111,112], as

well as m6A[18]. Then, the ViennaRNA package adopts a different strategy to accommodate for the effects of these modified bases by utilizing a hard- and soft-constraints framework[113] to modify upon the RSS energy computations made with normal Turner's nearest neighbor parameters, only when there is a modified nucleotide and its corresponding parameters are available. This strategy waives the need for a huge multi-dimensional lookup table that is typically required to store a complete set of nearest neighbor parameters covering all possible cases, which is more memory efficient and better aligned with the sparse nature of the energy parameter data for modified bases. Existing programs in the ViennaRNA package, like RNAfold[56] etc., have been extended with this functionality. At the time of writing, this work presents the largest set of RNA modification energy parameters for secondary structure prediction.

5 Datasets

Table 4: A summary of representative datasets discussed in this review.

	Type	Data description	URLs
ArchiveII[114]	RSS	Contains 3 975 seqeunces and structures ranging from 10 RNA families.	Website
bpRNA-1m[115]	RSS	Contains 102 318 sequences and structures of approximately 2500 RNA families, mainly collected from Rfam 12.2. Does not remove redundant sequences, but a processed subset, bpRNA-1m(90), does.	Website
Rfam[116]	RSS	A comprehensive collection consisting of 90 190 RNA sequences of >4000 RNA families and all RNA types. Various sequence lengths from <20 short miRNAs to several thousands-nt-long lncRNAs.	Website
RNAStralign[117]	RSS	Contains 30 451 sequences from 8 RNA families. Sequence length ranges from 30 to 1800+.	Website
TORNADO data[64]		In total, contains 5 387 sequences. Contains two different collections A (sourced from literature, 11 families) and B (from Rfam, 22 families), each further splitting to TrainSet and TestSet. RNAs in A have longer lengths (10 to 700+) than in B (27 to 200+), <70% sequence identity, and different structure distributions.	Data link
RNAcentral[118]	RSS	Contains comprehensive information about non-coding RNAs, including RNA sequences and structures, mRNA interactions, and RNA family classifications.	Website
NNDB[103,53]	Thermodynamic Parameters	Contains nearest neighbor parameters for normal and m6A modified base as well as DNA folding parameters, in the form of free energy changes, which is used for RSS predictions.	Data link
PDB[119]	3D structure	Provides RNA 3D structures, able to be used as a RNA folding benchmark, as was the case with Szikszai et al.[120].	Website
MODOMICS[17]	RNA Modification	Contains information about chemical structures, biological pathways, and sequences, enzymes of RNA modifications. Currently has more than 170 different RNA modifications, 429 different RNA modified residues (335 natural ones), and 1925 RNA sequences.	Website
ENCORI[121]	CLIP	Contains information about RNA-RNA and Protein-RNA interactions from CLIP-Seq data. Also has analysis of impact of interactions on gene expression in 32 cancer types.	Website
eCLIP data[122]	CLIP	Using the eCLIP protocol, currently has 119 RBPs and 102 RBPs interaction with RNA in the K562 and HepG2 cell lines respectively.	Website

Note: RNAStralign dataset is also commonly available in several follow-up studies with different versions (e.g., MXFOLD2 data, E2EFOLD data, HuggingFace).

5.1 RSS datasets

ArchiveII ArchiveII [114] is an RNA secondary structure dataset compiled by Mathews Lab that is commonly used for the training and testing of RSS prediction methods containing a total of 3,975 RNA sequences and their structures. This dataset is an expansion of the previous RNA secondary structure dataset, also compiled by Mathews et al. [123], updating and including new structures. It contains the sequences and structures of 10 RNA families, such as small subunit ribosomal RNA, large subunit ribosomal RNA, 5S ribosomal RNA, Group I self-splicing introns, RNase P RNA, signal recognition particle RNA, tRNA, and tmRNA. The structural data was collected from databases such as RNA STRAND v2.0, 5S ribosomal RNA database, Rfam 9.1, and tmRDB. The dataset contains redundant sequences, or the same sequence of RNAs from different species, which have not been removed from the dataset.

bpRNA-1m bpRNA-1m [115] is another RNA secondary structure meta-database that has been compiled from multiple data sources, such as Comparative RNA Web (CRW) [124], tmRNA database [125], tRNAdb [126], Signal Recognition Particle database [127], RNase P database [128], tRNAdb 2009 database [129], and RCSB Protein Data Bank [130], and RFAM 12.2 [131]. bpRNA-1m is a comprehensive database that contains 102 318 RNA sequences and structures and approximately 2500 different RNA families. A less redundant version exists, bpRNA-1m(90), which removes sequences with greater than 90% sequence similarity with at least 70% alignment coverage. As a subset of bpRNA-1m, bpRNA-1m(90) contains fewer sequences or 28 370 sequences and structures. bpRNA-1m could also be used under an alias. MXFold2 uses this dataset and splits the dataset into three different subset datasets, TR0, VL0, and TS0, where they correspond to the training, validation, and testing datasets, respectively.

TORNADO dataset Rivas et al. [64] curated a dataset in the development of their RSS prediction method, TORNADO. This dataset then serves as a benchmark widely used to measure the performance of a number of RSS prediction models later. The dataset consists of 4 different sets, TrainSetA, TestSetA, TrainSetB, TestSetB. TrainSetA and TestSetA were constructed by collecting sequences and structures from trusted literature. The TestSetA specifically ensures that there is no sequence redundancy by removing nearly identical sequences. This process results in a dataset with low sequence similarity, but as it is from 11 RNA families, the structural similarity is high. To combat this, TrainSetB and TestSetB were constructed by including 22 RNA families from the Rfam database.

RNAStrAlign RNAStrAlign [117], the most recent dataset by Mathews Lab, was constructed as a benchmark for the RSS folding algorithm, TurboFold II. The dataset consists of 8 different RNA families, 5S ribosomal RNA, Group I intron, tmRNA, tRNA, 16S ribosomal RNA, Signal Recognition Particle (SRP) RNA, RNase P RNA, and telomerase RNA, totaling 30 451 sequences, taken from disparate online databases.

RNAcentral While RNAcentral [118] is a database that specifically contains data about non-coding RNA (ncRNA) sequences. The most recent version has started integrating and visualizing known RNA secondary structures of tRNA sequences, imported from GtRNAdb [132]. While the database does not provide a curated list of RNA secondary structures, it contains RNA secondary structure data of known ncRNA sequences, which can be used as a starting point for creating an RSS benchmark specialized in ncRNA folding.

Rfam Rfam is a comprehensive RNA database that contains sequences and alignments from a wide variety of RNA families [116]. For each family, a seed multiple-sequence alignment is curated from a small set of representative sequences with a corresponding conserved secondary structure annotation. When such alignment and secondary structure annotation are not available from the literature, the Rfam team will generate them using programs like RNAalifold [79] mentioned above, with manual adjustment. The seed alignment is further used to build a covariance model, and subsequently, a full alignment with more sequences scored above a cutoff by the covariance model is added to the seed alignment. In general, Rfam is widely used in the field as the gold standard for training and assessing the accuracy of structure prediction programs. At the time of writing, Rfam holds 4178 families across different species and for both coding and non-coding RNAs.

NNDB Different from the other data reviewed above, the nearest neighbor database (NNDB) [53,103], as the name indicates, stores the nearest neighbor thermodynamic parameters for RNA and DNA from several experiments. As the backbone of energy-based methods, these parameters are the key to RSS prediction and have been widely used, which is why we also include NNDB here. Currently, the database contains both the 1999 [58] and 2004 [51] versions

of the Turner's nearest neighbor parameters, the m6A modified parameters [18] as introduced earlier (section 4.3), and a set of DNA folding parameters. For the three RNA parameter sets, free energy changes are stored as parameter lookup tables. The Turner 2004 version also has parameter tables for enthalpy changes. In addition to providing the values of the parameters, NNDB also summarizes the rules and provides representative examples of how to use these parameters, which can be easily adapted to dynamic programming and other methods of development.

PDB data Although not secondary structure data, PDB-derived datasets may potentially provide indirect RSS information from the 3D structures. So, we still include them in this review for a brief discussion. Recent research demonstrates the usefulness of them in enhancing tertiary RNA structure prediction. RNA3DB, developed by Szikszai et al. [120], utilizes PDB data to construct a comprehensive dataset optimized for training and evaluating deep learning models in RNA structure prediction. This dataset solves issues related to the restricted availability and diversity of experimentally determined RNA structures, providing a more solid basis for the development of computational tools. Additionally, the MARS and RNAcmap3 databases broaden the scope of RNA folding research by integrating RNA sequences from many sources to improve multiple sequence alignments, an essential process for precise secondary and tertiary structure predictions [133]. In total, these integrating PDB resources may provide potential source for addressing the data scarcity challenges associated with RNA secondary structure prediction.

5.2   RSS related RNA modification datasets

MODOMICS MODOMICS [17] is a comprehensive database that provides information about the chemical structure of RNA modifications, biological pathways, and RNA sequence location of modification, links to human diseases, and the participating RNA modification enzymes. This database collects information about more than 170 different types of RNA modifications that are currently still being discovered with the development of new high-throughput technologies. MODOMICs currently contains a total of 429 different RNA modified residues with 335 natural ones among them, from RNA types such as tRNA and small nucleolar RNA (snoRNA). It also hosts 1925 different RNA sequences from multiple RNA families such as tmRNA, tRNA, rRNA, small nuclear RNA (snRNA), snoRNA, and Piwi-interacting RNA (piRNA). There are many more data sources dedicated for RNA modifications. However, since this review is not for pure RNA modifications, we only include MODOMICS here as an representative of this class.

ENCORI ENCORI [121], also known as starBase v2.0, is a database that hosts detailed information about RNA-RNA and Protein-RNA interaction networks by analyzing crosslinking immunoprecipitation sequencing (CLIP-Seq) studies. Although not RNA modification nor RSS data directly, the data in ENCORI may potentially provide indirect information to link RNA modification and RSS folding, through relevant protein-RNA interactions for example. So, we put it here as an representative. At the time of this writing, ENCORI has analyzed and hosted interactions from 2725 CLIP-Seq datasets, resulting in millions of miRNA, RBP, RNA interactions with ncRNA and mRNA. In addition to the study of miRNAs' impact on mRNAs through CLIP-Seq datasets, due to the advancements of Degradome sequencing, ENCORI also hosts analyses of miRNA-RNA interactions through the degradome-seq datasets. ENCORI also allows its users to study the effects of these RNA-RNA and protein-RNA interactions on gene expression in 32 different cancer types.

eCLIP data in ENCODE eCLIP data [122] utilizes a more efficient and robust method of CLIP-Seq, called enhanced CLIP (eCLIP). We include this dataset here for the same reason as for ENCORI. With more efficient sample requirements and while retaining single-nucleotide resolution, eCLIP reduces the high failure rates of previous CLIP protocols. To find RBP-RNA interactions, eCLIP experiments were carried out on two human cell lines, K562 and HepG2. For the K562 cells, 119 RBPs were studied, while for the HepG2 cells, 102 RBPs were studied. The eCLIP dataset is publicly available on the ENCODE project website.

6   Challenges and opportunities

Predicting RNA secondary structures is challenging due to several key factors. First, there are significantly fewer known RNA structures compared to protein structures. This lack of data makes it hard to train machine learning models effectively, leading to biased results and lower prediction accuracy. Alternative rule-based dynamic programming algorithms are widely used but suffer from several issues, like the scalability for long sequences, etc. Second, pseudoknots and noncanonical interactions also introduce further difficulties, as many traditional DP-based methods

struggle to account for them. Third, limited thermodynamic parameters and challenges in sampling different RNA shapes add to the problem. Finally, RNA folding is a complex process. The secondary structures can significantly affect the tertiary structures, making prediction more complicated. These issues highlight the urgent need for better models and techniques in RNA secondary structure prediction.

Recent advancements in machine learning, especially those inspired by deep learning and natural language processing, greatly improve the prediction of RNA secondary structures. These techniques can use large RNA sequence datasets to boost prediction accuracy. Combining machine learning with traditional thermodynamic and physics-based models can help us better understand how RNA folds. The rise of RNA language models, such as RNAErnie[81], RNA-FM[14], and UNI-RNA[134], along with experimental data like chemical probing results[135], provides valuable resources for refining prediction models, particularly for longer RNA sequences. Collaboration between theoretical and experimental researchers can spark innovation. Moreover, considering environmental factors such as temperature, ligands, and ions can lead to predictions that are more relevant to biological contexts. These opportunities point to a promising future for RNA secondary structure prediction, with important implications for computational methods and biological insights.

Moreover, RNA secondary structure refers to the arrangement of base pairs formed through hydrogen bonding between nucleotides, which plays a crucial role in determining the tertiary structure and functionality of RNA molecules. In the context of mRNA vaccines, modifications such as converting uridine residues to N1-methylpseudouridine (m1Ψ) are strategically employed to enhance RNA stability and reduce immunogenicity, both of which are vital for vaccine efficacy[136]. These modifications can significantly alter the free energy parameters and base-pairing interactions, thereby complicating the prediction of secondary structures. Therefore, developing RNA secondary structure prediction methods that accurately account for these modifications is essential for advancing RNA drug discovery and therapeutic applications[16]. Accurate predictions enable the rational design of RNA molecules with specific properties, ultimately improving the effectiveness of RNA-based therapeutics. Consequently, integrating advanced prediction methodologies is critical for optimizing the therapeutic potential of RNA technologies.

RNA secondary structure prediction can be employed to understand the regulatory mechanisms of stress response and virulence in foodborne pathogens. For instance, RNA thermometers are non-coding RNA elements that regulate gene expression in response to temperature changes[137], playing a crucial role in the virulence of pathogens like *Listeria monocytogenes*[138], *Escherichia coli*[139] and *Salmonella Typhimurium*[140]. Additionally, small RNAs (sRNAs) have been shown to influence the expression of virulence factors in foodborne pathogens, with studies indicating their involvement in stress response and pathogenicity[141]. Furthermore, RNA secondary structure prediction is critical in developing gene therapy. The RSS tools help identify structural elements influencing RNA stability, translation efficiency, and interactions with RBPs. For example, RBPs such as OAS proteins can inhibit gene expression by binding to specific secondary structures within therapeutic RNA molecules, thereby reducing their therapeutic efficacy[142]. By using RNA secondary structure prediction, researchers can modify RNA molecules to prevent such inhibitory interactions with RBP binding sites[143]. Additionally, RNA modifications, such as m6A, Ψ, or m5C, can improve the efficiency of gene therapy by enhancing RNA stability, reducing immune activation, and improving translational output[144].

Lastly, in the context of RNA tertiary structure prediction, recent developments such as AlphaFold3 have garnered significant attention due to their ability to model heterogeneous macromolecular systems, including large RNA molecules. A recent study[145], Structure Prediction of Large RNAs with AlphaFold3 Highlights its Capabilities and Limitations, provides a comprehensive assessment of AlphaFold3's performance on RNA structures of up to 5000 nucleotides. The study highlights both the potential and the limitations of this tool, particularly for predicting large RNA molecules whose experimental dimensions are known. While AlphaFold3 can generate plausible models, challenges persist, including severe steric clashes, occasional breaks in the phosphodiester backbone, and excessive sphericalization of structures, with this effect becoming more pronounced as RNA length increases. Notably, hydrodynamic radii calculated from AlphaFold3 models are substantially larger than experimental measurements under low-salt conditions but align more closely with experimental results in the presence of polyvalent cations. These findings suggest that while AlphaFold3 can be used for RNA structure prediction, especially for RNAs up to 2000 nucleotides, it may be required to identify geometrically accurate predictions free of structural artifacts. These limitations suggest that AlphaFold3 provides a useful starting point for RNA structure modeling; nonetheless, it requires further optimization

and complementary approaches, such as experimental data integration or RNA-specific prediction tools, to achieve reliable RNA structure predictions. The static nature of PDB structure data, which captures only a single RNA conformation, also presents limitations to AlphaFold3.

7 Conclusion

In this review, we explored the advances in RNA secondary structure predictions, RNA modifications, and their interplays, highlighting the progression of the methodology and the connection between RNA structure and modification. RNA secondary structure prediction methods have evolved drastically over the past decades, moving from dynamic programming algorithms to sophisticated learning-based approaches that account for complex structure patterns such as pseudoknots and long-range interactions. Meanwhile, advances in RNA modification prediction tools have leveraged the progress in experimental data and various machine learning and deep learning paradigms to analyze the functional roles of many modification types. The integration of RNA secondary structural information into modification prediction models, and vice versa, has expanded our understanding of the critical role of RNA in regulating gene expression, RNA-protein interactions, and other important biological processes. Nevertheless, significant challenges remain in the field, such as the scarcity of data compared to the related field of protein structures. Addressing them will not only supply more accurate prediction methods but also pave the way for novel therapeutic applications and breakthroughs in quantitative biology.

References

1. Spitale RC, Incarnato D. Probing the dynamic RNA structurome and its functions. Nature Reviews Genetics. 2023;24(3):178-96.
2. Assmann SM, Chou HL, Bevilacqua PC. Rock, scissors, paper: How RNA structure informs function. The Plant Cell. 2023;35(6):1671-707.
3. Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM. The roles of structural dynamics in the cellular functions of RNAs. Nature reviews Molecular cell biology. 2019;20(8):474-89.
4. Edwards TE, Klein DJ, Ferré-D'Amaré AR. Riboswitches: small-molecule recognition by gene regulatory RNAs. Current opinion in structural biology. 2007;17(3):273-9.
5. Fu XD. Non-coding RNA: a new frontier in regulatory biology. National science review. 2014;1(2):190-204.
6. Crooke ST, Witztum JL, Bennett CF, Baker BF. RNA-targeted therapeutics. Cell metabolism. 2018;27(4):714-39.
7. Nowakowski J, Tinoco Jr I. RNA structure and stability. In: Seminars in virology. vol. 8. Elsevier; 1997. p. 153-65.
8. Tinoco Jr I, Bustamante C. How RNA folds. Journal of molecular biology. 1999;293(2):271-81.
9. Jones CP, Ferré-D'Amaré AR. RNA quaternary structure and global symmetry. Trends in biochemical sciences. 2015;40(4):211-20.
10. Mathews DH, Moss WN, Turner DH. Folding and finding RNA secondary structure. Cold Spring Harbor perspectives in biology. 2010;2(12):a003665.
11. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. nature. 2021;596(7873):583-9.
12. Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nature communications. 2019;10(1):5407.
13. 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 research. 2022;50(3):e14-4.
14. Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. arXiv preprint arXiv:220400300. 2022.
15. Gilbert WV, Bell TA, Schaening C. Messenger RNA modifications: form, distribution, and function. Science. 2016;352(6292):1408-12.
16. Tanzer A, Hofacker IL, Lorenz R. RNA modifications in structure prediction–Status quo and future challenges. Methods. 2019;156:32-9.
17. Cappannini A, Ray A, Purta E, Mukherjee S, Boccaletto P, Moafinejad SN, et al. MODOMICS: a database of RNA modifications and related information. 2023 update. Nucleic Acids Research. 2024;52(D1):D239-44.

18. Kierzek E, Zhang X, Watson RM, Kennedy SD, Szabat M, Kierzek R, et al. Secondary structure prediction for RNA sequences including N6-methyladenosine. Nature communications. 2022;13(1):1271.

19. Lewis CJ, Pan T, Kalsotra A. RNA modifications and structures cooperate to guide RNA–protein interactions. Nature reviews Molecular cell biology. 2017;18(3):202-10.

20. Boo SH, Kim YK. The emerging role of RNA modifications in the regulation of mRNA stability. Experimental & molecular medicine. 2020;52(3):400-8.

21. Wang S, Lv W, Li T, Zhang S, Wang H, Li X, et al. Dynamic regulation and functions of mRNA m6A modification. Cancer cell international. 2022;22(1):48.

22. Cui L, Ma R, Cai J, Guo C, Chen Z, Yao L, et al. RNA modifications: importance in immune cell biology and related diseases. Signal transduction and targeted therapy. 2022;7(1):334.

23. Liu WW, Zheng SQ, Li T, Fei YF, Wang C, Zhang S, et al. RNA modifications in cellular metabolism: implications for metabolism-targeted therapy and immunotherapy. Signal Transduction and Targeted Therapy. 2024;9(1):70.

24. Ramos J. RNA modifications: an overview of select web-based tools. RNA. 2022;28(11):1440-5.

25. Song Z, Huang D, Song B, Chen K, Song Y, Liu G, et al. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nature communications. 2021;12(1):4011.

26. Qiao J, Jin J, Yu H, Wei L. Towards retraining-free RNA modification prediction with incremental learning. Information Sciences. 2024;660:120105.

27. Chen T, Wu T, Pan D, Xie J, Zhi J, Wang X, et al. TransRNAm: identifying twelve types of RNA modifications by an interpretable multi-label deep learning model based on Transformer. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2023.

28. Xuan J, Chen L, Chen Z, Pang J, Huang J, Lin J, et al. RMBase v3. 0: decode the landscape, mechanisms and functions of RNA modifications. Nucleic Acids Research. 2024;52(D1):D273-84.

29. Wang X, Zhang Y, Chen K, Liang Z, Ma J, Xia R, et al. m7GHub V2. 0: an updated database for decoding the N7-methylguanosine (m7G) epitranscriptome. Nucleic Acids Research. 2024;52(D1):D203-12.

30. Pham NT, Rakkiyapan R, Park J, Malik A, Manavalan B. H2Opred: a robust and efficient hybrid deep learning model for predicting 2'-O-methylation sites in human RNA. Briefings in Bioinformatics. 2024;25(1):bbad476.

31. Harun-Or-Roshid M, Pham NT, Manavalan B, Kurata H. Meta-2OM: a multi-classifier meta-model for the accurate prediction of RNA 2'-O-methylation sites in human RNA. PloS One. 2024;19(6):e0305406.

32. Geng YQ, Lai FL, Luo H, Gao F. Nmix: a hybrid deep learning model for precise prediction of 2'-O-methylation sites based on multi-feature fusion and ensemble learning. Briefings in Bioinformatics. 2024;25(6):bbae601.

33. Pham NT, Terrance AT, Jeon YJ, Rakkiyappan R, Manavalan B. ac4C-AFL: A high-precision identification of human mRNA N4-acetylcytidine sites based on adaptive feature representation learning. Molecular Therapy-Nucleic Acids. 2024;35(2).

34. Jia Y, Zhang Z, Yan S, Zhang Q, Wei L, Cui F. Voting-ac4C: Pre-trained large RNA language model enhances RNA N4-acetylcytidine site prediction. International Journal of Biological Macromolecules. 2024;282:136940.

35. Su W, Xie XQ, Liu XW, Gao D, Ma CY, Zulfiqar H, et al. iRNA-ac4C: a novel computational method for effectively detecting N4-acetylcytidine sites in human mRNA. International Journal of Biological Macromolecules. 2023;227:1174-81.

36. Liu R, Zhang Y, Wang Q, Zhang X. TransAC4C—a novel interpretable architecture for multi-species identification of N4-acetylcytidine sites in RNA with single-base resolution. Briefings in Bioinformatics. 2024;25(3):bbae200.

37. Hasan MM, Tsukiyama S, Cho JY, Kurata H, Alam MA, Liu X, et al. Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. Molecular Therapy. 2022;30(8):2856-67.

38. Kurata H, Harun-Or-Roshid M, Hasan MM, Tsukiyama S, Maeda K, Manavalan B. MLm5C: A high-precision human RNA 5-methylcytosine sites predictor based on a combination of hybrid machine learning models. Methods. 2024;227:37-47.

39. Abbas Z, ur Rehman M, Tayara H, Zou Q, Chong KT. XGBoost framework with feature selection for the prediction of RNA N5-methylcytosine sites. Molecular Therapy. 2023;31(8):2543-51.

40. Su Q, Phan LT, Pham NT, Wei L, Manavalan B, et al. MST-m6A: A Novel Multi-Scale Transformer-based Framework for Accurate Prediction of m6A Modification Sites Across Diverse Cellular Contexts. Journal of Molecular Biology. 2024:168856.

41. Xia Y, Zhang Y, Liu D, Zhu YH, Wang Z, Song J, et al. BLAM6A-Merge: Leveraging Attention Mechanisms and Feature Fusion Strategies to Improve the Identification of RNA N6-methyladenosine Sites. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2024.

42. Zhao Y, Jin J, Gao W, Qiao J, Wei L. Moss-m7G: A Motif-Based Interpretable Deep Learning Method for RNA N7-Methlguanosine Site Prediction. Journal of Chemical Information and Modeling. 2024;64(15):6230-40.

43. Shoombuatong W, Basith S, Pitti T, Lee G, Manavalan B. THRONE: a new approach for accurate prediction of human RNA N7-methylguanosine sites. Journal of Molecular Biology. 2022;434(11):167549.

44. Bugnon LA, Edera AA, Prochetto S, Gerard M, Raad J, Fenoy E, et al. Secondary structure prediction of long noncoding RNA: review and experimental comparison of existing approaches. Briefings in Bioinformatics. 2022;23(4):bbac205.

45. Sato K, Hamada M. Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. Briefings in Bioinformatics. 2023;24(4):bbad186.

46. Schneider B, Sweeney BA, Bateman A, Cerny J, Zok T, Szachniuk M. When will RNA get its AlphaFold moment? Nucleic Acids Research. 2023;51(18):9522-32.

47. Wu KE, Zou JY, Chang H. Machine learning modeling of RNA structures: methods, challenges and future perspectives. Briefings in Bioinformatics. 2023;24(4):bbad210.

48. Zhang J, Fei Y, Sun L, Zhang QC. Advances and opportunities in RNA structure experimental determination and computational modeling. Nature methods. 2022;19(10):1193-207.

49. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015;31(20):3377-9.

50. Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proceedings of the National Academy of Sciences of the USA. 1980;77(11):6309-13.

51. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences. 2004;101(19):7287-92.

52. Schroeder SJ, Turner DH. Optical melting measurements of nucleic acid thermodynamics. In: Methods in enzymology. vol. 468. Elsevier; 2009. p. 371-87.

53. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic acids research. 2010;38(suppl_1):D280-2.

54. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic acids research. 1981;9(1):133-48.

55. Dieterich1 C, Stadler PF. Computational biology of RNA interactions. Wiley Interdisciplinary Reviews: RNA. 2012;4:107-20.

56. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for molecular biology. 2011;6:1-14.

57. Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of molecular biology. 1999;285(5):2053-68.

58. Mathews D, Sabina J, Zuker M, Turner D. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. J Mol Biol 1999b;288.

59. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers: Original Research on Biomolecules. 1990;29(6-7):1105-19.

60. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC bioinformatics. 2010;11:1-9.

61. Ali SE, Mittal A, Mathews DH. RNA Secondary Structure Analysis Using RNAstructure. Current protocols. 2023 July;3(7):e846. Available from: https://europepmc.org/articles/PMC11267465.

62. Xia T, SantaLucia Jr J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, 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.

63. Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. Nucleic acids research. 2006;34(17):4912-24.

64. Rivas E, Lang R, Eddy SR. A range of complex probabilistic models for RNA secondary structure prediction that include the nearest neighbor model and more. RNA. 2012;18(2):193-212.

65. 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):i295-304.

66. Andronescu MS. Algorithms for predicting the secondary structure of pairs and combinatorial sets of nucleic acid strands. University of British Columbia; 2003.

67. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics. 2007;23(13):i19-28.

68. Sükösd Z, Knudsen B, Kjems J, Pedersen CN. PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data. Bioinformatics. 2012;28(20):2691-2.

69. Sükösd Z, Knudsen B, Værum M, Kjems J, Andersen ES. Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. BMC bioinformatics. 2011;12:1-8.

70. Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic acids research. 2004;32(16):4925-36.

71. Pedersen JS, Forsberg R, Meyer IM, Hein J. An evolutionary model for protein-coding regions with conserved RNA structure. Molecular biology and evolution. 2004;21(10):1913-22.

72. Saman Booy M, Ilin A, Orponen P. RNA secondary structure prediction with convolutional neural networks. BMC bioinformatics. 2022;23(1):58.

73. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006 07;22(14):e90-8. Available from: https://doi.org/10.1093/bioinformatics/btl246.

74. Chen X, Li Y, Umarov R, Gao X, Song L. RNA Secondary Structure Prediction By Learning Unrolled Algorithms. In: International Conference on Learning Representations; 2020. Available from: https://openreview.net/forum?id=S1eALyrYDH.

75. Chen CC, Chan YM. REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network. BMC bioinformatics. 2023;24(1):122.

76. 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(17):2589-600.

77. Akiyama M, Sato K, Sakakibara Y. A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. Bioinformatics. 2017;33(22):3373-9.

78. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. Nature communications. 2021;12(1):941.

79. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC bioinformatics. 2008;9:1-13.

80. Sato K, Hamada M, Asai K, Mituyama T. CENTROIDFOLD: a web server for RNA secondary structure prediction. Nucleic acids research. 2009;37(suppl_2):W277-80.

81. Wang N, Bian J, Li Y, Li X, Mumtaz S, Kong L, et al. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. Nature Machine Intelligence. 2024:1-10.

82. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics (Oxford, England). 1999;15(6):446-54.

83. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic acids research. 2003;31(13):3423-8.

84. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution. 1981;17:368-76.

85. Sakai I. Syntax in universal translation. In: Proceedings of the International Conference on Machine Translation and Applied Language Analysis; 1961. .

86. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer; 2016. p. 630-45.

87. Hochreiter S. Long Short-term Memory. Neural Computation MIT-Press. 1997.

88. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing. 1997;45(11):2673-81.

89. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations; 2016. .

90. Shen T, Hu Z, Peng Z, Chen J, Xiong P, Hong L, et al.. E2Efold-3D: End-to-End Deep Learning Method for accurate de novo RNA 3D Structure Prediction; 2022. Available from: https://arxiv.org/abs/2207.01586.

91. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015. p. 234-41.

92. Hao Z, et al. A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming.[J]. Frontiers in genetics. 2019;10:467.

93. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

94. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2017. p. 11-9.

95. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics. 2009;25(4):465-73.

96. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Computational approaches for RNA energy parameter estimation. RNA. 2010;16(12):2304-18.

97. Zhang Y, Wang Z, Zhang Y, Li S, Guo Y, Song J, et al. Interpretable prediction models for widespread m6A RNA modification across cell lines and tissues. Bioinformatics. 2023;39(12):btad709.

98. Höfler S, Duss O. Interconnections between m6A RNA modification, RNA structure, and protein–RNA complex assembly. Life Science Alliance. 2024;7(1).

99. El Allali A, Elhamraoui Z, Daoud R. Machine learning applications in RNA modification sites prediction. Computational

and Structural Biotechnology Journal. 2021;19:5510-24.

100. Xiang S, Liu K, Yan Z, Zhang Y, Sun Z. RNAMethPre: a web server for the prediction and query of mRNA m6A sites. PloS one. 2016;11(10):e0162707.

101. Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. N 6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. Nature. 2015;518(7540):560-4.

102. Brümmer A, Yang Y, Chan TW, Xiao X. Structure-mediated modulation of mRNA abundance by A-to-I editing. Nature communications. 2017;8(1):1255.

103. Mittal A, Turner DH, Mathews DH. NNDB: An Expanded Database of Nearest Neighbor Parameters for Predicting Stability of Nucleic Acid Secondary Structures. Journal of Molecular Biology. 2024:168549.

104. Szabat M, Prochota M, Kierzek R, Kierzek E, Mathews DH. A test and refinement of folding free energy nearest neighbor parameters for RNA including N6-methyladenosine. Journal of Molecular Biology. 2022;434(18):167632.

105. Varenyk Y, Spicher T, Hofacker IL, Lorenz R. Modified RNAs and predictions with the ViennaRNA Package. Bioinformatics. 2023;39(11):btad696.

106. Richardson KE, Znosko BM. Nearest-neighbor parameters for 7-deaza-adenosine· uridine base pairs in RNA duplexes. RNA. 2016;22(6):934-42.

107. Wright DJ, Rice JL, Yanker DM, Znosko BM. Nearest Neighbor Parameters for Inosine·Uridine Pairs in RNA Duplexes. Biochemistry. 2007;46(15):4625-34.

108. Wright DJ, Force CR, Znosko BM. Stability of RNA duplexes containing inosine· cytosine pairs. Nucleic Acids Research. 2018;46(22):12099-108.

109. Hudson GA, Bloomingdale RJ, Znosko BM. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. RNA. 2013;19(11):1474.

110. Jolley EA, Znosko BM. The loss of a hydrogen bond: Thermodynamic contributions of a non-standard nucleotide. Nucleic acids research. 2017;45(3):1479-87.

111. Dalluge JJ, Hashizume T, Sopchik AE, McCloskey JA, Davis DR. Conformational flexibility in RNA: the role of dihydrouridine. Nucleic acids research. 1996;24(6):1073-9.

112. Chou FC, Kladwang W, Kappel K, Das R. Blind tests of RNA nearest-neighbor energy prediction. Proceedings of the National Academy of Sciences. 2016;113(30):8430-5.

113. Lorenz R, Hofacker IL, Stadler PF. RNA folding with hard and soft constraints. Algorithms for Molecular Biology. 2016;11:1-13.

114. Sloma MF, Mathews DH. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. RNA. 2016;22(12):1808-18.

115. 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 research. 2018;46(11):5381-94.

116. Ontiveros-Palacios N, Cooke E, Nawrocki EP, Triebel S, Marz M, Rivas E, et al. Rfam 15: RNA families database in 2025. Nucleic Acids Research. 2024:gkae1023.

117. Tan Z, Fu Y, Sharma G, Mathews DH. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. Nucleic acids research. 2017;45(20):11570-81.

118. Consortium TR. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic acids research. 2021;49(D1):D212-20.

119. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000;28(1):235-42.

120. Szikszai M, Magnus M, Sanghi S, Kadyan S, Bouatta N, Rivas E. RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction. Journal of Molecular Biology. 2024. Advance online publication.

121. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic acids research. 2014;42(D1):D92-7.

122. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nature methods. 2016;13(6):508-14.

123. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of molecular biology. 1999;288(5):911-40.

124. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC bioinformatics. 2002;3:1-31.

125. Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, et al. The tmRDB and SRPDB resources. Nucleic acids research. 2006;34(suppl_1):D163-8.

126. Zwieb C, Gorodkin J, Knudsen B, Burks J, Wower J. tmRDB (tmRNA database). Nucleic acids research. 2003;31(1):446-7.

127. Rosenblad MA, Gorodkin J, Knudsen B, Zwieb C, Samuelsson T. SRPDB: signal recognition particle database. Nucleic acids research. 2003;31(1):363-4.

128. Brown JW. The ribonuclease P database. Nucleic acids research. 1998;26(1):351-2.

129. Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic acids research. 2009;37(suppl_1):D159-62.

130. Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic acids research. 2016:gkw1000.

131. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic acids research. 2015;43(D1):D130-7.

132. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic acids research. 2016;44(D1):D184-9.

133. Chen K, Litfin T, Singh J, Zhan J, Zhou Y. MARS and RNAcmap3: The Master Database of All Possible RNA Sequences Integrated with RNAcmap for RNA Homology Search. Genomics, Proteomics & Bioinformatics. 2024;22(1):qzae018.

134. Wang X, Gu R, Chen Z, Li Y, Ji X, Ke G, et al. UNI-RNA: universal pre-trained models revolutionize RNA research. bioRxiv. 2023:2023-07.

135. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proceedings of the National Academy of Sciences. 2009;106(1):97-102.

136. Karikó K, Buckstein M, Ni H, Weissman D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. Immunity. 2005;23(2):165-75.

137. Johansson J, Mandin P, Renzoni A, Chiaruttini C, Springer M, Cossart P. An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. Cell. 2002;110(5):551-61.

138. Hanes R, Zhang F, Huang Z. Protein interaction network analysis to investigate stress response, virulence, and antibiotic resistance mechanisms in Listeria monocytogenes. Microorganisms. 2023;11(4):930.

139. Zhang F, Graham J, Zhai T, Liu Y, Huang Z. Discovery of MurA inhibitors as novel antimicrobials through an integrated computational and experimental approach. Antibiotics. 2022;11(4):528.

140. Liu Y, Zhang F, Hawkins JL, Elder JR, Baranzoni GM, Huang Z, et al. Comparative Gene Expression Analysis of Salmonella Typhimurium DT104 in Ground Chicken Extract and Brain Heart Infusion Broth. Microorganisms. 2024;12(7):1461.

141. Padalon-Brauch G, Hershberg R, Elgrably-Weiss M, Baruch K, Rosenshine I, Margalit H, et al. Small RNAs encoded within genetic islands of Salmonella typhimurium show host-induced expression and role in virulence. Nucleic acids research. 2008;36(6):1913-27.

142. Zhai T, Song Z, Warga E, Elmer J, Huang Z. Investigation of Small Molecule Inhibitors for OAS 1 to Enhance Gene Therapy. In: 2023 AIChE Annual Meeting. AIChE; 2023. .

143. Roundtree IA, Evans ME, Pan T, He C. Dynamic RNA modifications in gene expression regulation. Cell. 2017;169(7):1187-200.

144. Zaccara S, Ries RJ, Jaffrey SR. Reading, writing and erasing mRNA methylation. Nature reviews Molecular cell biology. 2019;20(10):608-24.

145. McDonnell RT, Henderson AN, Elcock AH. Structure Prediction of Large RNAs with AlphaFold3 Highlights its Capabilities and Limitations. Journal of Molecular Biology. 2024;436(22):168816.