. SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI

SISSA Digital Library

Machine Learning for RNA Secondary Structure Prediction: a review of current methods and challenges

*Original*
Machine Learning for RNA Secondary Structure Prediction: a review of current methods and challenges / Sacco, G., Bussi, G., Sanguinetti, G.. - In: RNA. - ISSN 1355-8382. - 32:4(2026), pp. 443-456. [10.1261/rna.080840.125]

*Availability:*
This version is available at: 20.500.11767/149890 since: 2026-01-27T07:46:03Z

*Publisher:*

*Published*
DOI:10.1261/rna.080840.125

*Terms of use:*

Testo definito dall'ateneo relativo alle clausole di concessione d'uso

*Publisher copyright*
Cold Spring Harbor

This version is available for education and non-commercial purposes.

note finali coverpage

(Article begins on next page)

13 October 2026

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

# Machine Learning for RNA Secondary Structure Prediction: a review of current methods and challenges

Giuseppe Sacco, Giovanni Bussi, Guido Sanguinetti[1]
All authors: Scuola Internazionale Superiore di Studi Avanzati, SISSA, Trieste, Italy, 34136

2026-01-22

**Running head title:** ML for RNA Secondary Structure Prediction

**Keywords:** RNA secondary structure prediction, machine learning, foundation models, deep learning

## Abstract

Predicting the secondary structure of RNA is a core challenge in computational biology, essential for understanding molecular function and designing novel therapeutics. The field has evolved from foundational but accuracy-limited thermodynamic approaches to a new data-driven paradigm dominated by machine learning and deep learning. These models learn folding patterns directly from data, leading to significant performance gains. This review surveys the modern landscape of these methods, covering single-sequence, evolutionary-based, and hybrid models that blend machine learning with biophysics. A central theme is the field's "generalization crisis," where powerful models were found to fail on new RNA families, prompting a community-wide shift to stricter, homology-aware benchmarking. In response to the underlying challenge of data scarcity, RNA foundation models have emerged, learning from massive, unlabeled sequence corpora to improve generalization. Finally, we look ahead to the next set of major hurdles—including the accurate prediction of complex motifs like pseudoknots, scaling to kilobase-length transcripts, incorporating the chemical diversity of modified nucleotides, and shifting the prediction target from static structures to the dynamic ensembles that better capture biological function. We also highlight the need for a standardized, prospective benchmarking system to ensure unbiased validation and accelerate progress.

[1] Corresponding author: gsanguin@sissa.it

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

# Introduction

## The Expanding World of RNA

Ribonucleic acid (RNA) molecules are fundamental biomolecules that fulfill a wide range of biological functions, extending far beyond their traditional role as genetic information carriers from DNA to proteins (Doudna and Cech 2002; Morris and Mattick 2014). In recent years, the diverse functionalities of non-coding RNAs (ncRNAs)—RNA molecules that are not translated into proteins—have been increasingly recognized, impacting processes such as development, cell differentiation, and disease (Statello et al. 2021). The sheer variety of RNA species continues to expand, with many cataloged in comprehensive databases like Rfam and RNAcentral (Kalvari et al. 2021; The RNAcentral Consortium 2019). Long non-coding RNAs (lncRNAs), in particular, are gaining attention for their critical roles in various intracellular regulatory processes in eukaryotes, including humans (Mattick 2001; Kopp and Mendell 2018). Their association with diseases like cancer and neurodegenerative disorders has positioned them as promising new targets for drug discovery (Peng et al. 2017; Wu et al. 2013).

## The Centrality of Secondary Structure

The functional capabilities of RNA molecules are inextricably linked to their intricate structures (Brion and Westhof 1997). RNA structure is typically described through a hierarchical model, beginning with the primary structure, which is the one-dimensional sequence of nucleotides (Adenine, Cytosine, Guanine, and Uracil). The secondary structure then emerges from the primary sequence through the formation of hydrogen bonds between complementary bases. These interactions primarily involve Watson-Crick base pairs (A-U and G-C) and wobble base pairs (G-U), which are the most common in RNA secondary structures (Tinoco and Bustamante 1999). This secondary structure forms rapidly from the primary sequence, accompanied by a significant loss of energy, and critically serves as a foundational scaffold that guides the subsequent folding of the RNA molecule into its complex three-dimensional (tertiary) structure (Brion and Westhof 1997). Consequently, a thorough understanding of RNA secondary structure is paramount for deciphering RNA functions, developing RNA-based therapeutics, and accurately predicting the molecule's final three-dimensional conformation. RNA secondary structures are known to be evolutionarily conserved among RNA species (Eddy and Durbin 1994). RNA secondary structures can be formally represented in several ways. One common method is using binary-valued triangular matrices, where a value of $y_{ij} = 1$ indicates that bases at positions $i$ and $j$ are paired. Another widely used representation is the dot-bracket notation. In its simplest form, this notation uses matching parentheses '(' and ')' to denote paired bases and dots '.' for unpaired bases. This convention is sufficient for representing secondary structures with entirely nested base pairs, which are known as pseudoknot-free structures. However, many biologically important RNAs contain pseudoknots, which are structural motifs characterized by non-nested base pairs where bases within a loop form pairs with bases outside of that loop (Bellaousov and Mathews 2010). Because these

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

crossing interactions cannot be described with a single type of parenthesis, the dot-bracket notation is extended to include additional bracket types (e.g., '[' and ']', or '{' and '}') to represent these more complex topologies.

## Experimental Limits and the Sequence-Structure Gap

Despite the critical importance of RNA structure, its experimental determination, particularly for secondary and tertiary structures, remains a slow, costly, and technically demanding endeavor (Tinoco and Bustamante 1999; Holbrook 2008; Cruz and Westhof 2009; Strobel et al. 2018). High-resolution methods like X-ray crystallography, nuclear magnetic resonance (NMR), and cryogenic electron microscopy (cryo-EM), while powerful, inherently suffer from low throughput. This means that only a minute fraction of the vast number of known RNAs have had their structures experimentally elucidated (Tinoco and Bustamante 1999; Holbrook 2008; Cruz and Westhof 2009). Furthermore, even these gold-standard techniques often provide data that is ensemble- or time-averaged. Since RNA molecules can exist as a heterogeneous ensemble of conformations, this averaging can obscure the presence of less populated but functionally important alternative structures (Cruz and Westhof 2009). Consequently, computational methods are often required to deconvolute this experimental information and model the full structural ensemble. Chemical probing techniques like Selective 2'-Hydroxyl Acylation Analyzed by Primer Extension (SHAPE) (Merino et al. 2005) and Dimethyl Sulphate (DMS) (Peattie and Gilbert 1980) are routinely used to improve the accuracy of predictions based on thermodynamic models (Weeks 2010; Mustoe et al. 2018; Strobel et al. 2018). However, the information content of chemical probing experiments is limited (Sükösd et al. 2013), making the determination of the structure of large RNA molecules still very difficult. This, together with the decreasing cost of sequencing technologies, has led to a significant "sequence-structure gap": an enormous volume of RNA sequence data is continuously generated, yet the number of experimentally determined RNA structures remains severely limited. For instance, less than 0.01% of the millions of non-coding RNAs listed in RNAcentral have experimentally validated structures (Strobel et al. 2018). This substantial gap underscores the urgent and persistent need for accurate, cost-effective, and high-throughput computational prediction methods.

## A Brief History of Computational Methods

The computational prediction of RNA secondary structure has been a prominent area of research since the 1970s (Zuker and Stiegler 1981). Historically, the field was dominated by **thermodynamics-based methods**. These approaches describes the energy of a folded RNA using a nearest-neighbor energy model (Tinoco et al. 1973) and then identify either the minimum free-energy (MFE) structure or the entire partition function using dynamic programming (Nussinov et al. 1978; Zuker 2003; Mathews et al. 1999; Hofacker et al. 1994). However, the performance of these methods eventually plateaued due to fundamental limitations of the nearest-neighbor model and further simplifying assumptions that inherently precluded the prediction of complex but biologically important features such as pseudoknots and tertiary contacts. Moreover, these approaches depend on a fixed catalog

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

of energy parameters obtained through labor-intensive experiments. While the thermodynamic paradigm was central, other classical approaches also made important contributions. **Co-evolutionary methods** leveraged multiple sequence alignments (MSAs) to identify conserved base pairs through correlated mutations (Eddy and Durbin 1994; Nawrocki and Eddy 2013). While powerful when applicable, this approach is fundamentally constrained by the "homology bottleneck": it requires a deep and diverse MSA to distinguish signal from noise, but constructing a meaningful MSA often required prior structural information. Furthermore, this approach is completely inapplicable to the vast number of "orphan" RNAs for which no homologs are known. **Stochastic Context-Free Grammars (SCFGs)** provided a formal probabilistic framework for modeling RNA structure (Sakakibara et al. 1994; Durbin et al. 1998), but, like their thermodynamic counterparts, standard implementations were typically restricted to pseudoknot-free structures and struggled to capture the full complexity of RNA folding without becoming computationally intractable. The limitations inherent in these classical paradigms created a clear need for new approaches. The first wave of **Machine Learning (ML)** emerged as a direct response to the shortcomings of the thermodynamic model. These methods sought to replace the fixed experimental energy parameters with richer, data-driven scoring functions, while still relying on the classical dynamic programming machinery (Do et al. 2006; Andronescu et al. 2010). More recently, the field has been revolutionized by rapid advancements in **Deep Learning (DL)** technologies and the increasing availability of large-scale RNA datasets. These methods represent a paradigm shift, moving from explicit physical or evolutionary models to learning the complex sequence-to-structure mapping directly from data. While many prominent DL approaches are designed to be end-to-end, another successful branch of research uses deep learning to create hybrid models that enhance classical frameworks (Singh et al. 2021; Fu et al. 2022). This data-driven revolution has led to notable improvements in prediction accuracy (Huang et al. 2019; Z. Wang et al. 2025). However, it has also introduced new challenges; these data-hungry models are often susceptible to overfitting and can struggle to generalize to novel RNA families (Szikszai et al. 2022), a critical limitation that remains a central focus of current research.

## Review Scope

This review surveys computational methods for RNA secondary structure prediction across three pillars: classical baselines, data and generalization, and modern deep learning. We first recap the thermodynamics-, evolutionary-, and grammar-based foundations to establish assumptions, strengths, and long-standing limitations that motivate data-driven approaches. We then examine datasets, curation pitfalls, and homology-aware evaluation, highlighting the field’s generalization crisis and the emerging norms for rigorous benchmarking. The core of the review classifies deep learning methods by input regime—single-sequence (ab initio), evolutionary (MSA-based), and biophysical hybrids—and synthesizes trends such as thermodynamic integration and end-to-end predictors. We discuss the emerging frontier of RNA foundation models and their potential to mitigate data scarcity. Finally, we map persistent challenges and opportunities: accurate modeling

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

of pseudoknots and non-canonical pairs, scaling to kilobase-length RNAs, representing chemical modifications and environmental context, and shifting targets from single MFE structures to dynamic ensembles that better capture the dynamic nature of RNA. We conclude with a forward-looking discussion on standards, including prospective community benchmarks to enable fair comparisons and sustained progress.

## Classical Methods

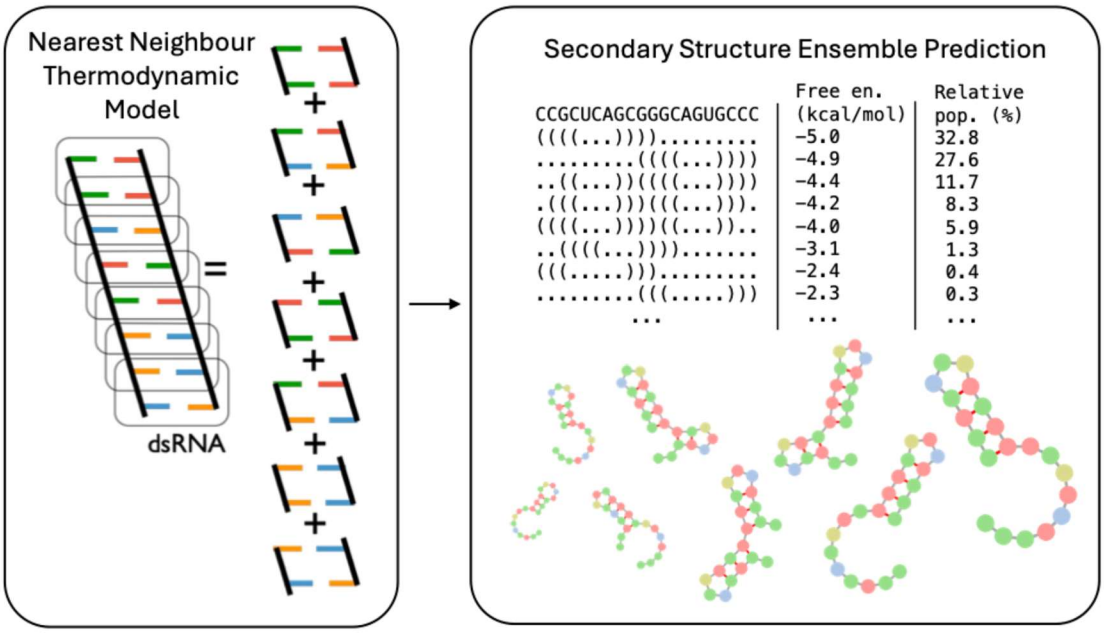


*Figure 1: Schematic representation of thermodynamics-based RNA secondary structure prediction. The free energy of a structure is computed with the Nearest Neighbor model (left panel) as the sum of contributions from individual structural elements, enabling efficient dynamic programming algorithms to enumerate and predict the relative population of all of the possible secondary structures for a given RNA sequence (right panel). Secondary structure visualization generated with Forna (Kerpedjiev et al. 2015).*

The computational prediction of RNA secondary structure has a long history, with methods that can be broadly categorized into four classical paradigms: thermodynamics-based, co-evolutionary, grammar-based, and early machine learning approaches.

### Thermodynamics-Based Methods

The earliest and most dominant approach to RNA secondary structure prediction is based on the principles of thermodynamics, as schematically illustrated in Figure 1. This model posits that the most stable RNA secondary structure is the one possessing the minimum free energy (MFE) (Tinoco et al. 1973; Brion and Westhof 1997; Holbrook 2005). Algorithms based on dynamic programming (DP), such as the Zuker-Stiegler algorithm (Zuker and Stiegler 1981), are designed to efficiently search for these optimal structures (Nussinov et al. 1978; Hofacker et al. 1994). This foundational approach is implemented in widely used

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

software packages such as **Mfold** (Zuker 2003) and its successor **UNAFold** (Markham and Zuker 2008), which compute the MFE structure and can also generate suboptimal foldings. The **ViennaRNA Package** (Lorenz et al. 2011), containing the key component **RNAfold**, implements a similar DP approach and is a benchmark tool in the field. These tools iteratively build optimal structures for subsequences in polynomial time, typically achieving a computational complexity of $O(L^3)$ for an RNA sequence of length $L$, which can be slow for long sequences. A significant limitation of these approaches is their heavy reliance on a fixed set of experimentally determined energy parameters (e.g., Turner's rules) (Turner and Mathews 2010). ViennaRNA and other softwares like **RNAstructure** (Reuter and Mathews 2010) are designed to seamlessly incorporate constraints from chemical probing experiments, which can significantly improve accuracy. A critical shortcoming of most traditional DP-based algorithms is their inherent restriction to predicting "nested" structures, meaning they cannot model pseudoknots. Pseudoknots are non-nested base-pair interactions that are biologically significant, occurring in roughly 40% of all RNAs (Staple and Butcher 2005; Holbrook 2008). RNAstructure, for instance, includes **ProbKnot** (Bellaousov and Mathews 2010), a Maximum Expected Accuracy (MEA) method that predicts the presence of pseudoknots, and **ShapeKnots** (Hajdin et al. 2013), which uses SHAPE data to guide pseudoknot prediction. The general problem of predicting the lowest free-energy structures that include arbitrary pseudoknots has been proven to be NP-complete or NP-hard, making it computationally prohibitive for exact energy-based methods. In addition, it is difficult to know by the sole secondary structure if a given pseudoknot would be achievable in a three-dimensional model, with heuristics that have been proposed to address this issue (Bon and Orland 2011).

## Stochastic Context-Free Grammars (SCFGs)

This foundational concept was established by Eddy & Durbin (Eddy and Durbin 1994), who introduced covariance models (CMs) as a way to describe both the sequence and structure consensus of an RNA family. The software **Infernal** (Nawrocki and Eddy 2013) is the engine that builds and uses these CMs, most famously to maintain the Rfam database (Kalvari et al. 2021). Prediction tools like **Pfold** (Knudsen and Hein 2003) and **EvoFold** (Pedersen et al. 2006) utilize advanced SCFGs that incorporate explicit evolutionary models to predict a consensus structure from an alignment.

## Co-evolutionary Methods

This paradigm leverages the principle that RNA secondary structures are often conserved across evolution, even when the primary sequence diverges. The "alignment folding" strategy is operationalized by tools like **RNAalifold** (Bernhart et al. 2008), part of the ViennaRNA suite. It computes a consensus structure by combining an averaged thermodynamic energy term with an additional score for covariation, the canonical signal of which is a compensatory mutation (e.g., a G-C pair mutating to an A-U pair). The align-then-fold approach relies heavily on the quality of the Multiple Sequence Alignment (MSA) used as input (Gardner and Giegerich 2004). This reliance is addressed by methods that utilize the Covariance Models (CMs), originally defined in the Stochastic Context-Free

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Grammar (SCFG) paradigm (Eddy and Durbin 1994; Rivas et al. 2012; Sato and Hamada 2023). The software Infernal (Nawrocki and Eddy 2013) is central to this, using CMs built from initial alignments or consensus structures (Nawrocki and Eddy 2013; Chen et al. 2024) to perform highly sensitive homology searches against sequence databases (Nawrocki and Eddy 2013; Zhang et al. 2021) and generate refined, structure-aware MSAs via its cmalign program (Chen et al. 2024; Nawrocki and Eddy 2013). Because CMs explicitly model both sequence and secondary structure conservation (Eddy and Durbin 1994; Nawrocki and Eddy 2013; Zhang et al. 2024), the resulting alignments provide superior input quality for subsequent structural inference methods, including those based on Direct Coupling Analysis (DCA) (Cuturello et al. 2020; Zhang et al. 2024; Pucci et al. 2020; De Leonardis et al. 2015). The "simultaneous" strategy is based on the Sankoff algorithm (Sankoff 2006), a DP method that simultaneously aligns sequences and infers a consensus structure. However, its full implementation is computationally intractable. Practical tools are therefore restricted versions; **Dynalign** (Mathews and Turner 2002) implements this for two sequences using a full thermodynamic model and is effective for divergent sequences, while **Foldalign** (Sundfeld et al. 2016) uses a simpler scoring scheme, often for finding short, conserved local motifs.

## Early Machine Learning Approaches

To overcome the inherent limitations of purely thermodynamics-based methods, early machine learning (ML) approaches were introduced to RNA secondary structure prediction. This data-driven approach allowed for the development of more accurate models by enabling a richer and more comprehensive parameterization than what was feasible through wet-lab experiments alone (Sato and Hamada 2023). This era saw a variety of ML techniques applied to learn better scoring functions. **SimFold** (Andronescu et al. 2007), for example, used a regularized linear model to optimize Turner's energy parameters to better fit training data. **TORNADO** (Rivas et al. 2012), a flexible framework for exploring complex Stochastic Context-Free Grammars, used Maximum Likelihood training to parameterize its probabilistic models. However, the most influential methods were based on discriminative training. A landmark example is **CONTRAfold** (Do et al. 2006), which utilized conditional log-linear models (CLLMs). Instead of learning a generative model of the joint probability of a sequence and structure like an SCFG, CONTRAfold directly models the conditional probability of a structure given a sequence. Its approach releases the strict constraints of a formal grammar, allowing for a more flexible, feature-rich scoring system. The model learns weights for these features from data, and the resulting scores are then optimized using a DP algorithm analogous to the classic Zuker algorithm. A critical evolutionary step was taken by **ContextFold** (Zakov et al. 2011), which provided a powerful proof-of-concept for "rich parameterization." It demonstrated that the field's performance plateau could be broken by abandoning the constraint of a small, physically-derived parameter set. By using a discriminative online learning algorithm, ContextFold effectively trained a model with approximately 70,000 parameters describing fine-grained structural and sequential contexts, leading to a nearly 50% reduction in prediction error over the then-state-of-the-art. Here, sequential context denotes the identities and positions of nucleotides at fixed

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

offsets around a structural element (for example, bases flanking a hairpin loop or closing pair), so identical motifs can be scored differently depending on their local surroundings. This hand-crafted, fine-grained context anticipated the learned features of modern deep models, which automatically capture similar local patterns and long-range dependencies via convolutions and attention. This work fundamentally shifted the focus from meticulously measuring energy parameters to designing expressive, data-hungry statistical models, serving as a direct intellectual precursor to later deep learning methods. While these early ML methods resulted in higher prediction accuracy, particularly on datasets structurally similar to their training data, they also introduced a significant risk of overfitting (Sato and Hamada 2023; Szikszai et al. 2022). This overfitting often manifested as a substantial drop in accuracy when these models were applied to predict secondary structures for RNA families that were not represented in their training data, thereby limiting their practical utility for newly discovered RNAs (Sato and Hamada 2023; Szikszai et al. 2022) We will discuss the issue of overfitting and homology-aware benchmarks in more detail in the next section after a brief discussion on the history of the data used in the field.

# Datasets and generalization

## The Evolution of Data

The history of data in RNA secondary structure prediction mirrors the field's methodological evolution, from physics-based models to data-driven paradigm. This progression can be understood as a series of distinct eras, each defined by the nature and scale of the data that enabled its core computational approaches.

The first paradigm to establish itself was the **thermodynamic** one, which was defined not by datasets of examples, but by "data as parameters" (Andronescu et al. 2010). The dominant minimum free energy (MFE) models, pioneered by Zuker and others, relied on a set of thermodynamic parameters that quantified the energetic cost or benefit of forming specific structural motifs like stacks and loops (Zuker and Stiegler 1981; Mathews et al. 1999). These parameters were not learned but meticulously measured through low-throughput optical melting experiments on short, synthetic RNAs (Mathews et al. 2004) This crucial information was compiled and disseminated through resources like the Nearest Neighbor Database (NNDB) (Turner and Mathews 2010), with the Turner rules becoming the *de facto* standard (Mathews et al. 1999, 2004). However, the accuracy of thermodynamic-based prediction was fundamentally bottlenecked by the precision of these physical measurements, and the model's simplifying assumptions highlighted the need for data derived from actual biological molecules (Andronescu et al. 2010; Szikszai et al. 2022).

This led to the development of **comparative and statistical paradigm**, which shifted the focus to curating "gold standard" biological structures. These ground truths were sourced from high-resolution experiments like X-ray crystallography and, more scalably, from comparative sequence analysis, which identifies conserved pairings through co-varying

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

mutations (Eddy and Durbin 1994). Foundational databases like the Comparative RNA Web (CRW) Site (Cannone et al. 2002) and the comprehensive RNA STRAND meta-database (Andronescu et al. 2008) aggregated thousands of these trusted structures, providing the first large-scale corpora for objective benchmarking. This wealth of structural data enabled a powerful feedback loop, allowing the original thermodynamic parameters to be retrained and refined using biological examples (Andronescu et al. 2010). Concurrently, it fueled the development of the first statistical predictors, such as CONTRAfold, which used conditional log-linear models trained on sequence-structure pairs derived from the Rfam database to learn scoring functions directly from data, demonstrating that a purely statistical approach could rival the accuracy of physics-based models (Do et al. 2006; Kalvari et al. 2021).

Concurrent with the curation of structural databases, a new data modality emerged from **chemical probing** experiments, which provide nucleotide-resolution information about the local structural environment of each base *in vitro* and *in vivo* (Strobel et al. 2018; Weeks 2010; Rouskin et al. 2014). Rather than defining a complete structure, this data acts as a set of soft experimental constraints to guide computational predictions (Hajdin et al. 2013; Reuter and Mathews 2010). Influential methods include SHAPE (Selective 2'-hydroxyl acylation analyzed by primer extension) (Merino et al. 2005), which uses reagents to modify the ribose backbone at conformationally flexible nucleotides, and DMS (Dimethyl Sulfate) (Peattie and Gilbert 1980), which modifies the Watson-Crick face of unpaired adenine and cytosine bases. The power of this approach lies in its direct integration with thermodynamic folding algorithms; the experimental reactivity scores are typically converted into soft energy constraints, thereby guiding the MFE search toward an experimentally supported conformation (Deigan et al. 2009; Reuter and Mathews 2010; Hajdin et al. 2013). The coupling of these techniques with next-generation sequencing created high-throughput methods like SHAPE-Seq and DMS-Seq, enabling transcriptome-wide structural interrogation and generating large-scale datasets of experimental constraints (Rouskin et al. 2014; Ding et al. 2014; Mustoe et al. 2018).

The advent of the **Deep Learning paradigm** created an unprecedented demand for data at a massive scale. The millions of parameters in deep neural networks required far larger and more diverse datasets for effective training. This need was met by the creation of key benchmarks like RNAStralign (Tan et al. 2017) and ArchiveII (Saman Booy et al. 2022), and most significantly, the bpRNA-1m (Danaee et al. 2018) database. Aggregating over 100,000 structures, bpRNA-1m provided the necessary scale to train data-hungry models, and its standardized, non-redundant splits (TR0 for training, TS0 for testing) became the community standard for fair evaluation (Singh et al. 2019). However, this new power exposed a critical weakness: many models that performed well on TS0 failed to generalize to novel RNA families not seen during training, a problem that became known as the "generalization crisis" (Szikszai et al. 2022). This realization mandated the development of more rigorous, homology-aware benchmarks. Datasets like bpRNA-new (Sato et al. 2021), composed of entirely new RNA families, and the stringent PDB-derived test sets (e.g., TS2), were created specifically to assess a model's ability to generalize beyond its training distribution (Sato et al. 2021; Singh et al. 2021; Franke et al. 2024).

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Most recently, the **Foundation Model paradigm** entered the field, which leverages the vast, unlabeled sequence space of entire transcriptomes for self-supervised pre-training. The central data source for this paradigm is RNAcentral (The RNAcentral Consortium 2019), a meta-database containing tens of millions of non-coding RNA sequences. By pre-training on this massive corpus, models like RNA-FM aim to learn the fundamental "language" of RNA without direct structural supervision (Chen et al. 2022). The frontier of scale is being pushed even further by efforts like the MARS database (Chen et al. 2024) and the Uni-RNA project (X. Wang et al. 2023), which aim to aggregate over a billion nucleotide sequences from genomic and metagenomic sources. This approach has created a two-tiered data ecosystem: massive, unlabeled sequence corpora are used for computationally intensive pre-training to build generalist models, while the smaller, high-quality labeled datasets like bpRNA are repurposed for the crucial tasks of fine-tuning and rigorous evaluation.

## The Generalization Crisis and the Mandate for Homology-Aware Benchmarking

The challenge of generalizing predictions to novel RNA families has been a long-standing issue in machine learning for RNA structure prediction. Early work on richly parameterized models like ContextFold had already demonstrated that while performance was high within known families, accuracy dropped considerably when tested on unseen ones, suggesting that models were learning family-specific features (Zakov et al. 2011; Rivas et al. 2012).

Despite these early insights, the initial wave of modern deep learning models, often reverted to less rigorous benchmarking based on simple sequence similarity cutoffs. This led to reports of impressive but misleading accuracy (Szikszai et al. 2022; Justyna et al. 2023). The subsequent "generalization crisis" was the widespread realization that these powerful new models were highly susceptible to overfitting and performed poorly when subjected to proper cross-family evaluation. This was starkly investigated and quantified by Szikszai et al. (2022), who demonstrated that a simple deep learning model's accuracy could plummet by 36% in F1-score when moving from a flawed intra-family test to a rigorous inter-family one.

The response to this crisis was a community-wide re-adoption and formalization of rigorous, "homology-aware" benchmarking. Building on the principles laid out by earlier studies, Szikszai et al. (2022) re-emphasized and standardized **family-fold cross-validation**—where entire RNA families are held out for testing—as the necessary gold standard for the deep learning era. As discussed, several strategies have been investigated in order to help address this challenge beyond improved evaluation. These include developing hybrid models like MXfold2 that ground deep learning in biophysical principles (Sato et al. 2021), leveraging evolutionary information as in SPOT-RNA2 (Singh et al. 2021), and designing models like RNAformer with meticulous homology-aware data pipelines from the ground up (Franke et al. 2024). These approaches represent a concerted effort to build more generalizable and biologically faithful predictive tools.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

## Deep Learning Methods

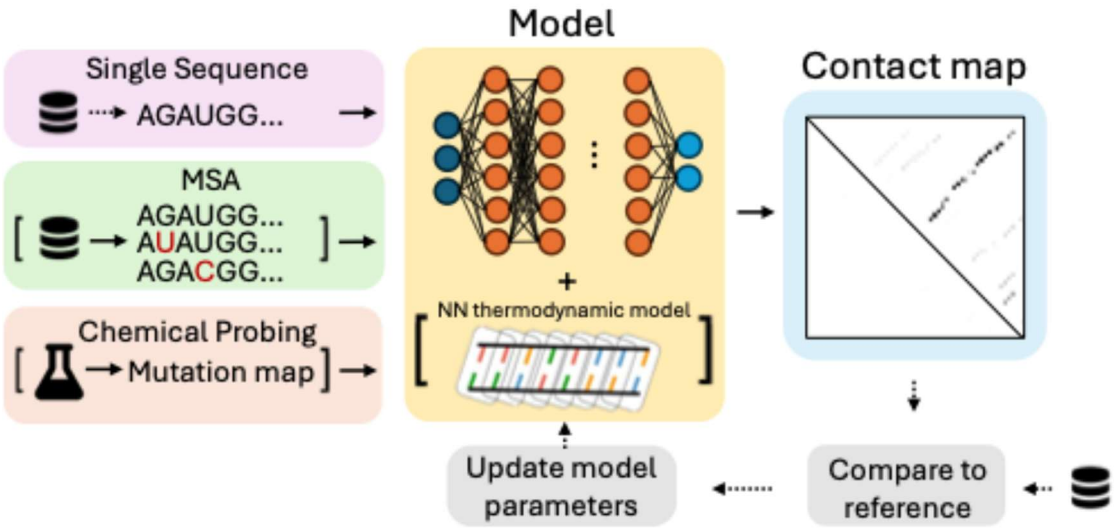


*Figure 2: Schematic representation of deep learning methods for RNA secondary structure prediction (not including foundation models). Dotted arrows indicate steps that are only included in training, and squared brackets indicate optional inputs. Ab initio methods predict structure from a single RNA sequence only; evolutionary methods leverage multiple sequence alignments (MSA) to capture co-evolutionary signals; hybrid methods integrate deep learning with thermodynamic models or experimental data.*

The limitations of classical and early ML methods, together with the success of discriminative, richly parameterized models (e.g., CONTRAfold and ContextFold), set the stage for the current era dominated by deep learning. Deep learning continues this data-driven line, moving from physics-based or statistically-tuned scoring functions to models that learn to predict a base-pairing contact map directly from the input sequence.

We can group these methods by the type of data they integrate. The first category includes ab initio methods that predict structure from a single RNA sequence alone; the second leverages evolutionary information from multiple sequence alignments (MSAs); and the third consists of hybrid methods that combine deep learning with biophysical models or experimental data. Figure 2 illustrates these workflows, from input data through training to final prediction. We dedicate a separate subsection to **Foundation Models** (FMs), a rapidly evolving paradigm that pre-trains large neural networks on massive unlabeled RNA sequence datasets to learn the fundamental "language" of RNA, then fine-tunes them for specific tasks like secondary structure prediction, as shown in Figure 3.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

## Ab Initio (Single Sequence) Prediction

These methods aim to predict the secondary structure using only a single RNA sequence as input, making them universally applicable and crucial for studying orphan RNAs. A pioneering method, **SPOT-RNA** (Singh et al. 2019), was directly inspired by the success of deep learning in protein contact map prediction. It applied the contact-map representation to RNA, modeling secondary structure as an adjacency matrix. The model utilized an ensemble of deep neural networks combining Residual Networks (Resnets) and 2D-Bidirectional LSTMs to predict this matrix, enabling it to model canonical, non-canonical, and pseudoknotted pairs. While it achieved a significant leap in performance on within-family test sets, subsequent independent studies revealed that it struggled to generalize to novel RNA families not seen during training (Szikszai et al. 2022; Justyna et al. 2023). **UFold** (Fu et al. 2022) further advanced the image-based approach with a novel input representation that makes all potential interactions explicit. It converts the sequence into a multi-layered map where each of the 16 possible dinucleotide pairings is represented on its own L x L grid. This "image" is then processed by a U-Net architecture. This proved highly effective, showing substantial performance improvements and superior pseudoknot prediction on within-family datasets, though its ability to generalize to new families was also questioned in later studies (Szikszai et al. 2022). **E2Efold** (Chen et al. 2020) integrated a Transformer model with an "unrolling algorithm," a technique that embeds hard structural constraints directly into the deep learning architecture. While it showed strong performance and high recall for pseudoknots on its original benchmarks, it was later found by multiple independent studies to be highly prone to overfitting, failing to generalize to new RNA families (Sato et al. 2021; Fu et al. 2022). More recently, Transformer-based architectures, inspired by their transformative success in protein structure prediction with AlphaFold (Jumper et al. 2021), have gained prominence. **RNAformer** (Franke et al. 2024) features a lean architecture with axial-attention blocks to efficiently capture long-range dependencies. Its main contribution is a novel homology-aware data pipeline that ensures a clean separation between training and test sets. This rigorous training and evaluation scheme was designed specifically to address the generalization crisis and allowed the model to achieve state-of-the-art performance on cross-family benchmarks, demonstrating strong generalization capabilities. A novel generative approach is taken by **RNADiffFold** (Z. Wang et al. 2025), which uses a discrete diffusion model to progressively denoise an initially random contact map into a final, coherent structure. This process is guided by a conditional control component that fuses features from the sequence and, crucially, from pre-trained foundation models like RNA-FM (which we will discuss in a separate section dedicated to Foundation Models). This method shows competitive performance across both within- and cross-family datasets and aims at capturing dynamic, multi-conformational aspects of RNA structure.

## Evolutionary (MSA-based) Prediction

To harness the powerful signal of co-evolution that proved so effective in classical comparative analysis, some of the most accurate deep learning models integrate

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

information from MSAs. **SPOT-RNA2** (Singh et al. 2021), an evolution of its single-sequence predecessor, exemplifies this approach. It enriches its input features with evolutionary information by incorporating a Position Specific Score Matrix (PSSM) and a two-dimensional Direct Coupling Analysis (DCA) map, both derived from an MSA generated by its 'RNAcmap' (Zhang et al. 2021) pipeline. This allows the model to learn from co-variation signals directly, showing a marked improvement over the original SPOT-RNA, particularly for complex interactions like non-canonical pairs. The method's accuracy scales directly with the number of available homologous sequences, achieving very high accuracy for RNAs with deep MSAs. However, this reliance on evolutionary data is also a key distinction; for orphan RNAs with very few homologous sequences, the original single-sequence SPOT-RNA can be more reliable. Furthermore, the method is computationally demanding and currently limited to sequences shorter than 1000 nucleotides due to its feature generation pipeline. While primarily aimed at 3D structure prediction, the success of **trRosettaRNA** (W. Wang et al. 2023) further underscores the power of this strategy. Its pipeline begins with an initial secondary structure predicted by SPOT-RNA, which is fed into a transformer network along with an MSA. The network then predicts a comprehensive set of 1D and 2D geometric restraints (contacts, distances, and orientations) that are used to guide the final 3D folding. Notably, this process can correct inaccuracies in the initial secondary structure prediction, identifying interactions missed by SPOT-RNA and removing false positives. However, this corrective ability is a double-edged sword; in cases where the initial prediction is already highly accurate, potential conflicts between the different data sources can lead to a slight decrease in the final secondary structure's F1-score. Nevertheless, the high accuracy of its final 3D models is predicated on the high precision of its MSA-driven refinement and expansion of the initial 2D structural information.

## Biophysical (Hybrid) Approaches

These methods seek the best of both worlds, combining the pattern-recognition strengths of deep learning with the rigorous framework of biophysical models. This is often done to improve generalization and ground the "black box" nature of deep learning in established physical principles. One major strategy is to learn a data-driven scoring function that augments or replaces the standard thermodynamic energy parameters. **MXfold2** (Sato et al. 2021) is a prime example of this synergy. It employs a deep neural network to compute four types of folding scores which are then combined with Turner's free energy parameters. A key innovation is its use of "thermodynamic regularization" during training, which encourages the learned scores to remain close to the physical parameters, thereby preventing overfitting. This hybrid approach has proven highly robust in its authors' benchmarks, showing strong performance on datasets of unseen families where purely end-to-end models like E2Efold have struggled. However, other independent, rigorous cross-family evaluations have suggested that it, too, can struggle to generalize, indicating that integrating thermodynamic knowledge is a promising but not complete solution to the overfitting problem (Szikszai et al. 2022). Furthermore, its reliance on a Zuker-like DP algorithm restricts it to predicting non-pseudoknotted structures. A second important hybrid strategy involves using machine learning to integrate diverse experimental data into

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

the folding process. The model proposed by **Calonaci et al.** (Calonaci et al. 2020) is a sophisticated example. It uses a convolutional network to learn a mapping from 1D chemical probing data (e.g., SHAPE) and 2D co-evolutionary data (DCA) to pseudo-energy penalties. These learned penalties are then integrated directly into the RNAfold algorithm. A key advantage of this design is that the entire pipeline is differentiable, allowing the thermodynamic model to be part of the end-to-end training procedure, which was shown to significantly boost the population of the native structure in their tests.

Other methods achieve hybrid status through their algorithmic design or parameterization. It is worth noting that the core innovations of the following methods are not in machine learning themselves, but in their algorithmic or statistical frameworks. They are classified as hybrid because they are designed as flexible engines that can be parameterized by scores from either thermodynamic models or machine learning-based methods. **LinearFold** (Huang et al. 2019), while primarily known for its linear-time complexity achieved via a beam search heuristic, can be parameterized with either traditional thermodynamic energies (LinearFold-V) or scores from machine learning models like CONTRAfold (LinearFold-C), making it a flexible hybrid tool. Finally, some methods combine different scoring schemes at the ensemble level. **CentroidFold** (Sato et al. 2009) operates on the principle of Maximum Expected Accuracy (MEA) rather than MFE. It predicts a "centroid" structure from a Boltzmann-weighted ensemble of possibilities by using a superior "g-centroid estimator." Its flexibility allows it to use parameters from Turner's model, CONTRAfold, or a combination, making it a statistical-mechanical hybrid that has demonstrated improved accuracy over pure MFE methods, though it is also limited to nested structures and has shown issues with generalization in cross-family tests (Szikszai et al. 2022).

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

## Foundation Models

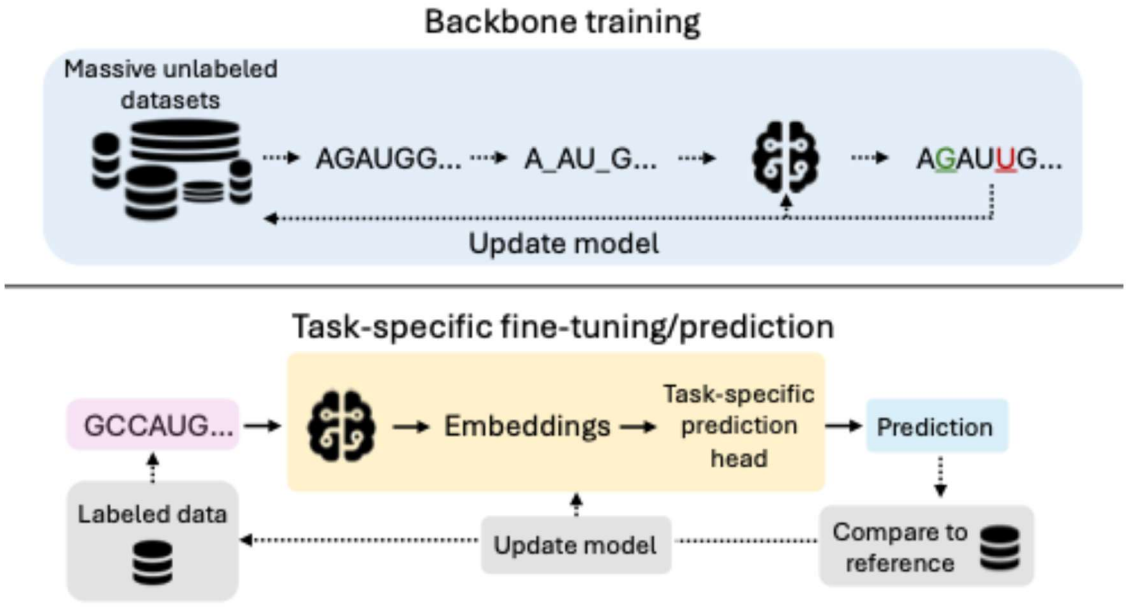


*Figure 3: Schematic representation of backbone training (above) and task-specific fine-tuning/prediction (below) for RNA foundation models. Dotted arrows indicate steps that are only included in training. During backbone training, the model learns general "RNA language" features by predicting masked nucleotides from their surrounding context on massive unlabeled sequence datasets. The pre-trained backbone can then be fine-tuned on smaller, labeled datasets for specific downstream tasks like secondary structure prediction.*

The generalization problem is inextricably linked to the "data bottleneck": the scarcity of diverse, high-quality experimental structures needed for supervised training. Foundation Models (FMs), or RNA Language Models (LMs), represent a recent and rapidly evolving paradigm to address this. Inspired by Large Language Models in natural language processing, this approach pre-trains large neural networks on millions of unlabeled RNA sequences to learn the fundamental "language" of RNA.

The pioneering **RNA-FM** (Chen et al. 2022) set the stage, sparking a fast-paced progression of more sophisticated strategies. To better capture structural information, models like **RNA-MSM** (Zhang et al. 2024) explicitly incorporate co-evolutionary signals from multiple sequence alignments. Others have focused on scaling up, such as **Uni-RNA** (X. Wang et al. 2023) which reportedly used a dataset of one billion sequences, and **HydraRNA** (Li et al. 2025), whose hybrid architecture is designed to handle full-length transcripts. Perhaps the most significant recent direction is structure-aware pre-training, exemplified by **structRFM** (Zhu et al. 2025), which integrates secondary structure information directly into its pre-training objective.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

As this field matures, early comparative reviews are emerging that offer critical insights. "Zero-shot" evaluations, which probe a model's intrinsic structural knowledge before fine-tuning, suggest that single-sequence RNA LMs have yet to achieve the same level of accuracy for structure prediction as their protein counterparts (H. Wang et al. 2025). It has been hypothesized that this is due to the lower information content of the four-letter RNA alphabet, which makes inferring co-evolutionary signals from a single sequence inherently more difficult than from a 20-letter protein alphabet. This challenge has sustained interest in MSA-based approaches, such as the recently developed RNA-MSM, which are designed to directly leverage this crucial evolutionary information and have shown promise in initial benchmarks (H. Wang et al. 2025). The true robustness of all these models, however, will only become clear as they are more widely and independently benchmarked over time.

# Core Deep Learning Challenges: From Dynamics to Interpretability

The integration of deep learning into RNA secondary structure prediction has brought significant gains but also new challenges concerning scientific validation, data scarcity, and the need for interpretable, biologically realistic models.

## Evolving Prediction Targets: From Static Blueprints to Dynamic Ensembles

The central target of RNA structure prediction has evolved beyond the single **Minimum Free Energy (MFE)** structure. This paradigm shift was driven by the recognition that the MFE is often misleading for functionally relevant RNAs, which typically exist as heterogeneous conformational ensembles where the most stable state may represent only a minor subpopulation. This realization necessitated a move towards outputs that could capture this structural diversity.

Thermodynamic models offer a path towards this goal by using partition function algorithms to generate a Boltzmann-weighted sample of structures, approximating the full ensemble with predicted populations. In principle, this is the ideal, full-information output. However, its practical utility is fundamentally constrained by the limitations of the underlying energy model: its imperfect accuracy and its general inability to handle pseudoknots, non-canonical pairs, or the vast chemical diversity of modified nucleotides.

As a more practical, albeit lossy, summary of the ensemble, these models are often used to compute a 2D matrix. This can be a **base-pairing probability matrix** (BPP), which has clear physical interpretability, or a more generalized **contact map**, which is the common output for deep learning models that trade physical meaning for the flexibility to represent any learned interaction (Fu et al. 2022). Both representations, however, obscure the correlational structure of the ensemble; the marginal probability of individual pairs does

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

not capture the co-occurrence or mutual exclusivity of different structural elements, resulting in a significant information loss.

To refine predictions, a powerful strategy is to incorporate external data as **pseudo-energy terms** that perturb the energy landscape, a technique used in hybrid models like **ShapeKnots** or **MXfold2** (Hajdin et al. 2013; Sato et al. 2021). Ultimately, the goal remains the accurate prediction of the full dynamic ensemble. Single-molecule chemical probing (e.g., SHAPE-MaP (Smola et al. 2015) and DMS-MaPseq (Zubradt et al. 2017)) records multiple modifications on individual molecules, preserving co-mutation patterns that reveal coexisting folds. **DREEM** (Tomezsko et al. 2020) clusters the reads obtained from each molecule to separate the ensemble-average profile into a small set of per-conformation reactivity profiles and their abundances; these profiles can then be used to constrain secondary-structure prediction. **DRACO** (Morandi et al. 2021) scales this idea to longer RNAs via a windowed co-mutation graph, spectral model selection, and fuzzy clustering to determine and merge conformations. **DANCE-MaP** (Olson et al. 2022) extends deconvolution by using a maximum likelihood clustering algorithm on MaP sequencing data to fit a Bernoulli mixture model. This approach simultaneously extracts per-nucleotide reactivity, direct base pairing (PAIRs), tertiary interactions (RINGs), and populations for each conformational state. By assigning individual reads to specific states, it enables state-specific correlation analyses that resolve structural features obscured in ensemble-average measurements.

## Emerging Frontiers and Persistent Hurdles

Key challenges remain at the forefront of the field:

### *Chemical Complexity*

Most models operate on a simplified four-letter alphabet, ignoring the more than $150$ known post-transcriptional modifications observed *in vivo*. These modifications are not merely decorative; they directly alter base-pairing potential and are indispensable for stabilizing the complex tertiary folds essential for biological activity (Boccaletto et al. 2018). High-resolution structures (notably rRNAs and tRNAs) do contain numerous modifications, so structural data exist. However, the space of modification chemistries is broad and for most individual modifications the available statistics are sparse, which hampers robust parameterization, benchmarking, and integration into folding models and ML features (Tanzer et al. 2019). Incorporating this chemical diversity remains crucial for biological realism and will require community resources aggregating per-modification thermodynamic and structural effects at scale.

### *Kilobase-Scale RNAs*

Accurately predicting the global architecture of long RNAs ($> 1000$ nt) remains a major challenge, driven by both computational and biological complexity. Computationally, many algorithms scale poorly with sequence length, making predictions intractable. Biologically, the difficulty lies in capturing the correct hierarchy of local structures and the

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

few crucial long-range interactions that define the global fold from a combinatorially vast search space (Szikszai et al. 2022).

### *Interpretability*

Deep learning models often function as “black boxes,” making it difficult to discern if they have learned generalizable biophysical rules or are simply fitting statistical patterns in the data. This is a critical concern, as a lack of interpretability can be linked to poor generalization. Hybrid approaches aim to address this by grounding the model in established physical principles.

### *Pseudoknots and Non-Canonical Pairs*

The accurate prediction of complex structural motifs remains a major hurdle. Predicting pseudoknots within an energy minimization framework is an NP-complete problem, making exact solutions computationally intractable for all but the shortest sequences (Sato et al. 2021). Furthermore, the thermodynamic parameters governing their stability are poorly characterized. Non-canonical base pairs, which are essential for stabilizing tertiary structure, are likewise excluded from most models due to a scarcity of experimental data. While deep learning models are not bound by the same algorithmic constraints and can learn to predict these interactions, their accuracy remains limited, particularly for pseudoknotted base pairs where sensitivity is often low (Fu et al. 2022).

### *Environmental Agnosticism*

Thermodynamic models, by incorporating experimentally measured enthalpy changes, can predict structures at different temperatures (Mathews et al. 1999). In contrast, most machine learning models are trained on sequence-structure pairs without environmental context. They are therefore agnostic to physical parameters like temperature or ion concentration, limiting their ability to predict how an RNA’s structure might change in different cellular or experimental conditions. Also in this case, hybrid models that integrate learned scores with physical energy parameters, represent a promising step toward re-incorporating this biophysical realism.

### *Cellular Context and Ligand Binding*

RNA molecules fold *in vivo* within a crowded cellular environment and their structures are often modulated by interactions with proteins, ions (e.g., $Mg^{2+}$), and small-molecule ligands. Riboswitches, for example, undergo functionally critical conformational changes upon ligand binding. Most prediction methods, particularly single-sequence models, are blind to this context. A key strategy to overcome this is the integration of experimental data from *in vivo* chemical probing (e.g., SHAPE), which implicitly captures the effects of these cellular factors and can guide prediction algorithms toward more biologically relevant structures.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

### *Standardized Prospective Benchmarking*

While data curation and evaluation practices have matured significantly, the field still lacks a community-wide, prospective benchmarking system analogous to the Critical Assessment of protein structure prediction (CASP) (Kryshtafovych et al. 2023). The success and rapid progress in protein folding, including the validation of AlphaFold, were driven in large part by CASP's role as an independent arbiter. Establishing a regular, blind challenge for RNA secondary structure prediction could provide unbiased evaluation of true generalization capabilities, accelerate progress on persistent hurdles like pseudoknots and modified bases, and build community consensus on the genuine state-of-the-art.

## Discussion

The field has shifted from principled yet constrained physics-based energy models to more flexible, data-driven deep-learning approaches. This transition unlocked a new tier of predictive accuracy, yet simultaneously unveiled a profound challenge that has since reshaped the field's priorities: the generalization crisis. The initial enthusiasm for deep learning's performance was rightly tempered by the discovery that many models were not learning the fundamental principles of RNA folding, but rather overfitting to family-specific features within the training data. This rendered them unreliable for their most critical use case: the structural analysis of newly discovered or poorly understood RNAs.

This realization forced a necessary maturation within the community, compelling a move away from simplistic benchmarks toward rigorous, homology-aware validation standards. The widespread adoption of family-based cross-validation has become the new gold standard, ensuring that modern methods are evaluated on their ability to generalize to unseen RNA families, not merely interpolate within known ones. Strategies to address this challenge are now central to the field, including the development of hybrid models that ground learning in thermodynamic principles and the integration of co-evolutionary signals from multiple sequence alignments, which provide a powerful, albeit not universally available, source of structural constraint.

In response to both the generalization problem and the underlying scarcity of high-quality structural data, the current frontier is moving toward the use of foundation models. By pre-training on millions of unlabeled RNA sequences, these models aim to learn the intrinsic "language" of RNA, capturing the statistical patterns that govern its structure and function without direct supervision. This approach promises to create more robust and widely applicable predictors that are less dependent on the limited corpus of experimentally solved structures. While still an emerging area, the development of these models represents a significant hope for breaking through the current data bottleneck.

Despite this progress, formidable challenges remain that will define the next era of research. The field must continue to evolve beyond predicting a single, static Minimum Free Energy structure; the crucial next step is to fully characterize the dynamic,

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

conformational ensembles that define an RNA's functional landscape. Moreover, the accurate prediction of complex motifs like pseudoknots and non-canonical pairs, which are often algorithmically intractable for physics-based models and a weak point for deep learning, remains a major barrier. Similarly, predicting the global architecture of long, kilobase-scale RNAs is largely unsolved, hampered by both computational complexity and the combinatorial explosion of possible long-range interactions. To achieve true biological realism, models must also incorporate the vast chemical diversity of post-transcriptional modifications and account for environmental context, such as ion concentrations, temperature, and in vivo ligand binding, all of which are critical for function but ignored by most current models. Addressing this suite of challenges will require not only new modeling strategies but also a community-wide, prospective benchmarking system, akin to CASP, to ensure unbiased validation and accelerate progress toward capturing the true, dynamic nature of RNA in the cell.

# Acknowledgements

Guido Sanguinetti acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 - Project FAIR "Future Artificial Intelligence Research". This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22].

Andronescu, Mirela, Vera Bereg, Holger H. Hoos, and Anne Condon. 2008. "RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database." *BMC Bioinformatics* 9 (1): 340. https://doi.org/10.1186/1471-2105-9-340.

Andronescu, Mirela, Anne Condon, Holger H. Hoos, David H. Mathews, and Kevin P. Murphy. 2007. "Efficient Parameter Estimation for RNA Secondary Structure Prediction." *Bioinformatics* 23 (13): i19–28. https://doi.org/10.1093/bioinformatics/btm223.

Andronescu, Mirela, Anne Condon, Holger H. Hoos, David H. Mathews, and Kevin P. Murphy. 2010. "Computational Approaches for RNA Energy Parameter Estimation." *RNA (New York, N.Y.)* 16 (12): 2304–18. https://doi.org/10.1261/rna.1950510.

Bellaousov, Stanislav, and David H. Mathews. 2010. "ProbKnot: Fast Prediction of RNA Secondary Structure Including Pseudoknots." *RNA (New York, N.Y.)* 16 (10): 1870–80. https://doi.org/10.1261/rna.2125310.

Bernhart, Stephan H., Ivo L. Hofacker, Sebastian Will, Andreas R. Gruber, and Peter F. Stadler. 2008. "RNAalifold: Improved Consensus Structure Prediction for RNA Alignments." *BMC Bioinformatics* 9 (1): 1–13. https://doi.org/10.1186/1471-2105-9-474.

Boccaletto, Pietro, Magdalena A Machnicka, Elzbieta Purta, et al. 2018. "MODOMICS: A Database of RNA Modification Pathways. 2017 Update." *Nucleic Acids Research* 46 (D1): D303–7. https://doi.org/10.1093/nar/gkx1030.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Bon, Michaël, and Henri Orland. 2011. “TT2NE: A Novel Algorithm to Predict RNA Secondary Structures with Pseudoknots.” *Nucleic Acids Research* 39 (14): e93. https://doi.org/10.1093/nar/gkr240.

Brion, Philippe, and Eric Westhof. 1997. “HIERARCHY AND DYNAMICS OF RNA FOLDING.” *Annual Review of Biophysics* 26 (Volume 26, 1997): 113–37. https://doi.org/10.1146/annurev.biophys.26.1.113.

Calonaci, Nicola, Alisha Jones, Francesca Cuturello, Michael Sattler, and Giovanni Bussi. 2020. “Machine Learning a Model for RNA Structure Prediction.” *NAR Genomics and Bioinformatics* 2 (4): lqaa090. https://doi.org/10.1093/nargab/lqaa090.

Cannone, Jamie J., Sankar Subramanian, Murray N. Schnare, et al. 2002. “The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs.” *BMC Bioinformatics* 3 (1): 2. https://doi.org/10.1186/1471-2105-3-2.

Chen, Jiayang, Zhihang Hu, Siqi Sun, et al. 2022. “Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions.” In *arXiv.org*. Https://arxiv.org/abs/2204.00300v5.

Chen, Ke, Thomas Litfin, Jaswinder Singh, Jian Zhan, and Yaoqi Zhou. 2024. “MARS and RNAcmap3: The Master Database of All Possible RNA Sequences Integrated with RNAcmap for RNA Homology Search.” *Genomics, Proteomics & Bioinformatics* 22 (1): qzae018. https://doi.org/10.1093/gpbjnl/qzae018.

Chen, Xinshi, Yu Li, Ramzan Umarov, Xin Gao, and Le Song. 2020. *RNA Secondary Structure Prediction By Learning Unrolled Algorithms*. arXiv:2002.05810. arXiv. https://doi.org/10.48550/arXiv.2002.05810.

Cruz, José Almeida, and Eric Westhof. 2009. “The Dynamic Landscapes of RNA Architecture.” *Cell* 136 (4): 604–9. https://doi.org/10.1016/j.cell.2009.02.003.

Cuturello, Francesca, Guido Tiana, and Giovanni Bussi. 2020. “Assessing the Accuracy of Direct-Coupling Analysis for RNA Contact Prediction.” *RNA* 26 (5): 637–47. https://doi.org/10.1261/rna.074179.119.

Danaee, Padideh, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, and David Hendrix. 2018. “bpRNA: Large-Scale Automated Annotation and Analysis of RNA Secondary Structure.” *Nucleic Acids Research* 46 (11): 5381–94. https://doi.org/10.1093/nar/gky285.

De Leonardis, Eleonora, Benjamin Lutz, Sebastian Ratz, et al. 2015. “Direct-Coupling Analysis of Nucleotide Coevolution Facilitates RNA Secondary and Tertiary Structure Prediction.” *Nucleic Acids Research* 43 (21): 10444–55. https://doi.org/10.1093/nar/gkv932.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Deigan, Katherine E., Tian W. Li, David H. Mathews, and Kevin M. Weeks. 2009. “Accurate SHAPE-directed RNA Structure Determination.” *Proceedings of the National Academy of Sciences* 106 (1): 97–102. https://doi.org/10.1073/pnas.0806929106.

Ding, Yiliang, Yin Tang, Chun Kit Kwok, Yu Zhang, Philip C. Bevilacqua, and Sarah M. Assmann. 2014. “In Vivo Genome-Wide Profiling of RNA Secondary Structure Reveals Novel Regulatory Features.” *Nature* 505 (7485): 696–700. https://doi.org/10.1038/nature12756.

Do, Chuong B., Daniel A. Woods, and Serafim Batzoglou. 2006. “CONTRAfold: RNA Secondary Structure Prediction Without Physics-Based Models.” *Bioinformatics* 22 (14): e90–98. https://doi.org/10.1093/bioinformatics/btl246.

Doudna, Jennifer A., and Thomas R. Cech. 2002. “The Chemical Repertoire of Natural Ribozymes.” *Nature* 418 (6894): 222–28. https://doi.org/10.1038/418222a.

Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790492.

Eddy, Sean R., and Richard Durbin. 1994. “RNA Sequence Analysis Using Covariance Models.” *Nucleic Acids Research* 22 (11): 2079–88. https://doi.org/10.1093/nar/22.11.2079.

Franke, Jörg K. H., Frederic Runge, Ryan Köksal, Dominika Matus, Rolf Backofen, and Frank Hutter. 2024. *RNAformer: A Simple yet Effective Model for Homology-Aware RNA Secondary Structure Prediction*. bioRxiv. https://doi.org/10.1101/2024.02.12.579881.

Fu, Laiyi, Yingxin Cao, Jie Wu, Qinke Peng, Qing Nie, and Xiaohui Xie. 2022. “UFold: Fast and Accurate RNA Secondary Structure Prediction with Deep Learning.” *Nucleic Acids Research* 50 (3): e14. https://doi.org/10.1093/nar/gkab1074.

Gardner, Paul P., and Robert Giegerich. 2004. “A Comprehensive Comparison of Comparative RNA Structure Prediction Approaches.” *BMC Bioinformatics* 5 (1): 140. https://doi.org/10.1186/1471-2105-5-140.

Hajdin, Christine E., Stanislav Bellaousov, Wayne Huggins, Christopher W. Leonard, David H. Mathews, and Kevin M. Weeks. 2013. “Accurate SHAPE-directed RNA Secondary Structure Modeling, Including Pseudoknots.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (14): 5498–503. https://doi.org/10.1073/pnas.1219988110.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. “Fast Folding and Comparison of RNA Secondary Structures.” *Monatshefte Für Chemie / Chemical Monthly* 125 (2): 167–88. https://doi.org/10.1007/BF00818163.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Holbrook, Stephen R. 2005. “RNA Structure: The Long and the Short of It.” *Current Opinion in Structural Biology* 15 (3): 302–8. https://doi.org/10.1016/j.sbi.2005.04.005.

Holbrook, Stephen R. 2008. “Structural Principles From Large RNAs*.” *Annual Review of Biophysics* 37 (Volume 37, 2008): 445–64. https://doi.org/10.1146/annurev.biophys.36.040306.132755.

Huang, Liang, He Zhang, Dezhong Deng, et al. 2019. “LinearFold: Linear-Time Approximate RNA Folding by 5’-to-3’ Dynamic Programming and Beam Search.” *Bioinformatics* 35 (14): i295–304. https://doi.org/10.1093/bioinformatics/btz375.

Jumper, John, Richard Evans, Alexander Pritzel, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Justyna, Marek, Maciej Antczak, and Marta Szachniuk. 2023. “Machine Learning for RNA 2D Structure Prediction Benchmarked on Experimental Data.” *Briefings in Bioinformatics* 24 (3): bbad153. https://doi.org/10.1093/bib/bbad153.

Kalvari, Ioanna, Eric P Nawrocki, Nancy Ontiveros-Palacios, et al. 2021. “Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families.” *Nucleic Acids Research* 49 (D1): D192–200. https://doi.org/10.1093/nar/gkaa1047.

Kerpedjiev, Peter, Stefan Hammer, and Ivo L. Hofacker. 2015. “Forna (Force-Directed RNA): Simple and Effective Online RNA Secondary Structure Diagrams.” *Bioinformatics (Oxford, England)* 31 (20): 3377–79. https://doi.org/10.1093/bioinformatics/btv372.

Knudsen, Bjarne, and Jotun Hein. 2003. “Pfold: RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars.” *Nucleic Acids Research* 31 (13): 3423–28. https://doi.org/10.1093/nar/gkg614.

Kopp, Florian, and Joshua T. Mendell. 2018. “Functional Classification and Experimental Dissection of Long Noncoding RNAs.” *Cell* 172 (3): 393–407. https://doi.org/10.1016/j.cell.2018.01.011.

Kryshtafovych, Andriy, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. 2023. “Critical Assessment of Methods of Protein Structure Prediction (CASP)-Round XV.” *Proteins* 91 (12): 1539–49. https://doi.org/10.1002/prot.26617.

Li, Guipeng, Feifei Jiang, Junhao Zhu, Huanhuan Cui, Zefeng Wang, and Wei Chen. 2025. *HydraRNA: A Hybrid Architecture Based Full-Length RNA Language Model*. bioRxiv. https://doi.org/10.1101/2025.03.06.641765.

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner zu Siederdissen, et al. 2011. “ViennaRNA Package 2.0.” *Algorithms for Molecular Biology* 6 (1): 1–14. https://doi.org/10.1186/1748-7188-6-26.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Markham, Nicholas R., and Michael Zuker. 2008. “UNAFold.” In *Bioinformatics: Structure, Function and Applications*, edited by Jonathan M. Keith. Humana Press. https://doi.org/10.1007/978-1-60327-429-6_1.

Mathews, David H., Matthew D. Disney, Jessica L. Childs, Susan J. Schroeder, Michael Zuker, and Douglas H. Turner. 2004. “Incorporating Chemical Modification Constraints into a Dynamic Programming Algorithm for Prediction of RNA Secondary Structure.” *Proceedings of the National Academy of Sciences* 101 (19): 7287–92. https://doi.org/10.1073/pnas.0401799101.

Mathews, David H., Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. 1999. “Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure1.” *Journal of Molecular Biology* 288 (5): 911–40. https://doi.org/10.1006/jmbi.1999.2700.

Mathews, David H., and Douglas H. Turner. 2002. “Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences.” *Journal of Molecular Biology* 317 (2): 191–203. https://doi.org/10.1006/jmbi.2001.5351.

Mattick, J. S. 2001. “Non-Coding RNAs: The Architects of Eukaryotic Complexity.” *EMBO Reports* 2 (11): 986–91. https://doi.org/10.1093/embo-reports/kve230.

Merino, Edward J., Kevin A. Wilkinson, Jennifer L. Coughlan, and Kevin M. Weeks. 2005. “RNA Structure Analysis at Single Nucleotide Resolution by Selective 2‘-Hydroxyl Acylation and Primer Extension (SHAPE).” *Journal of the American Chemical Society* 127 (12): 4223–31. https://doi.org/10.1021/ja043822v.

Morandi, Edoardo, Ilaria Manfredonia, Lisa M. Simon, et al. 2021. “Genome-Scale Deconvolution of RNA Structure Ensembles.” *Nature Methods* 18 (3): 249–52. https://doi.org/10.1038/s41592-021-01075-w.

Morris, Kevin V., and John S. Mattick. 2014. “The Rise of Regulatory RNA.” *Nature Reviews Genetics* 15 (6): 423–37. https://doi.org/10.1038/nrg3722.

Mustoe, Anthony M., Steven Busan, Greggory M. Rice, et al. 2018. “Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing.” *Cell* 173 (1): 181–195.e18. https://doi.org/10.1016/j.cell.2018.02.034.

Nawrocki, Eric P., and Sean R. Eddy. 2013. “Infernal 1.1: 100-Fold Faster RNA Homology Searches.” *Bioinformatics* 29 (22): 2933–35. https://doi.org/10.1093/bioinformatics/btt509.

Nussinov, Ruth, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. 1978. “Algorithms for Loop Matchings.” *SIAM Journal on Applied Mathematics* 35 (1): 68–82. https://doi.org/10.1137/0135006.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Olson, Samuel W., Anne-Marie W. Turner, J. Winston Arney, et al. 2022. “Discovery of a Large-Scale, Cell-State-Responsive Allosteric Switch in the 7SK RNA Using DANCE-MaP.” *Molecular Cell* 82 (9): 1708–1723.e10. https://doi.org/10.1016/j.molcel.2022.02.009.

Peattie, D. A., and W. Gilbert. 1980. “Chemical Probes for Higher-Order Structure in RNA.” *Proceedings of the National Academy of Sciences of the United States of America* 77 (8): 4679–82. https://doi.org/10.1073/pnas.77.8.4679.

Pedersen, Jakob Skou, Gill Bejerano, Adam Siepel, et al. 2006. “Identification and Classification of Conserved RNA Secondary Structures in the Human Genome.” *PLoS Computational Biology* 2 (4): e33. https://doi.org/10.1371/journal.pcbi.0020033.

Peng, W-X, P Koirala, and Y-Y Mo. 2017. “LncRNA-mediated Regulation of Cell Signaling in Cancer.” *Oncogene* 36 (41): 5661–67. https://doi.org/10.1038/onc.2017.184.

Pucci, Fabrizio, Mehari B. Zerihun, Emanuel K. Peter, and Alexander Schug. 2020. “Evaluating DCA-based Method Performances for RNA Contact Prediction by a Well-Curated Data Set.” *RNA* 26 (7): 794–802. https://doi.org/10.1261/rna.073809.119.

Reuter, Jessica S., and David H. Mathews. 2010. “RNAstructure: Software for RNA Secondary Structure Prediction and Analysis.” *BMC Bioinformatics* 11 (1): 1–9. https://doi.org/10.1186/1471-2105-11-129.

Rivas, Elena, Raymond Lang, and Sean R. Eddy. 2012. “A Range of Complex Probabilistic Models for RNA Secondary Structure Prediction That Includes the Nearest-Neighbor Model and More.” *RNA* 18 (2): 193–212. https://doi.org/10.1261/rna.030049.111.

Rouskin, Silvi, Meghan Zubradt, Stefan Washietl, Manolis Kellis, and Jonathan S. Weissman. 2014. “Genome-Wide Probing of RNA Structure Reveals Active Unfolding of mRNA Structures in Vivo.” *Nature* 505 (7485): 701–5. https://doi.org/10.1038/nature12894.

Sakakibara, Yasubumi, Michael Brown, Richard Hughey, et al. 1994. “Stochastic Context-Free Grammers for tRNA Modeling.” *Nucleic Acids Research* 22 (23): 5112–20. https://doi.org/10.1093/nar/22.23.5112.

Saman Booy, Mehdi, Alexander Ilin, and Pekka Orponen. 2022. “RNA Secondary Structure Prediction with Convolutional Neural Networks.” *BMC Bioinformatics* 23 (1): 58. https://doi.org/10.1186/s12859-021-04540-7.

Sankoff, David. 2006. “Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems.” *SIAM Journal on Applied Mathematics*, ahead of print, July. https://doi.org/10.1137/0145048.

Sato, Kengo, Manato Akiyama, and Yasubumi Sakakibara. 2021. “RNA Secondary Structure Prediction Using Deep Learning with Thermodynamic Integration.” *Nature Communications* 12 (1): 941. https://doi.org/10.1038/s41467-021-21194-4.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Sato, Kengo, and Michiaki Hamada. 2023. “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* 24 (4): bbad186. https://doi.org/10.1093/bib/bbad186.

Sato, Kengo, Michiaki Hamada, Kiyoshi Asai, and Toutai Mituyama. 2009. “CentroidFold: A Web Server for RNA Secondary Structure Prediction.” *Nucleic Acids Research* 37 (suppl_2): W277–80. https://doi.org/10.1093/nar/gkp367.

Singh, Jaswinder, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. 2019. “RNA Secondary Structure Prediction Using an Ensemble of Two-Dimensional Deep Neural Networks and Transfer Learning.” *Nature Communications* 10 (1): 5407. https://doi.org/10.1038/s41467-019-13395-9.

Singh, Jaswinder, Kuldip Paliwal, Tongchuan Zhang, Jaspreet Singh, Thomas Litfin, and Yaoqi Zhou. 2021. “Improved RNA Secondary Structure and Tertiary Base-Pairing Prediction Using Evolutionary Profile, Mutational Coupling and Two-Dimensional Transfer Learning.” *Bioinformatics* 37 (17): 2589–600. https://doi.org/10.1093/bioinformatics/btab165.

Smola, Matthew J., Greggory M. Rice, Steven Busan, Nathan A. Siegfried, and Kevin M. Weeks. 2015. “Selective 2′-Hydroxyl Acylation Analyzed by Primer Extension and Mutational Profiling (SHAPE-MaP) for Direct, Versatile and Accurate RNA Structure Analysis.” *Nature Protocols* 10 (11): 1643–69. https://doi.org/10.1038/nprot.2015.103.

Staple, David W., and Samuel E. Butcher. 2005. “Pseudoknots: RNA Structures with Diverse Functions.” *PLOS Biology* 3 (6): e213. https://doi.org/10.1371/journal.pbio.0030213.

Statello, Luisa, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. 2021. “Gene Regulation by Long Non-Coding RNAs and Its Biological Functions.” *Nature Reviews Molecular Cell Biology* 22 (2): 96–118. https://doi.org/10.1038/s41580-020-00315-9.

Strobel, Eric J., Angela M. Yu, and Julius B. Lucks. 2018. “High-Throughput Determination of RNA Structures.” *Nature Reviews Genetics* 19 (10): 615–34. https://doi.org/10.1038/s41576-018-0034-x.

Sükösd, Zsuzsanna, M. Shel Swenson, Jørgen Kjems, and Christine E. Heitsch. 2013. “Evaluating the Accuracy of SHAPE-directed RNA Secondary Structure Predictions.” *Nucleic Acids Research* 41 (5): 2807–16. https://doi.org/10.1093/nar/gks1283.

Sundfeld, Daniel, Jakob H. Havgaard, Alba C. M. A. de Melo, and Jan Gorodkin. 2016. “Foldalign 2.5: Multithreaded Implementation for Pairwise Structural RNA Alignment.” *Bioinformatics* 32 (8): 1238–40. https://doi.org/10.1093/bioinformatics/btv748.

Szikszai, Marcell, Michael Wise, Amitava Datta, Max Ward, and David H Mathews. 2022. “Deep Learning Models for RNA Secondary Structure Prediction (Probably) Do Not

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Generalize Across Families." *Bioinformatics* 38 (16): 3892–99. https://doi.org/10.1093/bioinformatics/btac415.

Tan, Zhen, Yinghan Fu, Gaurav Sharma, and David H. Mathews. 2017. "TurboFold II: RNA Structural Alignment and Secondary Structure Prediction Informed by Multiple Homologs." *Nucleic Acids Research* 45 (20): 11570–81. https://doi.org/10.1093/nar/gkx815.

Tanzer, Andrea, Ivo L. Hofacker, and Ronny Lorenz. 2019. "RNA Modifications in Structure Prediction – *Status Quo* and Future Challenges." *Methods*, Mining the Epitranscriptome: Detection of RNA editing and RNA modifications, vol. 156 (March): 32–39. https://doi.org/10.1016/j.ymeth.2018.10.019.

The RNAcentral Consortium. 2019. "RNAcentral: A Hub of Information for Non-Coding RNA Sequences." *Nucleic Acids Research* 47 (D1): D221–29. https://doi.org/10.1093/nar/gky1034.

Tinoco, I., P. N. Borer, B. Dengler, et al. 1973. "Improved Estimation of Secondary Structure in Ribonucleic Acids." *Nature: New Biology* 246 (150): 40–41. https://doi.org/10.1038/newbio246040a0.

Tinoco, Ignacio, and Carlos Bustamante. 1999. "How RNA Folds." *Journal of Molecular Biology* 293 (2): 271–81. https://doi.org/10.1006/jmbi.1999.3001.

Tomezsko, Phillip J., Vincent D. A. Corbin, Paromita Gupta, et al. 2020. "Determination of RNA Structural Diversity and Its Role in HIV-1 RNA Splicing." *Nature* 582 (7812): 438–42. https://doi.org/10.1038/s41586-020-2253-5.

Turner, Douglas H., and David H. Mathews. 2010. "NNDB: The Nearest Neighbor Parameter Database for Predicting Stability of Nucleic Acid Secondary Structure." *Nucleic Acids Research* 38 (suppl_1): D280–82. https://doi.org/10.1093/nar/gkp892.

Wang, He, Yikun Zhang, Jie Chen, Jian Zhan, and Yaoqi Zhou. 2025. *A Comparative Review of RNA Language Models*. arXiv:2505.09087. arXiv. https://doi.org/10.48550/arXiv.2505.09087.

Wang, Wenkai, Chenjie Feng, Renmin Han, et al. 2023. "trRosettaRNA: Automated Prediction of RNA 3D Structure with Transformer Network." *Nature Communications* 14 (1): 7266. https://doi.org/10.1038/s41467-023-42528-4.

Wang, Xi, Ruichu Gu, Zhiyuan Chen, et al. 2023. *Uni-Rna: Universal Pre-Trained Models Revolutionize Rna Research*. bioRxiv. https://doi.org/10.1101/2023.07.11.548588.

Wang, Zhen, Yizhen Feng, Qingwen Tian, Ziqi Liu, Pengju Yan, and Xiaolin Li. 2025. "RNADiffFold: Generative RNA Secondary Structure Prediction Using Discrete Diffusion Models." *Briefings in Bioinformatics* 26 (1): bbae618. https://doi.org/10.1093/bib/bbae618.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

Weeks, Kevin M. 2010. “Advances in RNA Structure Analysis by Chemical Probing.” *Current Opinion in Structural Biology* 20 (3): 295–304. https://doi.org/10.1016/j.sbi.2010.04.001.

Wu, Ping, Xialin Zuo, Houliang Deng, Xiaoxia Liu, Li Liu, and Aimin Ji. 2013. “Roles of Long Noncoding RNAs in Brain Development, Functional Diversification and Neurodegenerative Diseases.” *Brain Research Bulletin* 97 (August): 69–80. https://doi.org/10.1016/j.brainresbull.2013.06.001.

Zakov, Shay, Yoav Goldberg, Michael Elhadad, and Michal Ziv-ukelson. 2011. “Rich Parameterization Improves RNA Structure Prediction.” *Journal of Computational Biology* 18 (11): 1525–42. https://doi.org/10.1089/cmb.2011.0184.

Zhang, Tongchuan, Jaswinder Singh, Thomas Litfin, Jian Zhan, Kuldip Paliwal, and Yaoqi Zhou. 2021. “RNAcmap: A Fully Automatic Pipeline for Predicting Contact Maps of RNAs by Evolutionary Coupling Analysis.” *Bioinformatics (Oxford, England)* 37 (20): 3494–500. https://doi.org/10.1093/bioinformatics/btab391.

Zhang, Yikun, Mei Lang, Jiuhong Jiang, et al. 2024. “Multiple Sequence Alignment-Based RNA Language Model and Its Application to Structural Inference.” *Nucleic Acids Research* 52 (1): e3. https://doi.org/10.1093/nar/gkad1031.

Zhu, Heqin, Ruifeng Li, Feng Zhang, et al. 2025. *A Fully-Open Structure-Guided RNA Foundation Model for Robust Structural and Functional Inference*. bioRxiv. https://doi.org/10.1101/2025.08.06.668731.

Zubradt, Meghan, Paromita Gupta, Sitara Persad, Alan M. Lambowitz, Jonathan S. Weissman, and Silvi Rouskin. 2017. “DMS-MaPseq for Genome-Wide or Targeted RNA Structure Probing in Vivo.” *Nature Methods* 14 (1): 75–82. https://doi.org/10.1038/nmeth.4057.

Zuker, Michael. 2003. “Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction.” *Nucleic Acids Research* 31 (13): 3406–15. https://doi.org/10.1093/nar/gkg595.

Zuker, Michael, and Patrick Stiegler. 1981. “Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information.” *Nucleic Acids Research* 9 (1): 133–48. https://doi.org/10.1093/nar/9.1.133.

Downloaded from rnajournal.cshlp.org on January 26, 2026 . Published by Cold Spring Harbor Laboratory Press

# Machine Learning for RNA Secondary Structure Prediction: a review of current methods and challenges

Giuseppe Sacco, Giovanni Bussi and Guido Sanguinetti

*RNA* published online January 23, 2026

**P<P** Published online January 23, 2026 in advance of the print journal.

**Accepted Manuscript** Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

**Open Access** Freely available online through the *RNA* Open Access option.

**Creative Commons License** This article, published in *RNA*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *RNA* go to:
**http://rnajournal.cshlp.org/subscriptions**

Published by Cold Spring Harbor Laboratory Press for the RNA Society