All-at-once RNA folding with 3D motif prediction framed by evolutionary information

Aayush Karan* & Elena Rivas*†
Department of Molecular and Cellular Biology,
Harvard University, Cambridge, MA 02138, USA

Abstract

Structural RNAs exhibit a vast array of recurrent short 3D elements found in loop regions involving non-Watson-Crick interactions that help arrange canonical double helices into tertiary structures. We present CaCoFold-R3D, a probabilistic grammar that predicts these RNA 3D motifs (also termed modules) jointly with RNA secondary structure over a sequence or alignment. CaCoFold-R3D uses evolutionary information present in an RNA alignment to reliably identify canonical helices (including pseudoknots) by covariation. We further introduce the R3D grammars, which also exploit helix covariation that constrains the positioning of the mostly non-covarying RNA 3D motifs. Our method runs predictions over an almost-exhaustive list of over fifty known RNA motifs (*everything*). Motifs can appear in any non-helical loop region (including 3-way, 4-way and higher junctions) (*everywhere*). All structural motifs as well as the canonical helices are arranged into one single structure predicted by one single joint probabilistic grammar (*all-at-once*). Our results demonstrate that CaCoFold-R3D is a valid alternative for predicting the all-residue interactions present in a RNA 3D structure. CaCoFold-R3D is fast and easily customizable for novel motif discovery and shows promising value both as a strong input for deep learning approaches to all-atom structure prediction as well as towards guiding RNA design as drug targets for therapeutic small molecules.

* Equal contribution.
Contact: †elenarivas@fas.harvard.edu

Availability: The source code can be downloaded from the website rivaslab.org, the git https://github.com/EddyRivasLab/R-scape, as well as from the supplementary materials associated to this manuscript.
Supplementary information: Supplementary materials (data and code) are provided with this manuscript, and at rivaslab.org.

Introduction

Many noncoding RNAs (ncRNAs) play essential roles in cellular processes by means of conserved 3D structures [28]. Accurately determining the 3D structure of an RNA is a window into inferring its molecular mechanism of action.

RNA structure is hierarchical. Canonical base pairs (cis-Watson-Crick A:U, G:C and G:U wobble pairs) stack together as double helices and pseudoknots, forming the secondary structure. Critical loops and junctions connect these helices and arrange them into a 3D structure. These non-helical linker regions, called RNA 3D motifs [47] or modules [14], have been extensively studied in the literature [91, 98, 36, 97, 45, 46, 2, 58, 89, 30, 59, 39, 32, 18, 24, 16] for their importance in accurately characterizing full RNA structure. RNA 3D motifs have recurrent properties: they are typically short; they include recurrent patterns of non-Watson-Crick base pairs resulting in complex and distinctive 3D architectures; and often they also display conserved sequence patterns. Their structural properties are usually independent of the helices they are connected to; thus, identifying 3D motifs alongside secondary structure provides important additive clues that guide the assembly of a full RNA structure from its sequence.

RNA 3D motifs (modules) are inherently difficult to detect due to their short size (often between 4 to 20 nucleotides), sequence variability within motif types, and their sheer variety (more than 30 well categorized motifs have been identified in RNA crystal structures [2] from the PDB [5]). They can also be discontinuous in linear sequence, and they can appear in internal loops or junctions where the fragments composing the motif are hundreds of nucleotides apart. Important efforts have been developed to extract RNA 3D motifs from

crystal structures, and to create databases of RNA 3D motifs, such as: RAG [25, 103], FR3D Motif library [80], RNA3DMotif [18], RNA FRABASE [66], RNA 3D Motif Atlas [65], RNA Bricks [10], CaRNAval [68], LORA [6], D-ORB [19], and ARTEM [3]. Based on this knowledge, several important efforts exist to predict RNA 3D motifs from sequence such as RMDetect [14], JAR3D [104], RMfam [27], and BayesPairing2 [79].

However, these methods are not fully integrated with secondary structure prediction. Several methods [14, 104, 79, 52] are indirectly guided by secondary structures predicted by standard thermodynamic methods [51, 71]. But because those thermodynamic methods cannot incorporate similar parameters for the 3D motifs, the prediction of motifs cannot be integrated together with that of canonical base pairs. In fact, the inputs required can be quite strong: e.g., [104] requires that the loop regions testing for the presence of motifs are provided, while [79] trains over annotated motifs in one family for prediction, getting the most competitive results only when the train and test family are the same. Furthermore, previous techniques [14, 27, 104] are computationally expensive, making independent predictions for one motif at a time. This also restricts the diversity of motifs predicted over, often relegated to hairpin and internal loop motifs [14, 104].

Here, we introduce CaCoFold-R3D, a computationally fast probabilistic model that simultaneously predicts the joint RNA 3D motifs and secondary structure present in a structural RNA. CaCoFold-R3D is grounded on the power of covariation in alignments as inputs. While covariation is not prominent in RNA 3D motifs, the covariation found in canonical helices constraints the space where these 3D motifs can occur, and R-scape's covariation analysis [74] assigns statistical significance as to whether its predictions are evolutionarily conserved RNA structures [73]. Methods such as RMDetect [14] and BayesPairing2 [79] also use alignments, but they do not provide statistical significance for their predictions.

Another important feature of CaCoFold-R3D is the exclusive use of probabilistic modeling which naturally facilitates the integration of the prediction of RNA 3D motifs with that of the RNA secondary structure. Several existing methods use probabilistic modeling of RNA 3D motifs, but they do not integrate those with the predictions of canonical base pairs [14, 27, 104]. CaCoFold-R3D deploys an array of stochastic context-free grammars (SCFGs), to model the structural architecture, and profile hidden Markov models (HMMs), to model sequence homology, that incorporate a large variety of motifs–accounting for sequence variability, we predict over 96 motifs total present in any loop region including hairpins, bulges, internal loops, and multiloops. In addition, the CaCoFold-R3D grammar is designed to generate not just individual sequences but probabilistic sequences representing the columns of an alignment. This important feature allows the modeling of sequence variations within the motif.

CaCoFold-R3D serves as a structural paradigm for a new class of probabilistic RNA folding algorithms that directly integrates the prediction of multiple RNA 3D motifs with that of canonical helices, as well as triplets and other long-range interactions, all of that constrained by the covariation found in the input alignments.

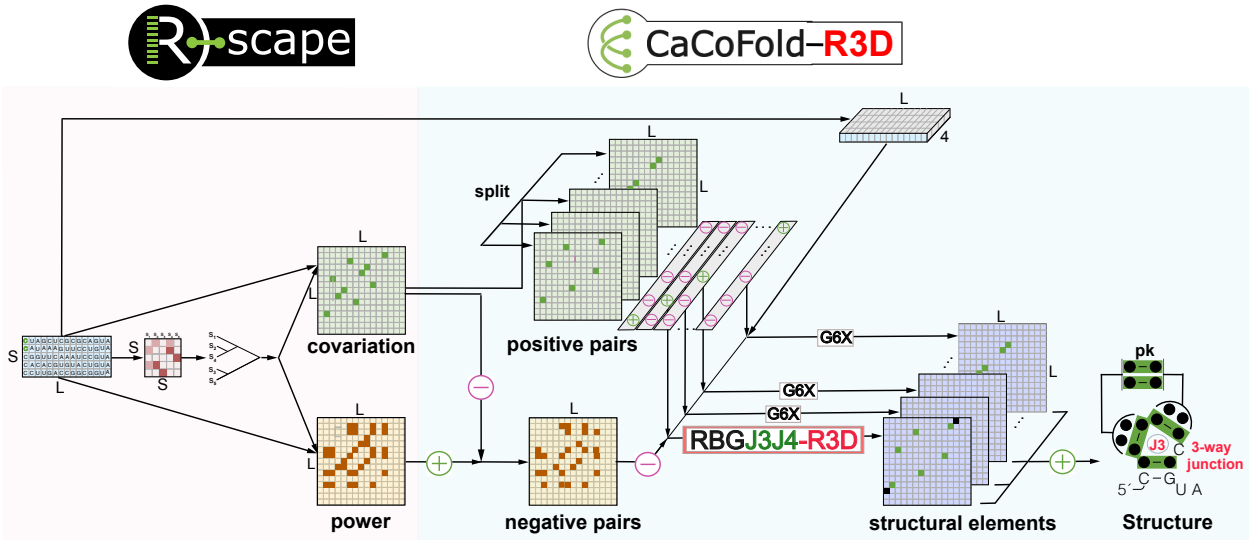
Results

CaCoFold-R3D: Prediction of RNA 3D motifs constrained by covariation

Figure 1 describes the overall CaCoFold-R3D method. The input is a sequence or alignment, and the output is an RNA structure that includes RNA 3D motifs, canonical helices (both nested and pseudoknotted), as well as other tertiary base pairing interactions, provided that they have covariation evidence.

From an alignment, R-scape identifies a set of positive base pairs that significantly covary above phylogenetic expectation and a set of negative pairs that are not expected to form because their variability is not reflective of them being base paired [74, 75]. We have previously shown that the accuracy of RNA structure prediction improves significantly by using covariation information as prediction constraints [72]. Crucially though, CaCoFold-R3D not only uses covariation to constrain secondary structure prediction, but it further uses covariation-bound secondary structure to further constrain the location of RNA 3D motifs via an integrated stochastic context-free grammar (SCFG).

Specifically, CaCoFold-R3D splits the covarying pairs into layers each with the maximum number of nested pairs until all positive pairs have been taken into account. The first layer includes the maximal number of covarying nested base pairs, and is folded into the main secondary structure. The rest of the layers are expected to identify helices of pseudoknotted canonical helices and other tertiary base pair interactions provided that they have covariation support. CaCoFold-R3D introduces a novel SCFG called RGBJ3J4-R3D to describe the first layer where the main structure is predicted. RGBJ3J4-R3D jointly infers the collection of nested canonical helices along with the RNA 3D motifs found within the loop regions (Figure 1) via a maximum probability parsing facilitated by dynamic programming.



‘

Figure 1: **The CaCoFold-R3D algorithm for the joint prediction of 3D RNA structural motifs integrated with canonical RNA helices.** We show the end-to-end method using a toy alignment of length $L = 15$ and $S = 5$ sequence. R-scape identifies significantly covarying (above phylogenetic expectation) base pairs, the positive pairs, as well as negative pairs that have evidence against being base paired. CaCoFold-R3D produces a structure that includes canonical helices as well as 3D motifs using a layered approach that uses different probabilistic RNA folding grammars. The first layer includes the largest set of positive pairs that are nested with each other and uses the RGBJ3J4-R3D grammar to predict a secondary structure including 3D motifs. The rest of the layers use the G6X grammar [75] which adds pseudoknots and other tertiary base pair interactions with covariation evidence. The toy alignment has five significantly covarying base pairs (green) and 13 negative base pairs. CaCoFold-R3D needs to use two layers, and the resulting structure includes three nested helices, another helix forming a pseudoknot, and has one annotated 3-way junction “J3”. The black dots in the consensus structure indicate that there is not a consensus nucleotide for that position.

Comparison to other methods

Our method builds on several previous methods to identify RNA structural elements beyond Watson-Crick base pairs forming a nested secondary structure.

Some methods attempt to extend the repertoire of interactions beyond canonical Watson-Crick base pairs. Methods such as MC-Fold [62] and RNA-MoIP [99] (which includes MC-Fold) build a secondary structure with a model that introduces special pseudo-energy scores for some small whole hairpin loops and internal loops. However, the output does not identify any structural 3D motifs per se, and for example, it would not be able to distinguish a K-turn from any other 2x5 internal loop. Another method that includes non-WC interactions is the method RNAwolf [105] which also introduces terms to predict some triplets. However, RNAwolf cannot tell whether those triplets belong or constitute any of the known 3D motifs. CaCoFold directly predicts 3D motifs, thus a direct comparison is not possible.

Methods that try to identify specific 3D motifs include JAR3D [104] and RMfam [27]. These methods require to be given which are the loops in which to look for the 3D motifs, and the search is done for a small number of motifs and one at the time.

The closest methods to CaCoFold-3D are RMDetect [14] and BayesPairing2 [79]. These methods model specific 3D motifs, but the number of motifs predicted is limited, and the motifs are identified one at the time. Both methods require the previous specification or prediction of a secondary structure. The methods are also computationally expensive. BayesPairing2, the fastest of the two approaches, has to process RNAs longer 300 nucleotides by windowing. CaCoFold-R3D integrates the prediction of the 3D motifs with that of the secondary structure, and it can deal with the largest rRNA molecules. With respect to methodology to describe the 3D motifs, RMDetect using Bayesian Networks which precisely describe all non-Watson-Crick in the motif. This approach however is computationally expensive. BayesPairing2 improves time performance by using much faster regular expression to preselect loops in which to apply the Bayesian Networks. Alternatively, CaCoFold-R3D uses SCFGs with profile HMMs to model the motifs. These SCFGs lack the detail of the Bayesian Networks of describing every single base pair, but still describe correlations between loops effectively, and importantly they can be integrated into standard dynamical programming algorithms for RNA secondary structure prediction

such as CaCoFold. Thus resulting on CaCoFold-R3D has the same time complexity than standard method for just secondary structure prediction.

CaCoFold-R3D has a collection of uniquely defining properties: (1) The method can handle any 3D motif occurring in hairpin loops, internal loops, 3-way and 4-way junctions, as well as independent sequence motifs in branches of higher-order junctions. (2) All motifs are predicted at once and under one unique probabilistic model. (3) The model can folds entire alignments, taking into account RNA 3D motif sequence variability even within a given structural RNA family. (4) The method is computationally efficient, and (5) it has the potential to facilitate novel motif discovery. (6) Importantly, CaCoFold-R3D uses the evolutionary signal found in the alignment to inform the prediction both of canonical helices and 3D motifs which significantly improves performance, and provides statistical confidence on the predictions depending on the amount of covariation observed in the helices bounding the motif.

RBGJ3J4-R3D: Joint prediction of nested helices and 3D motifs with one single SCFG

The RGBJ3J4-R3D model described in Figure 2 is an SCFG that simultaneously infers the secondary structure of nested canonical helices as well as the RNA 3D motifs present in any of the loop regions. It combines together a grammar called RBGJ3J4 (Methods and supplemental Figure S1) with a library called R3D of RNA 3D motif grammars described in the next Section. RBGJ3J4 is unique in that it has specific descriptions for 3-way and 4-way junctions which are the most frequent of the multiloop structures found in RNA structures, which form many different RNA 3D motifs present in important RNA molecules such as the hammerhead 3-way junction [31] and the four-way junction of the hepatitis C virus IRES [55].

RGBJ3J4-R3D creates specific R3D grammar models (*i.e.* grammar non-terminals) for each of the different loop motifs. To incorporate these motif non-terminals into the RBGJ3J4 grammar, we simply add the motif SCFGs are additional productions along with a generic loop motif (Figure 2). Motif designs are added for six classes of loops: hairpin (HL), bulge (BL), and internal loops (IL) as well as 3-way (J3) and 4-way (J4) junctions, and general branch motifs (BS) that can appear in any branch of any higher order multiloop.

Training of the RBGJ3J4 probabilistic parameters. The RBGJ3J4 grammar (Figure 2a) depends on parameters that describe the probabilities of the different rules allowed for a given nonterminal, as well as emission probabilities to describe single unpaired or in loops residues, as well as paired and stacked paired residues. As the RBG grammar in CaCoFold [72], the RBGJ3J4 grammar has been trained by maximum likelihood using TORNADO [76] on a large and diverse dataset of known RNA structures and sequences. This dataset (TrainSetA + 2× TrainSetB) introduced in TORNADO includes a total of over different 4,000 sequence/structure pairs, that include over 100 different RNA structures (for more details, see Figure 2 in TORNADO [76]). The large diversity of the training set guarantees that SCFGs trained on it do not overfit (on a TestSetA+TestSetB also introduced in TORNADO) relative to thermodynamic models not trained on sequence data (Figure 5 in TORNADO [76]).

The main difference between the RBG and RBGJ3J4 grammars is in the multiloop (`ML`) nonterminal (Figure S1). The RBG single rule for `ML` splits into three rules for RBGJ3J4 representing the relative frequencies of J3, J4 versus other higher order junctions. For the training set above, 51% of all multiloops are 3-way junctions, and 32% are 4-way junctions, which stresses the importance of these lower-order junctions amongst all possible multiloops. The collection of all probabilistic parameters for the RBGJ3J4 and RBG SCFGs are provided as part of the code in file `covgrammars.c`.

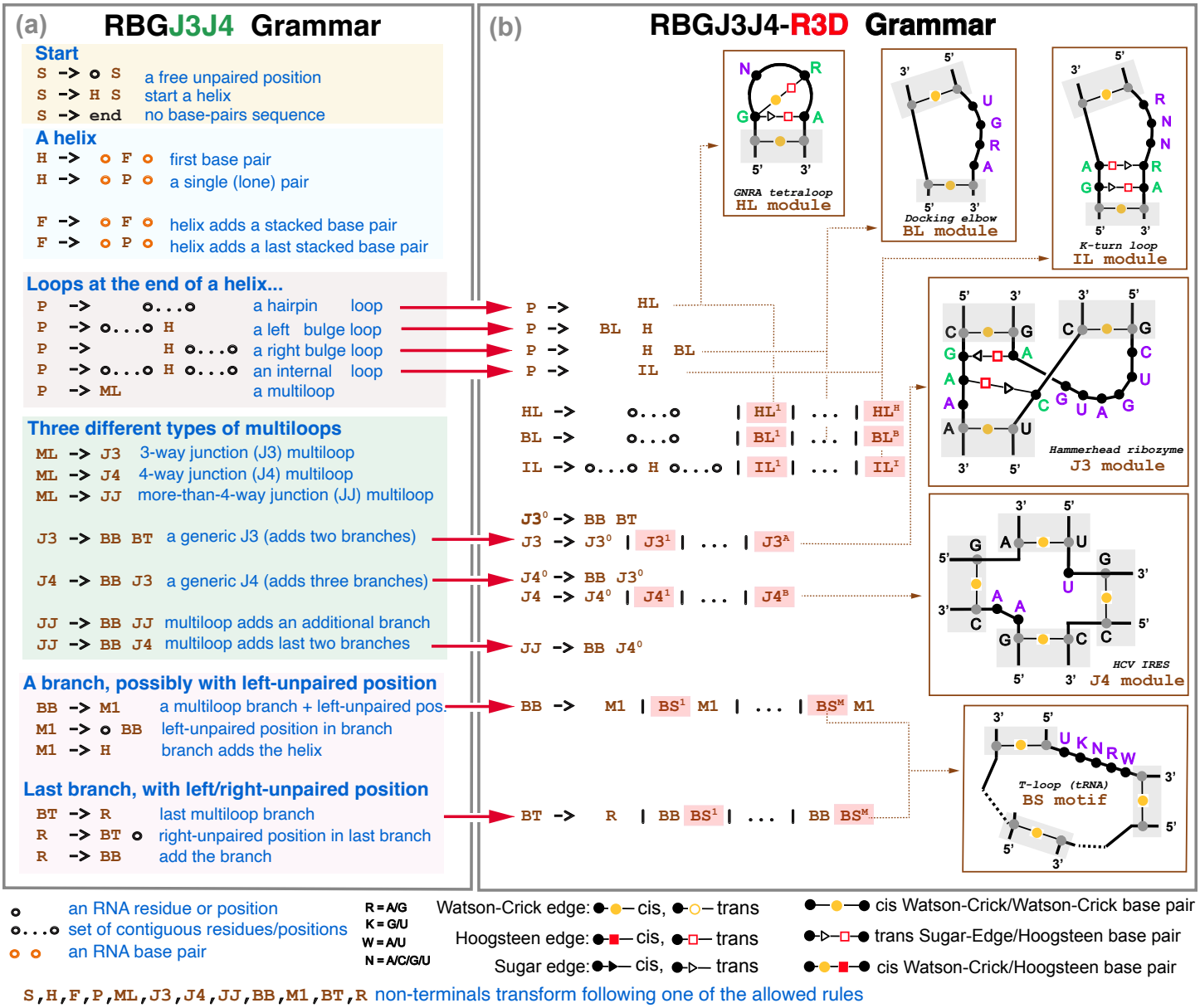


Figure 2: **The RBGJ3J4-R3D grammar.** The extended RBGJ3J4-R3D grammar with the modified elements describing specific RNA 3D motifs highlighted in red. Individual examples of RNA structural motifs and their corresponding appearance in the model are given as inserts. We depict in gray Watson-Crick base paired residues in the canonical helices setting the bounds, but not being part of the RNA motif; in green positions of the motif involved in non-Watson-Crick base pairing, in purple motif positions not paired. The RBGJ3J4-R3D grammar can be used to describe alignments not just individual sequences. The BB/M1 non-terminals (as well as the BT/R) are redundant for the RBGJ3J4 grammar in **(a)** (see supplemental Figure S1 for a simplied description), but they become different entities in the RBGJ3J4-R3D grammar in **(b)**. The RBGJ3J4-R3D grammar is unambiguous, that is, a given alignment with a particular arrangement of base pairs and 3D motifs can only be generated one way by the grammar.

Parameterization of RBGJ3J4-R3D probabilistic parameters. The RBGJ3J4-R3D grammar modifies the nine rules in RBGJ3J4 that are responsible for the incorporation of loops into the structure (depicted by red arrows in Figure 2b). These rules describing generic loops are expanded into also describing specific 3D motifs in that class. The additional transition probability parameters associated to those new 3D motif-specific rules could in principle be calculated by maximum likelihood as the rest of the parameters before in RBGJ3J4. Unfortunately, there is not yet a reliable and large enough database of RNA sequences/structures that extensively annotates 3D motifs along with the secondary structure.

While such database of RNA secondary structures and 3D motifs is produced, we have proceeded as fol-

lows: the RBGJ3J4 probability of a given loop type is split between the generic loop state and the whole R3D motif class, which gets assigned a fraction of it. Those fractions set for now by human curation are $(0.4, 0.4, 0.5, 0.2, 0.2, 0, 2)$ for the HL, BL, IL, J3, J4, and BS motif types respectively. Moreover and to avoid any possible overfitting, we use the principle of maximum entropy and assign the same probability to each of the specific 3D motifs of a particular class given in the descriptor. For instance, the probability of forming a generic hairpin loop in the trained RBGJ3J4 is 0.3475, thus RBGJ3J4-R3D assigns 0.2085 ($= 0.3475 \times 0.6$) to the generic hairpin loop, and the remaining 0.1390 ($= 0.3475 \times 0.4$) is equally distributed over all defined hairpin loop motifs (15 in the current implementation). Thus each 3D HL motif gets assigned a probability of 0.0093 ($= 0.1390/15$).

Next we describe the specific R3D models for all six different loop classes.

R3D: Six architectures to describe 3D motifs in all types of RNA loops

Now we introduce the R3D grammars, which incorporate an arbitrary number of 3D motifs in any arbitrary loop region into the folding grammar. Integrating the R3D grammars with the RBGJ3J4 grammar gives one SCFG jointly modeling both secondary structure and motifs (Figure 2).

The key insight behind the R3D grammars is to realize that RNA 3D motifs have a structural component determined by the set of (mostly conserved) non-Watson-Crick pairs that characterize the motif, and also a sequence-based component as many 3D motifs also conserve particular residue identities. The R3D grammars describe the structural component of a motif using profile SCFGs specific for each type of motif (Figure 3a-3f), and the sequence component with customized profile hidden Markov models that allow for sequence variability (Figure 3g).

The key that makes the R3D grammar affordable is that unlike other methods like RMDetect [14] or BayesPairing2 [79], R3D does not attempt to model each of the actual non-Watson-Crick base pairs individually (which can be quite complicated and non nested). R3D instead models groups of residues that are correlated because of their underlying non-Watson-Crick base pairing. This induces a segmentation of a motif into continuous subsequences (modeled by profile HMMs) involved in specific correlations (modeled by the SCFGs). This decomposition allows one model to describe all motifs of one given type, giving rise to a generalized R3D grammar per motif type (Figure 3). The SCFG states can generate multi-residue long strings using specific profile HMMs. For each motif, the individual nucleotide bases that constitute each segment of the profile motif are of course dependent on the consensus sequences of the motifs.

We consider six different types of structural motifs, based on whether they occur in hairpin (HL), bulge (BL), or Internal loops (IL), as well as in 3-way (J3), 4-way (J4), or Branch Segments (BS) that can occur in any junction. Each of the six general R3D SCFG models in Figure 3a-3f have a particular SCFG architecture describing the interactions present in each motif.

We decompose a motif into helical and non helical interactions. Helical interactions are modeled by classic Watson-Crick pairing. The non-helical interactions constitute the actual motif and there are broken down into contiguous subsequences (sequence segments) that belong to loops; *e.g.* just one segment in one loop capping a helix corresponds to a hairpin loop motif, while three separate loops connecting three helices correspond to a 3-way junction. We now detail the segmentation method per type of motif, along with the corresponding grammar rules.

Hairpin Loop motifs. 3D motifs in hairpin loops (`HL`) motifs include both residues paired through non-Watson-Crick interactions as well as unpaired ones. For instance, the GNRA tetraloop [33] is a frequent hairpin loop motif in which the first G base forms two non-Watson Crick interactions with the R and A bases, which provides extra thermodynamic stability to the tetraloop [33]. The GNRA R3D SCFG models the correlated occurrence of the G and the NA base pairs (but it does not model the type of base pairing involved in that correlation), as well as the unpaired N residue (Figure 3a).

R3D designs a generic HL 3D motif by an arbitrary number of left/right correlated segments and a final loop segment of residues not correlated elsewhere. Figure 3a shows the general model. The R3D-HL motif assigns a profile HMM to the loop sequence, as well as to all the allocated left and right segments which will consist of the contiguous subsequences that pair through non-Watson-Crick interactions.

Bulge Loop motifs. R3D bulge loop (`BL`) motifs are described in Figure 3b, and they have similar properties to the `HL` motifs. Notice that a `BL` motif can appear in a left of right bulge depending on which of the two ends of the motif is continuous and which inserts itself with the rest of the structure. Figure 3b shows only one of the two possibilities (called variants) for the `BL` motif. We generalize the concept of motif variants in the following sections and supplemental Figure S2.

Internal Loop motifs. For an internal loop (IL) motif (Figure 3c), R3D assumes the presence of 2 loop regions with an inner stem and an outer stem region which are emitted correlatedly by the SCFG. As with the HL motifs, the actual sequences in the loops and left/right inner and outer stem sequences (all of which can be potentially empty) are modeled by profile HMMs.

For instance, the K-turn (or Kink turn) is a common internal loop motif featuring two G-A hydrogen bonded Sugar-Hoogsteen edge interactions that help induce an axial bend [39]. The K-turn R3D SCFG models these two correlated interactions. The internal loop portion of the K-turn has three unpaired nucleotides with consensus RNN, so the R3D grammar adds a profile HMM for the right bulge RNN sequence and treats the left bulge as empty (Figure 3c).

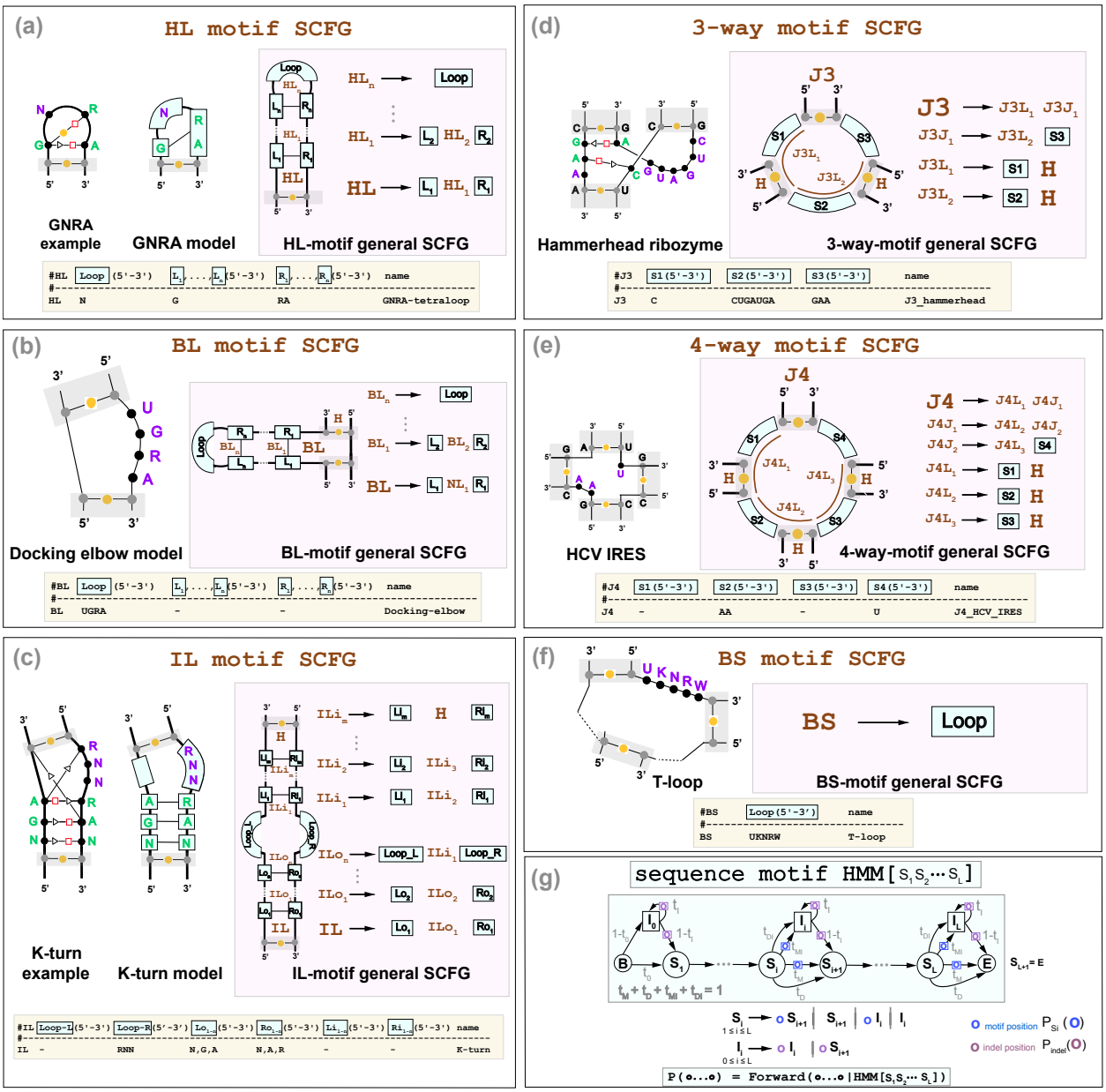


Figure 3: **The R3D grammars describes six types of RNA 3D motifs.** The motif types are: **(a)** Hairpin Loop (HL), **(b)** Bulge Loop (BL), **(c)** Internal Loop (IL), **(d)** 3-way Junction (J3), **(e)** 4-way Junction (J4), **(f)** and Branch Segment (BS). For each motif, the bold terminal: **HL**, **BL**, **IL**, **J3**, **J4**, and **BS**, is the one that inserts into the RBGJ3J4-R3D grammar in Figure 2. Closing Watson-Crick base pairs are not included in the motif models. **(g)** In blue, we describe the sequence string modeled with profile HMMs. We also show in yellow the descriptor used to build any particular motif.

3-way and 4-way junction motifs. The R3D SCFG for a 3-way junction (J3) motif includes three sequence

segments that are emitted correlatedly (Figure 3d). For instance for the Hammerhead ribozyme, while R3D does not model the non canonical base pairs that occur within the junction, it does model the correlated emission of all three segments which include the base paired residues as well as those that are not paired, but part of the motif. Similarly, for a 4-way junction (`J4`) motif four arbitrary sequence segments are considered simultaneously (Figure 3e). As seen in the case of the HCV IRES 4-way junction, the correlated segments may include no nucleotides, thus indicating helix coaxial stacking [93].

Other multiloop motifs. The R3D grammar also introduces sequence (`BS`) motifs which can appear in any multiloop branch as described in Figure 3f. These motifs may describe particular protein-binding motifs such as the CsrA binding motifs of the CsrB RNA [50], as well as components of higher-order loop motifs. For instance, the Loop E that appears in a 3-way junction of the Glutamine riboswitch [70], which is interrupted by a one base pair pseudoknot, and R3D is able to model with two BS motifs.

The sequence-motif profile HMMs. Each interacting partner or loop in a RNA 3D motif consists normally of a conserved sequence with some variability. R3D models those sequence segments as short profile hidden Markov models (HMMs) described in Figure 3g. Each profile HMM has a consensus sequence, and by allowing mutations, insertions and deletions, it is able to accommodate sequence variability and to identify motif instances that have some variability relative to the consensus. The states of the profile HMM emit on transition, not on state. Motifs with sequence segments without residues, such as those occurring in multiloops bounded by coaxially stacked helices, are also possible. We model empty segments with a profile HMM to allow for the possibility of insertions relative to consensus.

Parameterization of R3D profile HMMs. The 3D motifs behave mostly independently of the RNA structure in which they appear. This adaptability has facilitated the identification of 3D motifs repeatedly found in the context of very different RNA structure [2], and now it facilitates the parameterization of the models describing these motifs in a way that naturally generalizes to any RNA structures in where they appear.

In this work, each individual motif is assigned a R3D profile SCFG that belongs to one of the six classes in Figure 3a-f. Each motif includes a particular set of profile HMMs describing the actual sequence motifs defining the motif. For instance, any HL motif has one `Loop`, and at least two correlated `L-R` profiles (Figure 3a). Each profile HMM has a consensus sequence. As an example, we design the GNRA tetraloop with a one-nucleotide profile HMM `Loop` = "N", an two correlated profiles `L`= "G" and `R` = "RA", as "G" interacts both with "R" and "A" (see supplemental Figure S3). Profile HMMs can have an empty consensus. For instance, the U-turn three-loop motif has a `Loop` = "URA" profile, but both the `L` and `R` profiles are empty.

A profile HMM assigns one Match state per consensus position (Figure 3g). Each Match assigns the same probability to all residues in the consensus, and mismatches to residues not in the consensus are allowed with a small probability of 10^{-4}. For instance, for the U-turn motif profile HMM `Loop` = "URA", the emission probabilities for the first position "U" are $p(U) = 0.9997$ and $p(A) = p(C) = p(G) = 0.0001$. The emission probabilities for the second position "R" are $p(A) = p(G) = 0.4999$ and $p(C) = p(U) = 0.0001$. With this parameterization, CaCoFold-R3D is able to identify for instance a U-turn motif "YAA" in the Metazoan SRP (see Figure 5).

For each consensus Match state, there is also one Insertion and Deletion states that allow to extend or shorten the size of a given motif in order to model the variability observed for the actual motifs found in the 3D structures [5]. In this implementation, the transition probabilities between states are parameterized such that the profile HMM generates sequences that on average exceed slightly the length of the consensus motif (adding a 0.1 per position) up to a max of 1.5 extra length per motif on average. With this parameterization, while we have designed the Loop-E IL motif to have 5 consensus positions on each branch, CaCoFold-R3D is able to identify the Loop-E in 5S rRNA which 5 and 6 residues per branch respectively (see Figure 5).

Another source of variability occurs because alignments can introduce large insertions in loop regions that extend beyond the actual 3D motif. By default, R-scape does not analyze positions in the alignment that have more than 75% gaps (by default). This effectively removes highly variable regions from analysis which helps identifying the actual 3D motifs which generally show a high level of conservation.

Motif variants. All RNA 3D motifs except `HL` motifs are bound by more than one helix, thus allowing different topological variants depending on which $5'/3'$ ends are selected to integrate the motif into the rest of the structure. Bulge and Internal loop motifs have two variants, and 3-way and 4-way junctions have three and four variants respectively. For instance, the two variants of any BL motif correspond to a left and right bulge motif respectively. Supplemental Figure S2 describes all motif variants with their SCFG rules. For any 3D motif entry in the R3D descriptor file, CaCoFold-R3D internally models all possible variants of the motif.

R3D-prototype: The importance of framing 3D motifs by evolutionary information

One of the keys to our approach is that the CaCoFold-R3D method bounds the search of RNA 3D motifs to the segments of the RNA molecule enclosed by helical regions with covariation support. This is important as, due to the small size of the motifs, their associated models have low information content and would otherwise produce large number of false positives.

To initially test the effect of adding covariation information into the prediction of RNA 3D motifs, we implemented a R3D-prototype that simultaneously produce a secondary structure and models two 3D motifs: the GNRA tetraloop (a hairpin motif) and the K-turn (an internal loop motif). This prototype uses a version of the RBG grammar (Figure S1a) that produces structural predictions directly on RNA sequences, and implements two R3D grammars (also on sequences) modeling GNRA loops and K-turns. The prototype uses this RBG-R3D grammar, and for each RNA sequence predicts a maximum probability secondary structure including GNRA loops and K-turn motifs. The R3D-prototype is independent of R-scape, and does not assess significant covariation. But as input to the prototype one can add, as an option, external information of covarying base pairs.

For each Rfam family, sequences are selected at random from their seed alignments, and covarying base pairs are extracted from the Rfam seed alignments. We record both the sensitivity, defined to be the percent of truth motifs successfully detected, as well as the average number of false positives per prediction. We perform this analysis both including and excluding covariation information to demonstrate the effectiveness of the model.

In Table 1, we present results from applying the R3D-prototype to structural RNAs from different Rfam families [61]. As positives, we tested the U3 small nuclear RNA and the spliceosomal U4 RNA which include two and one K-turns respectively [94], and the 5S rRNA which contains a GNRA tetraloop [64]. The U3 and U4 RNAs also serve as negative tests for the GNRA tetraloop, and 5S rRNA as negative for the K-turn. For an independent control, we selected the 6S RNA and the Ribosome modulation factor (RMF) RNA, which lack either of the tested motifs.

As hypothesized, adding covariation information vastly improves motif prediction accuracy despite the lack of covariation within the motifs themselves. Overall sensitivity on the detection of GNRA tetraloops and K-turns in the three positive RNAs increases after adding covariation from 84% to 95% (Table 1). Adding covariation also significantly reduces false positives for K-turn detection to similar levels to that of the GNRA tetraloop.

RNA family	# seqs	avg length	# bpairs covary/total	**Constrained by Covariation** Sensitivity (%)	False positives/seq GNRA	False positives/seq K-turn	**No covariation used** Sensitivity (%)	False positives/seq GNRA	False positives/seq K-turn
GNRA motif									
5S rRNA (RF00001)	50	117	32/34	100	0.14	0	80.65	0.29	0.24
K-turn motif									
U3 snoRNA (RF00012)	50	209	13/63	90.3	0.12	0.06	84.95	0.14	0.15
U4 snRNA (RF00015)	50	144	24/28	96.0	0.30	0.02	88.00	0.30	0.10
Controls									
6S RNA (RF00013)	50	180	35/52	–	0.02	0.06	–	0.04	0.12
RMF RNA (RF01755)	50	130	1/28	–	0	0	–	0	0
Random Shuffle	20×5	(from all 5 families)		–	0.19	0.12	–	0.07	0.44
Weighted Total				**95.4**	**0.15**	**0.06**	**84.5**	**0.14**	**0.24**
New K-turns									
drum bacterial RNA (RF02958)	100	113	22/29	93.0	–	0	89.5	–	0.09
Actinomyces-1 RNA (RF02928)	100	110	11/34	100	–	0.01	98.9	–	0.06
RAGATH-18 RNA (RF03064)	100	71	15/18	97.9	–	0	90.6	–	0.11
HOLDH RNA (RF02997)	10	401	6/95	100	–	0.64	80.0	–	0.60
Weighted Total				**97.8**	–	**0.03**	**92.6**	–	**0.10**

Table 1: **R3D-prototype accuracy on GNRA and K-turn motifs.** The tested RNAs include one of the two tested motifs: the GNRA tetraloop for 5S RNA sequences, and the K-turn internal loop for the U3 and U4 snRNA sequences, or none for the control 6S and RMF RNAs. Sensitivity measures the fraction of motifs identified correctly by the R3D-prototype, where correct identification requires exact matching of the ends of the motif. False positives measure the instances of the tested motif found at the wrong position or instances of the not tested motif. For each RNA, sequences are selected at random from the corresponding alignments. Weighted averages are calculated weighting by the fraction of sequences of a particular RNA type that were analyzed. Several RNAs in the "New K-turns" table also have GNRA loops, thus they cannot be used to estimate GNRA false positives. The location in the alignments of the motifs and covarying base pairs are given in the Supplemental Materials.

To further test the efficacy of our method, we applied it to four K-turns recently identified in bacterial RNAs

via structural prediction and X-ray crystallography [34]. The performance of our method on these alignments corroborates the high level of accuracy and low false positivity as demonstrated before (Table 1). This R3D-prototype shows that our approach is a reliable predictor of confirmed motif structure. We moved on to making a full implementation of the RBGJ3J4-R3D model, named **CaCoFold-R3D**, that incorporates a large collection of RNA 3D motifs found recurrently in RNA structures [46, 2, 32] and operates on alignments.

R3D SCFG profiles of over fifty recurrent RNA 3D motifs

For the presented version of CaCoFold-R3D, we wanted to test its capability to integrate a comprehensive collection of RNA 3D motifs. We searched the literature for methods that had created 3D motif models, from which we finally cataloged 51 different motif architectures with confirmed representation in multiple structured RNAs [2, 25, 32, 80, 18, 66, 16, 14, 65, 10, 27, 68, 6]. It should also be noted that these does not have to be considered a definite list. The selection of motifs in CaCoFold-R3D is extremely flexible, and it only requires to adjust or modify the descriptor file. A customized descriptor file can be provided using the option `--r3dfile <descriptor-file>`.

The R3D descriptor describing the 51 motifs is provided in Figure S3. For this particular R3D descriptor, the three variants for the J3 motif J3_typeB are redundant with each other, as well as the two variants of the IL motif J4/5-IL, and the total number of unique motif variants is 96. The R3D code, given a descriptor file, calculates automatically all possible variants and eliminates those that are redundant. The total number of total non-redundant RNA motifs used is reported as part of the standard output for further inspection.

Figure 4 includes a representation of 20 (out of 51) motifs included in this implementation. The full list of motifs can be found in Figure S3, which also provides the descriptive notation used in our input files to represent the motifs in our models (supplemental Figure 2). The method is customizable by simply changing the input file with the representations of new motifs to be considered.

Figure 4 also includes for each of the 20 motifs a positive example of a Rfam family documented to have the motif, accompanied by a detail of the CaCoFold-R3D full structural prediction correctly detecting the motif. It is worth noticing, that in the majority of cases, the Rfam 3D motif is bounded by helices that show some level of covariation, further supporting to our key design feature of informing motif detection with the evolutionary conservation of secondary structure helices that arrange into a 3D structure.

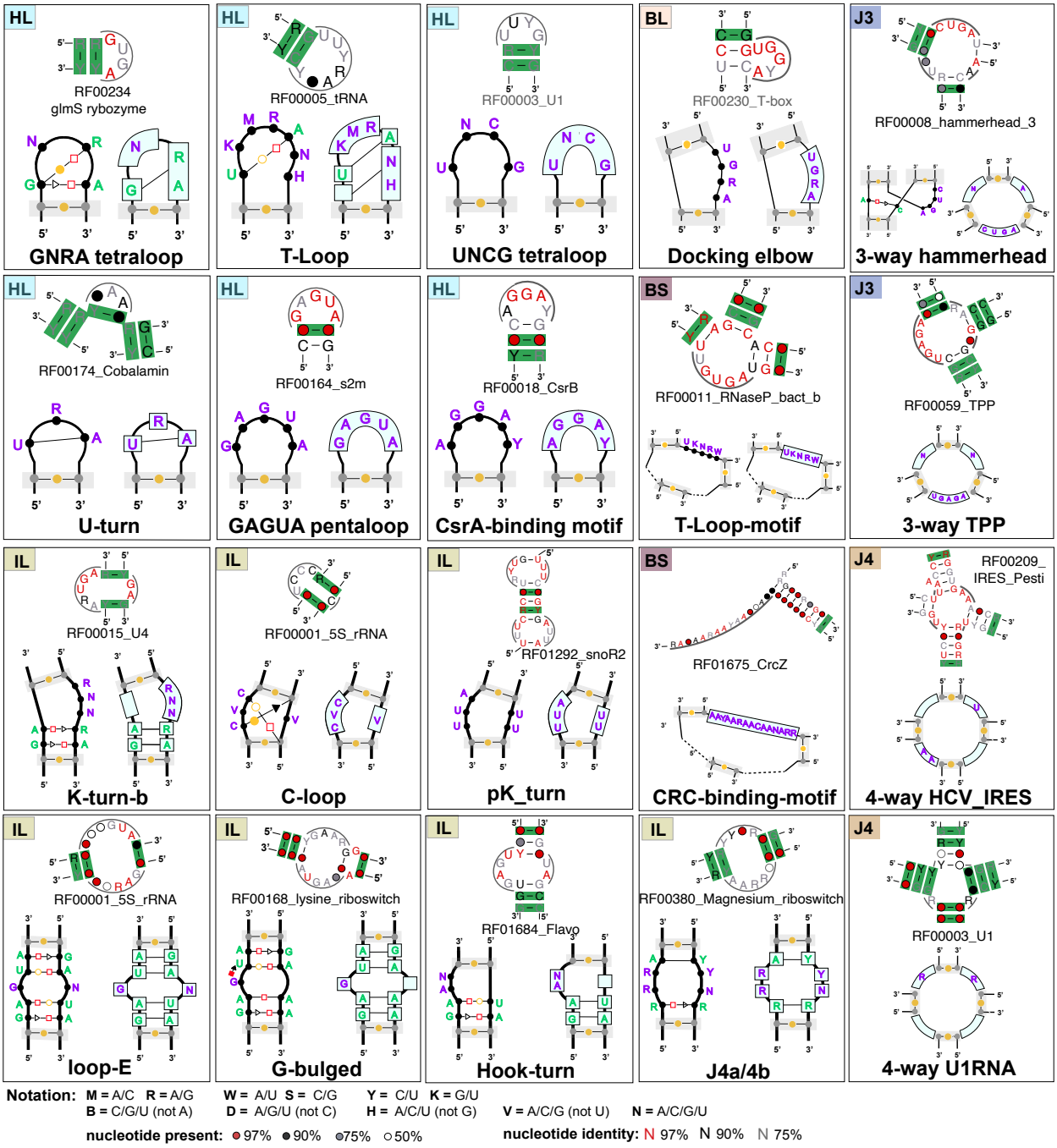


Figure 4: **Twenty RNA motifs with their R3D-grammar representation and detail from one Rfam family for which CaCoFold has identified a true instance of the motif with covariation support. Hairpin Loop motifs (HL)**: GNRA [33], T-loop [8], UNCG tetraloop [90, 9, 91]. U-turns [30], GAGAU pentaloop from conserved SARS region[77], and the CsrA binding motif [50]. **Bulge motifs (BL)**:Docking elbow [44]. **Internal loop motifs (IL)**: K-turn-b [39] (a small variant of the K-turn in Figure 3c), C-loop [48, 84], Loop E [23, 45], G-bulge [82], pK-turn [69, 56], Hook-turn [86], J4a/4b internal loop [15]. **3-way junctions (J3)**: the hammerhead ribozyme 3-way junction [54], and the 3-way junction of the TPP riboswitch [83]. **4-way junctions (J4)**: in the Hepatitis C virus internal ribosome entry site, or HCV IRES [63], and in the U1 spliceosomal RNA [42]. **Multiloop motif (BS)**: the T-loop domain [44] and the CRC binding domain [21]. Descriptors for the remaining 31 RNA 3D motifs included in the current version of CaCoFold-R3D are given in the supplemental Figure S3.

Results on RFAM alignments

We ran CaCoFold-R3D on all Rfam seed alignments. CaCoFold-R3D finds the K-turns in the alignments of the U3 snoRNA [53, 100], U4 snRNA [92], and the other four new K-turns [34] that we used in the R3D-prototype, as

well as the K-turn in the SAM riboswitch [57]. Figure 5 reports additional examples of full structure predictions with representative 3D motifs that have been reported in the literature.

CaCoFold-R3D also identifies the GNRA tetraloop in the 5S rRNA tested with the R3D-prototype (Table 1). In the 5S rRNA [12], we also observe the Loop E motif, the two G-bulge motifs in the T-box riboswitch [40], the J4a/4b 3D motif of the Magnesium riboswitch [15], and the T-loop motif in the TPP riboswitch as well as its characteristic 3-way junction [83]. Two interesting cases are the CsrB RNA that binds to the CsrA protein [50], for which we identify 12 binding motifs, and the Glutamine riboswitch, where two R3D branch segment (`BS`) motifs allow us to identify a confirmed Loop E motif occurring in a multiloop involving a pseudoknot instead of in an internal loop [70]. For the Metazoan SRP, CaCoFold-R3D identifies several of its characterized motifs (domain IV, C-loop, K-turn, U-turn, GNRAs) [81, 4, 29] (Figure 5). In the supplemental Table S1, we report the performance for all Rfam families. The collection of all Rfam predicted structures is provided in the supplemental material.

With regard to the distribution of detected 3D motifs, we observe that the GNRA tetraloop is the most frequently observed motif, followed by the K-turn. Most motifs of any other kind have between 10-50 instances in Rfam (supplemental Table S1). Because the CaCoFold-R3D predictions integrate the covariation information observed in the alignment base pairs, we use the covariation observed in the helices bounding the 3D motifs in order to assess our confidence in the predictions. Overall, we detect a total of 2,124 motifs, of which 1,460 have *covariation support*, defined here as a motif for which at least one of its bounding helices has one or more covarying base pairs. 591 of the Rfam families include 3D motifs with covariation support. For the two largest RNA structures SSU and LSU rRNA, we find 45 supported 3D motifs for eukaryotic SSU, and 62 for the eukaryotic LSU rRNA (see Table S1). The list of motifs detected for each Rfam family is also provided in the supplemental materials.

As a control, we obtain predictions for negative alignments obtained from the Rfam alignment by permuting the residues in each column (position) independently from each other. As a result of the shuffling, the covariation signal in the input alignment is altered, but the base composition of the positions remain unchanged, thus retaining the sequence signature of any potential motif. For these control alignments, we obtain 121 motifs supported by covariation, which compared to the $1,460$ motifs obtained for the Rfam alignments, indicate an estimated 8.3% false discovery rate (FDR) in our predictions supported by covariation. For motifs not supported by covariation support, the comparison between found in Rfam ($2,124 - 1,460 = 664$) versus those found in controls ($290 - 121 = 169$) report an FDR increases of 25.4%. Again, stressing the important effect of using covariation information in the detection of 3D motifs.

Notice that the control alignments report 733 helices out of $14,146$ with at least one covarying base pair. Since R-scape [74, 72] reports pair with a significance E-value cutoff of 0.05, this number (733) is in good agreement with the expected average number of helices with at least one covarying pair under the null hypothesis ($707.3 = 0.05 \times 14,146$).

The Rfam alignments are good-quality structural alignments built with the method Infernal [60] using diverse homologs of known structural RNAs. When building RNA alignments several issues need to be considered that could compromise the quality of the covariation signal. We have documented that pseudogenes and non-homologous sequences weaken the covariation signal; and the use of structural alignments built on unconfirmed predicted structures introduces a circular reasoning that can produce spurious covariation [26]. We have shown that structural alignments of bona-fide structural RNA sequences result in reliable 3D motif predictions informed by covariation found in alignments. Furthermore, even within a good structural alignment there is an important difference in confidence whether the motif prediction is bound by covariation or not.

Performance comparison to other methods

In Table 2, we present CaCoFold performance for an array of 3D motifs present in specific crystal structures.

Type	Motif name	PDB	Rfam homology	CaCoFold-R3D	Other Methods	Motif Model
HL	GNRA-tetraloop	7EI6	synthetic	170	–	Hendrix *et al.*, 2005 [32]
HL	U-turn	4WFL	RF00017_Metazoan_SRP	YES	–	Kempf *et al.*, 2014 [37]
HL	UNCG-tetraloop	1F7Y	synthetic	48	–	Hendrix *et al.*, 2005 [32]
HL	ANYA-tetraloop	6MSF	synthetic	18	–	Hendrix *et al.*, 2005 [32]
HL	CUYG-tetraloop	2L6I	synthetic	20	–	Hendrix *et al.*, 2005 [32]
HL	YGNN-tetraloop	1AFX	synthetic	13	–	Hendrix *et al.*, 2005 [32]
HL	GANC-tetraloop	3BWP	synthetic	15	–	Keating *et al.*, 2009 [38]
HL	UNAC-tetraloop	4A4R	synthetic	26	–	Zhao *et al.*, 2012 [101]
HL	T-loop-tetraloop	7SAM	RF01085_TLS-PK4	YES	–	RMfam [27]
HL	L8_RNaseP_bact_a	1U9S	RF00010_RNaseP_bact_a	YES	–	Das & Baker, 2010 [16]
HL	Sarcin-Ricin_loop	1SCL	RF02540_LSU_rRNA_archaea	YES	BayesPairing2	Szewczak *et al.*, 1993 [87]
HL	CsrA_binding	2MF0	RF00018_CsrB	YES	–	Duss *et al.*, 2014 [20]
HL	GAGUA_pentaloop_sars	1XJR	RF00164_Coronavirus s2m motif	YES	–	Das & Baker, 2010 [16]
IL	Kink-turn_SAM_riboswitch	2GIS	RF00162_SAM_riboswitch	YES	BayesPairing2	Das & Baker, 2010 [16]
IL	Kink-turn_rRNA	1JJ2	RF02540_LSU_rRNA_archaeal	YES	RMDetect	Das & Baker, 2010 [16]
IL	LoopE	354D	RF00001_5S_rRNA	YES	–	Das & Baker, 2010 [16]
IL	GAAA_Tetraloop-receptor	1GID	RF00028_Group I intron	YES	–	Cate *et al.*, 1996 [7]
IL	C-loop	1S72	RF02541_LSU_rRNA_bacterial	YES	RMDetect	Lescoute *et al.*, 2005 [48]
IL	G-Bulge	3DIL	RF00168_Lysine riboswitch	YES	RMDetect	Serganov *et al.*, 2008 [82]
IL	G-Bulge_Das	1Q9A	RF02541_Bacterial LSU_RNA	YES	–	Das & Baker, 2010 [16]
IL	tandem GA	1SA9	RF00177_SSU_rRNA_bacteria	YES	RMDetect	Cruz & Westhof, 2011 [14]
IL	Twist up	1J5E	RF02541_LSU_rRNA_bacterial	NO, 9	–	Zhong & Zhang, 2011 [102]
IL	UAA/GAN	6AHR	RF00009_RNaseP_nuc	YES	–	Lee *et al.*, 2006 [43]
IL	J4a/4b	2QBZ	RF00380_M-box_riboswitch	YES	–	Das & Baker, 2010 [16]
IL	J4/5	2R8S	RF00028_Group I intron	NO, 6	–	Das & Baker, 2010 [16]
IL	GUA/GG_RRE	1CSL	RF00055_HIV-1 RRE	too conserved, 10	–	Das & Baker, 2010 [16]
IL	SRP RNA DOMAIN IV	1CQ5	RF00017_Metazoan_SRP	YES	–	Schmitz *et al.*, 1999 [81]
IL	Hook_turn_GAAC/CUA	1MHK	synthetic	32	–	Das & Baker, 2010 [16]
IL	Docking elbow-IL	3Q1Q	RF00011_RNaseP_bact_b	YES	–	Lehmann *et al.*, 2012 [44]
IL	pK-turn	3OK7	RF00010_RNase P RNA	NO, 45	–	Meyer *et al.*, 2012 [56]
J3	Purine_riboswitch_J3	2EEW	RF00167_Purine riboswitch	YES	–	Das & Baker, 2010 [16]
J3	hammerhead_CUGA/A/C	359D	RF02275_Hammerhead_HH9	YES	–	Das & Baker, 2010 [16]
J3	J3_typeA	4RZD	RF02680_PreQ1-III	YES	–	Lescoute & Westhoff, 2006 [49]
J3	J3_typeB	1J5E	RF00177_SSU_rRNA_bacteria	YES	–	Lescoute & Westhoff, 2006 [49]
J3	J3_typeC	4LX6	RF03165_2dG-II riboswitch	YES	–	Lescoute & Westhoff, 2006 [49]
J3	J3_TPP	2GDI	RF00059_TPP riboswitch	YES	–	Serganov *et al.*, 2006[83]
J3	J3_groupII	5J01	RF02001_group-II-D1D4-3	YES	–	this work
J4	HCV_IRES_AA/0/U/0	1KH6	RF00061_IRES_HCV	YES	–	Das & Baker, 2010 [16]
J4	J4_U1RNA	3PGW	RF00003_U1 RNA	YES	–	Weber *et al.*, 2010 [95]
BS	T-loop	1EHZ	RF00005_tRNA	YES	–	Hendrix *et al.*, 2005 [32]
BS	LoopE-a	4FRN	RF01689_AdoCbl-II	YES	–	this work
BS	LoopE-b	4FRN	RF01689_AdoCbl-II	YES	–	this work
BS	CRC_binding	-	RF00195_RsmY	YES	–	Sonnleitner *et al.*, 2009 [85]
PK		5DDP	RF01739_Glutamine riboswitch	YES	–	Ren *et al.*, 2015 [70]

Table 2: **Non-canonical motif *de novo* prediction from alignments.** For each motif, we identify one crystal structure that includes the motif, and we test for the prediction of the motif on the corresponding Rfam alignment. We also report other methods that have identified the motif in the same family. For motifs associated to a crystallized synthetic oligo, or not identified by CaCoFold-R3D, we provide the number of other Rfam families that include that motif with covariation support. The specifications of the motifs are provided in the R3D descriptor file given in supplemental Figure S3. CaCoFold-R3D results for the Rfam seed aligments are given in the supplemental materials file `results/CaCoFold-R3D/CaCoFold-R3D_Rfam`.

With regard to comparison to the performance of related methods, RMDetect assesses 3D motif prediction in alignments for four IL motifs: the K-turn, C-loop, G-bulge and tandem GA motifs. The motifs are tested on SSU rRNA archaea and bacterial alignments (both of which include cases of all four motifs), SAM riboswitch and U4 RNA alignments (which include the K-turn motif), and a Lysine riboswitch alignment (which includes the G-bulge motif). All those motifs are detected by CaCoFold-R3D in the corresponding Rfam alignments. The method BayesPairing2 assesses prediction sensitivity for two motifs (the Sarcin-Ricin and K-turn loops) tested on three families: LSU archaeal (RF02540) and LSU bacterial (RF02541) and the SAM riboswitch (RF00162). BayesPairing2 identifies the Sarcin-Ricin motif in the LSU archaeal and LSU bacterial, and a K-tun in all three families. We identify the Sarcin-Ricin motif in all four LSU families in Rfam (including LSU eukarya and LSU microsporidia), as well as in three other families. As for the K-turn motif, CaCoFold-R3D identifies it in the SAM riboswitch as well as in all four LSU families in Rfam. Notice that these results are obtained with one family-independent parameterization of CaCoFold-R3D that generalizes to all families. This is unlike BayesPairing2, for which results vary a lot depending on which family has been used to train the model.

CaCoFold-R3D not only detects all motifs mentioned in the literature of comparable methods, but also detects many new motifs, especially internal loop and multiloop junction motifs, no other method has reported detecting.

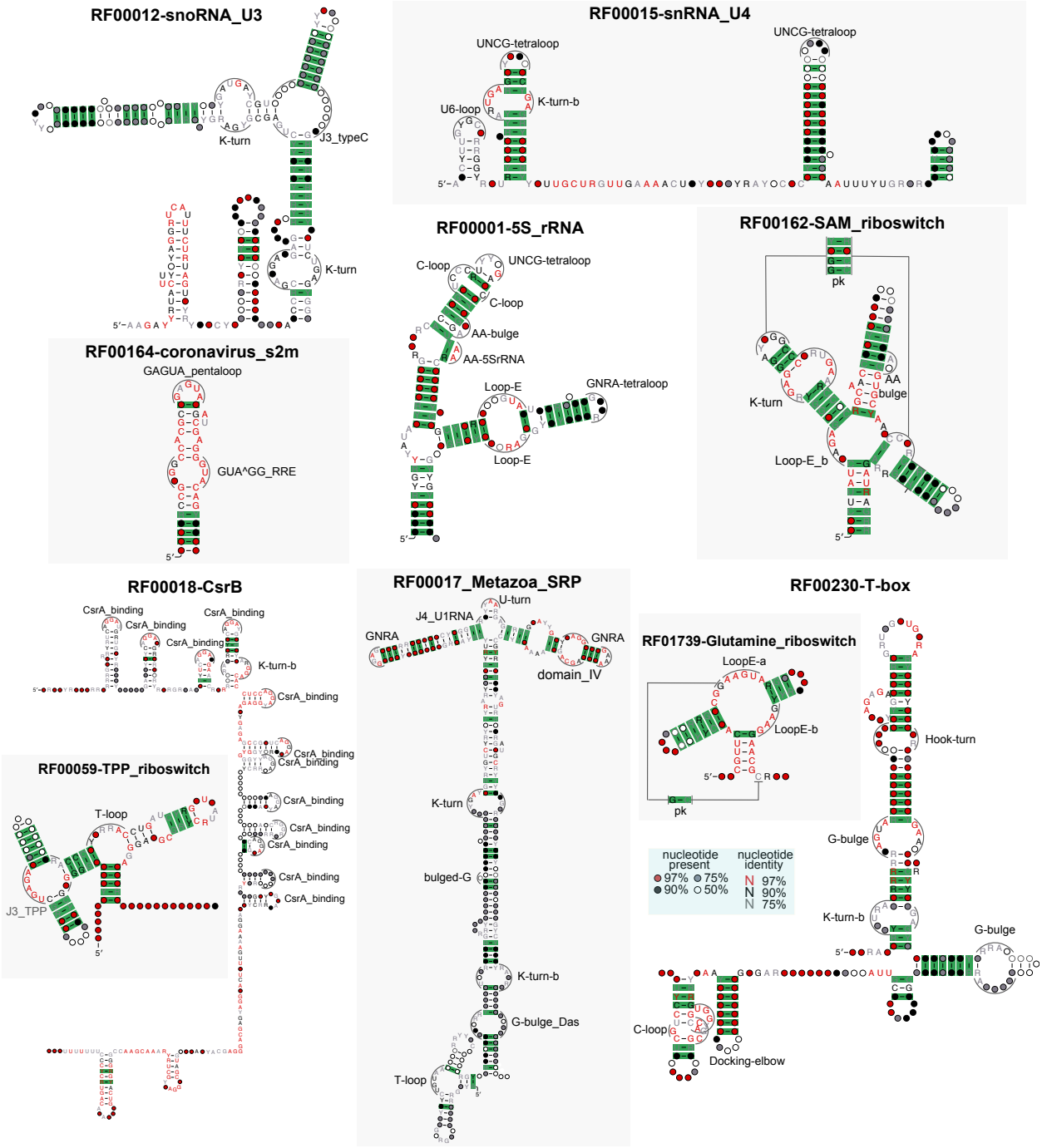


Figure 5: **CaCoFold-R3D structures confirmed by known 3D structures.** The examples of predicted consensus structures with 3D motifs are: the snoRNA U3 [53], snRNA U4 [92], the coronavirus stem-loop II motif (s2m) [77], the 5S rRNA [12], the SAM riboswitch [57], the CsrB RNA [50], the Metazoan SRP RNA with the domain IV motif [81], K-turn, U-turn [4] and T-loop [29], the Glutamine riboswitch [70], the T-box riboswitch [40], and the TPP riboswitch [83]. CaCoFold-R3D uses a customized version of R2R [96] that automatically draws the RNA 3D motifs in the context of the rest of the consensus structure and its covariation. The collection of predictions for all the Rfam RNA families is provided in the supplemental materials.

A new 3-way junction motif with high representation

As an example of the power of CaCoFold-R3D as a tool to discover new motifs, we turn to a loop in the Group II intron RNA for which Rfam describes a generic left bulge (Figure 6a). From the CaCoFold analysis of the Rfam seed alignment, we inferred that this is actually a 3-way junction that is very conserved in sequence and exquisitely framed by covariation in all three closing helices (Figure 6b). Two of the closing helices are adjacent,

and the third one is just a lone base pair. A Group II crystal structure [13] confirms the coaxial stacking of the two adjacent helices, as well as the lone pair; it also reports one non-Watson-Crick base pair within the 3-way junction (Figure 6c).

Because CaCoFold-R3D works on alignments, in the presence of a new potential motif, as in this case, the alignment already provides the potential consensus for the motif. Thus, the construction of R3D descriptor is automatic. In this case, the consensus sequence profiles of the three branches of the 3-way junction are $(5' - 3')$ "A", "RAA" and empty.

We created the R3D grammar for this novel J3 motif (Figure 6d), and introduced it into the model. We were surprised to find that this seems to be a recurrent motif also found in other structural RNAs. In fact, our analysis shows that it is the most frequent 3-way junction observed in Rfam as well as one of the top five most frequent motifs (Table S1). In Figure 6e, we show examples of other J3-groupII instances found in the CaCoFold-R3D structures for other Rfam families.

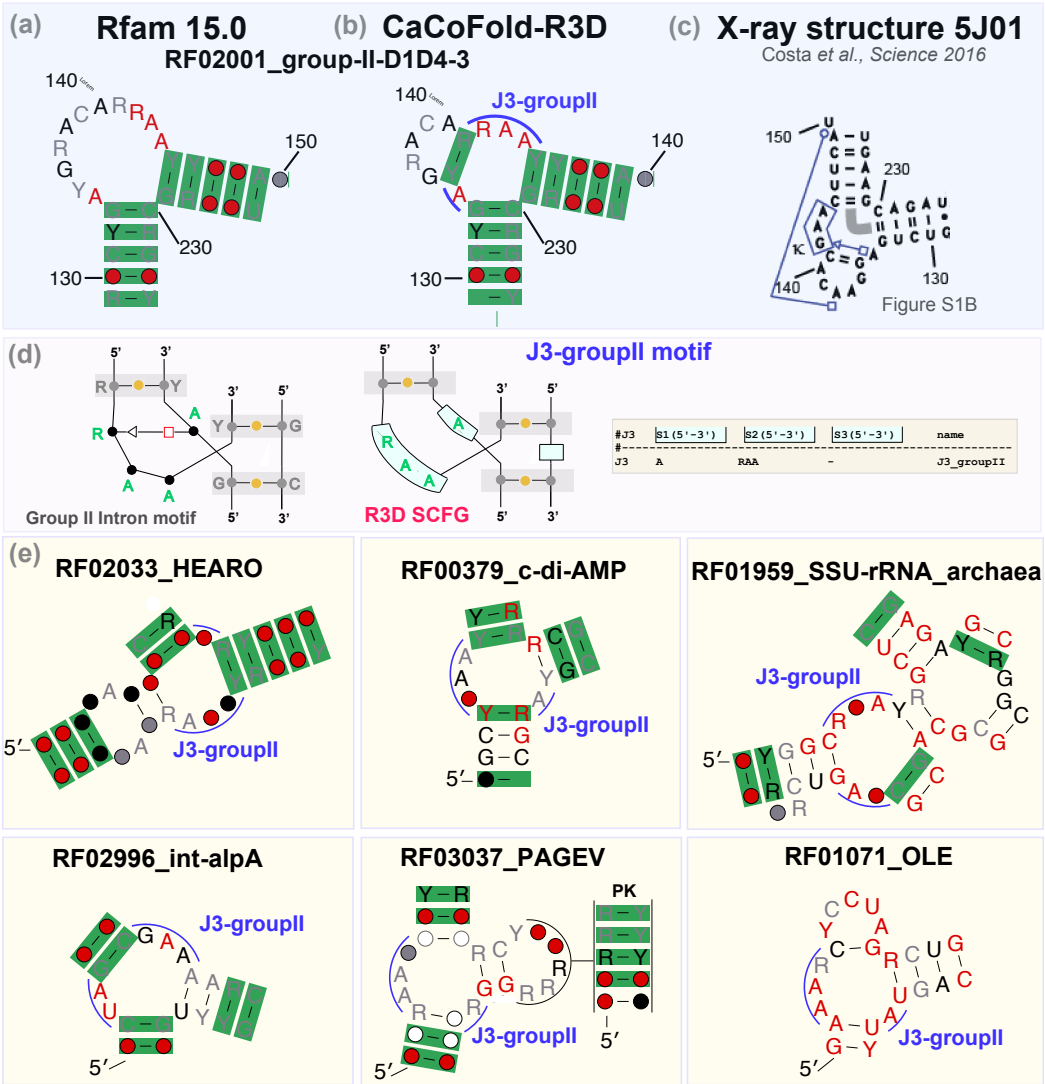


Figure 6: **Group II intron 3-way junction motif**. **(a)** Detail of a bulge described by the Rfam structure for the Group II intron RNA. (Coordinates are for reference to the crystal structure in **c**). **(b)** 3-way junction identified by CaCoFold with covariation support in all three helices. Two of the helices are adjacent, and one is a lone base pair. **(c)** Detail of a Group II intron crystal structure which confirms the presence of the 3-way junction, a coaxial stacking between two of the helices and the lone Watson-Crick base pair, as well as a non Watson-Crick base pair in the 3-way junction. **(d)** Description of the J3-groupII motif and its R3D model.**(e)** Examples of CaCoFold-R3D predictions for other Rfam families that also include the J3-groupII junction.

Time performance

CaCoFold-R3D is fast. On an Apple M3 Max (128 GB), 98% of the Rfam families (4079/4178) take less than 60 seconds to run CaCoFold-R3D end-to-end, and 95% of families take less than 30 secs. For the small and large subunits (SSU and LSU) of the rRNA–the two longest structured RNAs–it takes 32 minutes to analyze the eukaryotic SSU alignment (length 1,978 and 90 sequences) and 2.9 hours for the eukaryotic LSU rRNA alignment (length 3,680 and 88 sequences).

Moreover, while other methods have to run a different search for each motif and for each sequence and also calculate a secondary structure separately [14, 79], CaCoFold-R3D directly runs all 96 motifs together with the secondary structure in a single shot prediction, and reports a consensus structure including 3D motifs for the alignment. The all-at-once RBGJ3J4-R3D prediction CYK algorithm scales with $\mathcal{O}(L^3 \times M)$ for an alignment (or sequence) of length L, where M is the total number of nonterminals including both those for the RBGJ3J4 grammar (12) and those for the R3D grammars (96 in the tested implementation). Although due to the covariation constraints, we expect this to be a worse-case behavior.

Discussion

CaCoFold-R3D combines together several unique features that make the prediction of RNA 3D motifs accurate, fully integrated with secondary structure, and annotated with their expected reliability. The R3D grammar abstracts the different 3D motifs into six generalized designs, unlocking the ability to incorporate an arbitrary number and variety of motifs–we provide results using a total of 96 motifs (*everything*). The RBGJ3J4 grammar specifies all possible loops in an RNA molecule, allowing motif detection in any possible location within a sequence (*everywhere*). CaCoFold-R3D is fully probabilistic, so one can compute the joint probability of all structural motifs together with all nested helices, pseudoknots and triplets (*all at once*).

Because CaCoFold-R3D is flexible and easily customizable, we introduce a core of most representative 3D motifs including a new 3-way junction. Because it is computationally fast, we are able to present full predictions for all Rfam families, including the ribosomal RNA. Finally, Because CaCoFold-R3D is framed by the evolutionary information contained in the alignment, it provides information on predictive confidence as a function of the number of significant covarying base pairs extracted from the input alignment. These results demonstrate that the R3D grammar coupled with covariation information offers an accurate and reliable prediction paradigm for identifying crucial 3D motifs in structural RNA sequences, and an optimal tool to discover novel 3D motifs. CaCoFold-R3D will facilitate the construction of a database of structural elements present in RNA structures. This could a relevance similar to that of databases of protein domains [22].

Moreover, CaCoFold-R3D is a tool to provide valuable data to deep learning methods of RNA 3D structure prediction. Secondary structure is a crucial source of information to inform the prediction of RNA 3D structure in methods such as MC-Sym [62]. However, most methods ignore the rich information present in loops. Our method, able to predict jointly secondary and 3D motifs in one single predictive setting, opens the door for easily extracting this information, and using it with methods trained on RNA sequences and alignments for a more robust RNA 3D structure prediction. In particular, CaCoFold-R3D predictions for the Rfam RNA families will be used to provide more complete inputs for the training of deep learning methods for RNA 3D structure prediction [88]. Methods that predict RNA 3D structure such as AlphaFold3 [1] and RoseTTaFold All-Atom (RFAA) [41], which already use the Rfam data to inform their inputs, will benefit from the comprehensive information contained on the prevalent 3D recurrent motifs present in all RNA 3D structures provided by CaCoFold-R3D.

CaCoFold-R3D has the potential to have an impact in small molecule drug therapeutics that target RNA binding sites. RNA therapeutic compounds such as risdiplam (for spinal muscular atrophy) [67] rely of the presence of RNA loops forming favorably binding pockets [35, 17]. Our method is proven to hold important predictive information about key structural elements within RNA loops, which could greatly help guide the design of RNAs whose loop configurations serve as therapeutic targets for small molecule drugs.

Software and database versions

We introduce CaCoFold-R3D in the software package R-scape v2.5.7, which is provided as part of the supplemental materials, and can also be downloaded from `rivaslab.org`. The R3D-prototype python code is provided as part of the supplemental materials. We used TORNADO v0.6.0 [76], and Rfam v15.0 [61].

Author Contributions

E.R. conceived the research. A.K. and E.R. designed the algorithms. A.K. implemented the python R3D-prototype. E.R. implemented the CaCoFold-R3D method. A.K. and E.R. wrote the manuscript.

Funding

This work was supported by NIH grant R01-GM144423. A. Karan was funded by a Harvard Undergraduate Research Fellowship and a Paul and Daisy Soros Fellowship.

Acknowledgments

We thank Eric Westhof and Vladimir Reinharz for insights on RNA 3D motifs. We thank Liana Merk for bringing to our attention the 3-way junction motif in the Group II intron RNA. We thank Sean Eddy for a critical reading of the manuscript.

References

[1] J. Abramson, J. Adler, and J. et al. Dunger. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024.

[2] R. T. Batey, R. P. Rambo, and J. A. Doudna. Tertiary motifs in RNA structure and folding. *Angew Chem Int Ed Engl*, 38:2326–2343, 1999.

[3] E. F. Baulin, D. R. Bohdan, D. Kowalski, M. Serwatka, J. Swierczynska, Z. Zyra, and J. M. Bujnicki. ARTEM: a method for RNA and DNA tertiary motif identification with backbone permutations, and its example application to kink-turn-like motifs. *bioRxiv*, doi.org/10.1101/2024.05.31.596898, 2024.

[4] M. M. Becker, K. Lapouge, B. Segnitz, K. Wild, and I. Sinning. Structures of human SRP72 complexes provide insights into SRP RNA remodeling and ribosome interaction. *NAR*, 45:470–481, 2016.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[6] D. R. Bohdan, V. V. Voronina, J. M. Bujnicki, and E. F. Baulin. A comprehensive survey of long-range tertiary interactions and motifs in non-coding RNA structures. *RNA*, 51, 2023.

[7] J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, C. E. Kundrot, T. R. Cech, and J. A. Doudna. Crystal structure of a Group I ribozyme domain: Principles of RNA packing. *Science*, 273:1678–1685, 1996.

[8] C. W. Chan, B. Chetnanim, and A. Mondragón. Structure and function of the T-loop structural motif in noncoding RNAs. *Wiley Interdiscip Rev RNA*, 4:507–522, 2013.

[9] C. Cheong, G. Varani, and I. Tinoco, Jr. Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature*, 346:680–682, 1990.

[10] G. Chojnowski, T. Walen, and J. M. Bujnicki. RNA Bricks-a database of RNA 3D motifs and their interactions. *NAR*, 42:D123–D131, 2014.

[11] J. Cocke and J. T. Schwartz. Programming languages and their compilers: Preliminary notes. *Technical report, Courant Institute of Mathematical Sciences, New York University*, 1970.

[12] C.C. Correll, B. Freeborn, P.B. Moore, and T.A. Steitz. Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell*, 91:705–712, 1997.

[13] M. Costa, H. Walbott, D. Monachello, E. Westhof, and F. Michel. Crystal structures of a group II intron lariat primed for reverse splicing. *Science*, 354:aaf9258, 2016.

[14] J. A. Cruz and E. Westhof. Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nat. Methods.*, 8:513–521, 2011.

[15] C. E. Dann, C. A. Wakeman, C. L. Sieling, Baker S. C., I. Irnov, and W. C. Winkler. Structure and mechanism of a metal-sensing regulatory RNA. *Cell*, 130:878–892, 2007.

[16] R. Das, J. Karanicolas, and D. Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Meth.*, 7, 2010.

[17] M. D. Disney. The druggable transcriptome project: From chemical probes to precision medicines. *ACS, Biochemistry*, doi: 10.1021/acs.biochem.5c00006, 2025.

[18] M. Djelloul and A. Denise. Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14:2489–2497, 2008.

[19] M. J. Dupont and F. Major. D-ORB: A web server to extract structural features of related but unaligned RNA sequences. *JMB*, 435:168181, 2023.

[20] O. Duss, E. Michel, M. Yulikov, M. Schubert, G. Jeschke, and F. H.-T. Allain. Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature*, 509:588–992, 2014.

[21] M. J. Filiatrault, P. V. Stodghill, J. Wilson, B. G. Butcher, H. Chen, C. R. Myers, and S. W. Cartinho. CrcZ and CrcX regulate carbon source utilization in Pseudomonas syringae pathovar tomato strain DC3000. *RNA Biol.*, 10:245–255, 2013.

[22] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. Pfam: The protein families database. *Nucl. Acids Res.*, 42:D222–D230, 2014.

[23] G. E. Fox and C. R. Woese. 5S RNA secondary structure. *Nature*, 256:505–507, 1975.

[24] V. Fritsch and E. Westhof. The architectural motifs of folded RNAs. In G. Mayer, editor, *The Chemical Biology of Nucleic Acids*, pages 141–174. John Wiley & Sons, Ltd, 2010.

[25] H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *NAR*, 31:2926–2943, 2003.

[26] W. Gao, A. Yang, and E. Rivas. Thirteen dubious ways to detect conserved structural RNAs. *IUBMB Life*, 75:471–492, 2022.

[27] P. P. Gardner and H. Eldai. Annotating RNA motifs in sequences and alignment. *NAR*, 43:691–698, 2015.

[28] R. F. Gesteland, T. R. Cech, and J. F. Atkins, editors. *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*. Cold Spring Harbor Laboratory Press, New York, 2006.

[29] J. T. Grotwinkel, K. Wild, B. Segnitz, and I. Sinning. SRP RNA remodeling by SRP68 explains its rolein protein translocation. *Science*, 344:101–104, 2014.

[30] R. R. Gutell, J. J. Cannone, D. Konings, and D. Gautheret. Predicting U-turns in ribosomal RNA with comparative sequence. *J. Mol. Biol.*, 300:791–803, 2000.

[31] J. P. Haseloff and W. L. Gerlach. Simple RNA enzymes with new and highly specific endoribonuclease activities. *Nature*, 334:585–591, 1988.

[32] D. K. Hendrix, S. E. Brenner, and S. R. Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.*, 38:221–243, 2005.

[33] H. A. Heus and A. Pardi. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science*, 253:191–194, 1991.

[34] L. Huang, X. Liao, M. Li, J. Wang, X. Peng, T. J. Wilson, and Lilley D. M. Structure and folding of four putative kink turns identified in structured RNA species in a test of structural prediction rules. *Nucleic Acids Res*, 49:5916–5924, 2021.

[35] Y. Ishigami, M. S. Wong, C. Martí-Gómez, A. Ayaz, M. Kooshkbaghi, S. M. Hanson, D. M. McCandlish, A. R. Krainer, and Kinney. J. B. Specificity, synergy, and mechanisms of splice-modifying drugs. *Nat. Comm.*, 15:Article number: 1880, 20254.

[36] F. M. Jucker, H. A. Heus, P. F. Yip, E. H. Moors, and A. Pardi. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.*, 264:968–980, 1996.

[37] G Kempf, K. Wild, and I. Sinning. Structure of the complete bacterial SRP Alu domain. *Nucl. Acids Res.*, 42:12284–12294, 2014.

[38] K. S. Kevin S. Keating, N. Toor, and A. M. Pyle. The GANC tetraloop: A novel motif in the Group IIC Intron structure. *J. Mol. Biol.*, 383:475–481, 2009.

[39] D. J. Klein, T. M. Schmeing, P. B. Moore, and T. A. Steitz. The kink-turn: a new RNA secondary structure motif. *EMBO J.*, 20:4214–4221, 2001.

[40] K. D. Kreuzer and T. M. Henkin. The T box riboswitch: tRNA as an effector to modulate gene regulation. *Microbiol Spectr.*, 6:10.1128, 2019.

[41] R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. Rie Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. N. Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio, and D. Baker. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384:NO. 6693, 2024.

[42] D. A. P. Krummel, C. Oubridge, A. K. W. Leung, J. Li, and K. Nagai. Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature*, 458:475–480, 2009.

[43] J. C. Lee, R. R. Gutell, and R. Russell. The UAA/GAN internal loop motif: a new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions. *J Mol Biol*, 360:978–988, 2006.

[44] J. Lehmann, F. Jossinet, and D. Gautheret. A universal RNA structural motif docking the elbow of tRNA in the ribosome, RNAse P and T-box leaders. *NAR*, 41:5494–5502, 2012.

[45] N. B. Leontis and E. Westhof. A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J Mol Biol*, 283:571–583, 1998.

[46] N. B. Leontis and E. Westhof. The 5S rRNA loop E: chemical probing and phylogenetic data versus crystal structure. *RNA*, 4:1134–1153, 1998.

[47] N. B. Leontis and E. Westhof. Analysis of RNA motifs. *Curr Opin Struct Biol*, 13:300–308, 2003.

[48] A. Lescoute, N. B. Leontis, C. Massire, and E. Westhof. Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *RNA*, 33:2395–2409, 2005.

[49] A. Lescoute and E. Westhof. Topology of three-way junctions in folded RNAs. *RNA*, 12:83–93, 2006.

[50] M. Y. Liu, G. Gui, B. Wei, J. F. 3rd Preston, L. Oakford, U. Yuksel, D. P. Giedroc, and T. Romeo. The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in Escherichia coli. *J Biol Chem*, 272:17502–17510, 19971.

[51] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6:1748–7188, 2011.

[52] G. Loyer and V. Reinharz. Concurrent prediction of RNA secondary structures with pseudoknots and local 3D motifs in an integer programming framework. *Bioinformatic*, 40:btae022, 2024.

[53] N. Marmier-Gourrier, A. Clery, V. Senty-Segault, F. Charpentier, B. Schotter, F. Leclerc, R. Fournier, and C. Branlant. A structural, phylogenetic, and functional study of 15.5-kD/Snu13 protein binding on U3 small nucleolar RNA. *RNA*, 9:821–838, 2003.

[54] M. Martick and W. G. Scott. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell*, 126:309–320, 2006.

[55] S. E. Melcher, T. J. Wilson, and D. M. J. Lilley. The dynamic nature of the four-way junction of the hepatitis C virus IRES. *RNA*, 9:809–820, 2003.

[56] M. Meyer, E. Westhof, and B. Masquida. A structural module in RNase P expands the variety of RNA kinks. *RNA Biol*, 9:254–260, 2012.

[57] R. Montange and R. T. Batey. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, 441:1172–1175, 2006.

[58] P. B. Moore. Structural motifs in RNA. *Annu. Rev. Biochem*, 68:287300, 1999.

[59] U. Nagaswamy, N. Voss, Z. Zhang, and G. E. Fox. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res.*, 28:375–376, 2000.

[60] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29:2933–2935, 2013.

[61] N. Ontiveros-Palacios, E. Cooke, E. P. Nawrocki, S. Triebel, M. Marz, E. Rivas, S. Griffiths-Jones, A. I. Petrov, A. Bateman, and B. Sweeney. Rfam 15: RNA families database in 2025. *NAR*, gkae1023, 2024.

[62] M. Parisien and F. Major. The MC-fold and MC-sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–55, 2008.

[63] J. Perard, C. Leyrat, F. Baudin, E. Drouet, and M. Jamin. Structure of the full-length HCV IRES in solution. *Nature Communications*, 1612, 2013.

[64] M. Perbandt, A. Nolte, S. Lorenz, R. Bald, C. Betzel, and V. A. Erdmann. Crystal structure of domain E of Thermus flavus 5S rRNA: a helical RNA structure including a hairpin loop. *FEBS Lett*, 211:211–215, 1998.

[65] A. I. Petrov, C. L. Zirbel, and N. B. Leontis. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, 19:1327–1340, 2013.

[66] M. Popenda, M. Szachniuk, M. Blazewicz, S. Wasik, E. K. Burke, J. Blazewicz, and R. W. Adamiak. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*, 11:231, 2010.

[67] H. Ratni, R. S. Scalco, and A. H. Stephan. Risdiplam, the first approved small molecule splicing modifier drug as a blueprint for future transformative medicines. *ACS Med. Chem. Lett.*, 12:874–877, 2021.

[68] V. Reinharz, A. Soule, E. Westhof, J. Waldispuhl, and A. Denise. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. *Bioinformatics*, 46:38413851, 2018.

[69] N. J. Reiter, A. Osterman, A. Torres-Larios, K. K. Swinger, T. Pan, and A. Mondragón. Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature*, 468:784–789, 2010.

[70] A. Ren, Y. Xue, A. Peselis, A. Serganov, H. M. Al-Hashimi, and D.J. Patel. Structural and dynamic basis for low-affinity, high-selectivity binding of L-Glutamine by the Glutamine riboswitch. *Cell Rep.*, 13:1800–1813, 2015.

[71] J. S. Reuter and D. H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11:10, 2010.

[72] E. Rivas. RNA structure prediction using positive and negative evolutionary information. *PLOS Comput Biol*, 16(10):e1008387, 2020.

[73] E. Rivas. RNA covariation at helix-level resolution for the identification of evolutionarily conserved RNA structure. *PLOS Comput Biol*, 19(7):e1011262, 2023.

[74] E. Rivas, J. Clements, and S. R. Eddy. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, 14:45–48, 2017.

[75] E. Rivas, J. Clements, and S. R. Eddy. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*, 36:3072–3076, 02 2020.

[76] E. Rivas, R. Lang, and S. R. Eddy. A range of complex probabilistic models for RNA secondary structure prediction that include the nearest neighbor model and more. *RNA*, 18:193–212, 2012.

[77] H. P. Robertson, H. Igel, R. Baertsch, D. Haussler, M. Ares Jr., and W. G. Scott. The structure of a rigorously conserved RNA element within the SARS virus genome. *PLOS Biol*, 3:e5, 2004.

[78] I. Sakai. Syntax in universal translation. In *1961 International Conference on Machine Translation of Languages and Applied Language Analysis*, volume II, pages 593–608, Teddington, England, 1962. London: Her Majesty's Stationery Office.

[79] R. Sarrazin-Gendron, V. Reinharz, C. G. Oliver, N. Moitessier, and J. Waldispuhl. Automated, customizable and efficient identification of 3D base pair modules with BayesPairing. *NAR*, 47:3321–3332, 2019.

[80] M. Sarver, C.L. Zirbel, J. Stombaugh, A. Mokdad, and N.B. Leontis. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, 56:215–252, 2009.

[81] U. Schmitz, S. Behrens, D. M. Freymann, R. J. Keenan, P. Lukavsky, P. Walter, and T. L. James. Structure of the phylogenetically most conserved domain of SRP RNA. *RNA*, 5:1419–1429, 1999.

[82] A. Serganov, L. Huang, and D. J. Patel. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature*, 455:1263–1267, 2008.

[83] A. Serganov, A. Polonskaia, A. T. Phan, R. R. Breaker, and D. J. Patel. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, 441:1167–1171, 2015.

[84] A. Smirnov, N. Entelis, I. Krasheninnikov, R Martin, and I. Tarassov. Specific features of 5S rRNA Structure - its interactions with macromolecules and possible functions. *Biochemistry*, 73:1418–37, 12 2009.

[85] E. Sonnleitner, L. Abdou, and D. Haas. Small RNA as global regulator of carbon cataboliterepression in *Pseudomonas aeruginosa*. *PNAS*, 106:21866–21871, 2009.

[86] S. Szep, J. Jimin Wang, and P. B. Moore. The crystal structure of a 26-nucleotide RNA containing a hook-turn. *RNA*, 9:44–51, 2003.

[87] A. A. Szewczak, P. B. Moore, Y. L. Chang, and I. G. Wool. The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl. Acad. Sci. USA*, 90:9581–9585, 1993.

[88] M. Szikszai, M. Magnus, S. Sanghi, S. Kadyan, N. Bouatta, and E. Rivas. RNA3DB: A structurally-dissimilar dataset split for training and benchmarking deep learning models for RNA structure prediction. *J Mol Biol*, 436:168552, 2024.

[89] I. Tinoco and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293:271–281, 1999.

[90] C. Tuerk, P. Gauss, C. Thermes, D. R. Groebe, M. Gayle, N. Guild, G. Stormo, Y. D'Aubenton-Carafa, O. C. Uhlenbec, I. Tinoco Jr., E. N. Brody, and L. Gold. CUUCGG hairpins: Extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci. USA*, 85:1364–1368, 1988.

[91] O. C. Uhlenbeck. Tetraloops and RNA folding. *Nature*, 346, 1990.

[92] I. Vidovic, S. Nottrott, K. Hartmuth, R. Lührmann, and R. Ficner. Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol Cell*, 119:1331–1342, 2000.

[93] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Muller, D. H. Mathews, and M. Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci U S A.*, 91:9218–9222, 1994.

[94] N. J. Watkins, V. Segault, B. Charpentier, S. Nottrott, P. Fabrizio, A. Bachi, M. Wilm, M. Rosbash, C. Branlant, and R. Luhrmann. A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNA. *Cell*, 103:457–466, 2000.

[95] G. Weber, S. Trowitzsch, B. Kastner, Lúhrmann R., and M. C. Wah. Functional organization of the Sm core in the crystal structure of human U1 snRNP. *The EMBO Journal*, 29:4172–4184, 2010.

[96] Z. Weinberg and R. R. Breaker. R2R – software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, 12:3, 2011.

[97] E. Westhof, B. Masquida, and L. Jaeger. RNA tectonics: towards RNA design. *Fold Des.*, 1:R78–R88, 1996.

[98] C. R. Woese, S. Winker, and R. R. Gutell. Architecture of Ribosomal RNA: Constraints on the sequence of "Tetra-loops". *Proc. Natl. Acad. Sci. USA*, 87:8467–8471, 1990.

[99] J. Yao, V. Reinharz, F. Major, and J. Waldispuhl. RNA-MoIP: prediction of RNA secondary structure and local 3D motifs from sequence data. *NAR*, 45:W440–W444, 2017.

[100] L. Zhang, J. Lin, and K. Ye. Structural and functional analysis of the U3 snoRNA binding protein Rrp91–18768. *RNA*, 19:701–x711, 2013.

[101] Q. Zhao, H.-C. Huang, U. Nagaswamy, Y. Xia, X. Gao, and G. E. Fox. UNAC tetraloops: to what extent do they mimic GNRA tetraloops? *Biopolymers*, 97:617–628, 2012.

[102] C. Zhong and S. Zhang. Clustering rna structural motifs in ribosomal rnas using secondary structural alignment. *NAR*, 40:13071317, 2011.

[103] Q. Zhu, L. Petingi, and T. Schlick. RNA-As-Graphs motif atlas–dual graph library of RNA modules and viral frameshifting-element applications. *NAR*, 31:2926–2943, 2023.

[104] C. L. Zirbel, J. Roll, B. A. Sweeney, A. I. Petrov, M. Pirrung, and N. B. Leontis. Identifying novel sequence variants of RNA 3D motifs. *NAR*, 43:7504–7520, 2015.

[105] C. Höner zu Siederdissen, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. A folding algorithm for extended RNA secondary structures. *Bioinformatics*, 27:i129–136, 2011.

Methods

The RNA Basic Grammar including 3-way and 4-way junctions (RBGJ3J4)

The RNA basic grammar (RBG) [75] is a probabilistic stochastic context-free grammars (SCFG) modeling RNA folding, targeting secondary structural elements including stacked canonical base pairs forming helices, hairpin loops, bulge and internal loops, and multiloops.

SCFGs build correlated states at arbitrary distances and hence are well suited for expressing RNA base pairing. Moreover, the nonterminals directly correspond to the overarching secondary structures. For example, the non-terminal `F` builds a helix, while the non-terminal `P` is present after a helix has ended and is able to initiate all possible loop features such as a hairpin loop, internal loop, or multiloops (Figure S1).

Because many RNA 3D motifs occur in multiloops, especially the most frequent 3- and 4-way junctions, here we introduce the RBGJ3J4 grammar which singles out 3-way junctions (multiloops framed by 3 helices) as well as 4-way junctions (multiloops framed by four helices) as specific cases from any other higher order multiloop. See Figure 2a and Figure S1 for the full description of the RBGJ3J4 grammar.

Another unique feature of the RBG and RBGJ3J4 grammars as they are used in CaCoFold is that they fold an alignment, not a single sequence. A position does not represent one residue but a probability vector describing the frequency of each residue in the aligned column. This way, we can produce consensus secondary structures that include the information contained in all aligned sequences.

The RBGJ3J4-R3D joint grammar uses the CYK folding algorithm

The R3D motif SCFGs are integrated into the RBGJ3J4-R3D grammar as described in Figure 2. CaCoFold-R3D admits an arbitrary number of 3D motifs. Figure S3 describes the list of 51 motifs used in this manuscript. CaCoFold-R3D internally interprets the descriptor lists and implements a SCFG for each motif according to the general R3D grammars for each of the six types of motifs as described in Figure 3.

RNA 3D motifs bound by two or more helices, can appear in different configurations depending on which of the ends corresponds to the $5'/3'$ ends of the molecule, versus all the others for which the backbone is continuous. That is, `BLs` and `ILs` can have two configurations, while `J3s` and `J4s` can have three and four respectively (Figure S2). For a given representation of the motif, R3D internally implements and adds all possible configurations which get added to the RGBJ3J4-R3D grammar.

As with the RBGJ3J4 grammar, all R3D SCFGs describe a consensus motif in an alignment, not just a particular sequence motif. Thus, R3D is able to represent the variability observed in the motif. The input to the algorithm is not one nucleotide per sequence position, but L probability distributions of dimension 4 describing the frequency of each nucleotide per alignment position (a probabilistic sequence), and $L \times L$ distributions describing join 4×4 pair probabilities.

RBGJ3J4-R3D implements the CYK folding algorithm [78, 11] to report the consensus fold and consensus 3D motifs that maximize the probability of the alignment. This CYK algorithm is applied not to one single sequence, but to a probability sequence where for each position, there is not one given residue but a probability distribution representing the frequency of each nucleotide in each alignment position. This generalized CYK on probability sequences has the same algorithmic complexity as the original CYK on individual sequences.

The CaCoFold-R3D multilayered folding and 3D motif prediction method

CaCoFold-R3D uses a multilayered folding approach constrained by covariation. The first layer includes a maximal set of nested covarying base pairs and uses the RBGJ3J4-R3D CYK algorithm introduced here which replaces the RBG CYK algorithm of CaCoFold [72]. This first layer folding identifies the main secondary structure together with any 3D motif to produce a structure with the maximum probability using the computationally efficient CYK algorithm.

The treatment of the higher layers of CaCoFold-R3D is similar to CaCoFold. When the covariation found in the alignment by R-scape cannot be fully explained by a nested structure, those non-nested covarying pairs are incorporated in additional folding layers. The number of layer depends on the number additional clusters of nested covarying pairs that can be constructed. These subsequent folding layers of CaCoFold-R3D use a simpler folding grammar (named G6XS and described in [72]) to introduce additional helices supported by covariation that could be representing any combination of pseudoknots, alternative helices or triplet interactions with covariation support. In that form, CaCoFold-R3D is able to identify pseudoknotted helices as folds appearing at subsequent layers along with any 3D motif that the first CYK layer may have already identified for that loop.

As an example, the Glutamine riboswitch structure presented in Figure 5 identifies a pseudoknot (one single base pair with covariation support) and a LoopE 3D motif in the same three-way junction; both motifs are confirmed by the crystal structure [70].

There may be situations in which the secondary structure or the 3D motif annotation reported by the first layer overlaps with a pseudoknot or triplet reported by the higher layers. Those could represent conflicts or they could be alternative structures. We leave any of those unmodified for the user to see and make further assessments.

The sequence-motif profile HMMs

Each sequence motif is characterized by a consensus sequence $S_1 \dots S_L$, and it gets assigned a profile HMM (Figure 3g). The profile HMM introduces one consensus state per position S_i, for $1 \leq i \leq L$. Each consensus state is characterized by a consensus residue or residue type such as: A, or R (A or G), or Y (C or U) or others that determine the emission probability of the consensus. There is an error probability to allow any of the other residues not designated by the consensus. That is a $S = R$ position assigns $P_S(A) = P_S(G) = 0.5 - \epsilon/2$ and $P_S(C) = P_S(U) = \epsilon/2$. The profile HMM also includes one insertion states per position. All inserted position use the same emission residue probability distribution matching the residue frequencies of the training set. Emissions are done on transition, and transitions without emissions describe deletion events.

Each profile HMM is parameterized by a length distribution with an expected length closely matching that of the consensus motif, and allowing insertions and deletions relative to consensus. Empty sequence motifs are also modeled by a profile HMM to allow for the possibility of more divergent motif examples with insertions.

The profile HMMs described in Figure 3g can be used to score either individual sequence positions or alignment positions. Introducing a probability distribution per position o, $\{p_o(a)\}_{a=A,C,G,U/T}$, the consensus and indel emission probabilities of the profile HMM can be written as

$$P_{S_i}(o) = \sum_{a=A,C,G,U} p_o(a) P_{S_i}(a),$$
$$P_{\text{indel}}(o) = \sum_{a=A,C,G,U} p_o(a) P_{\text{indel}}(a).$$

For the case of an aligned position, p_o is the position base composition. For the case of a particular sequence, there is a single residue per position $o = a$, such that $p_o(o = a) = 1$, and

$$P_{S_i}(o = a) = P_{S_i}(a),$$
$$P_{\text{indel}}(o = a) = P_{\text{indel}}(a).$$

For each tested segment (in alignment or sequence), the profile HMM calculates the probability of the segment given the motif using the forward algorithm. The probability for a given subsequence is incorporated into the corresponding R3D SCFG where the sequence segment is included.

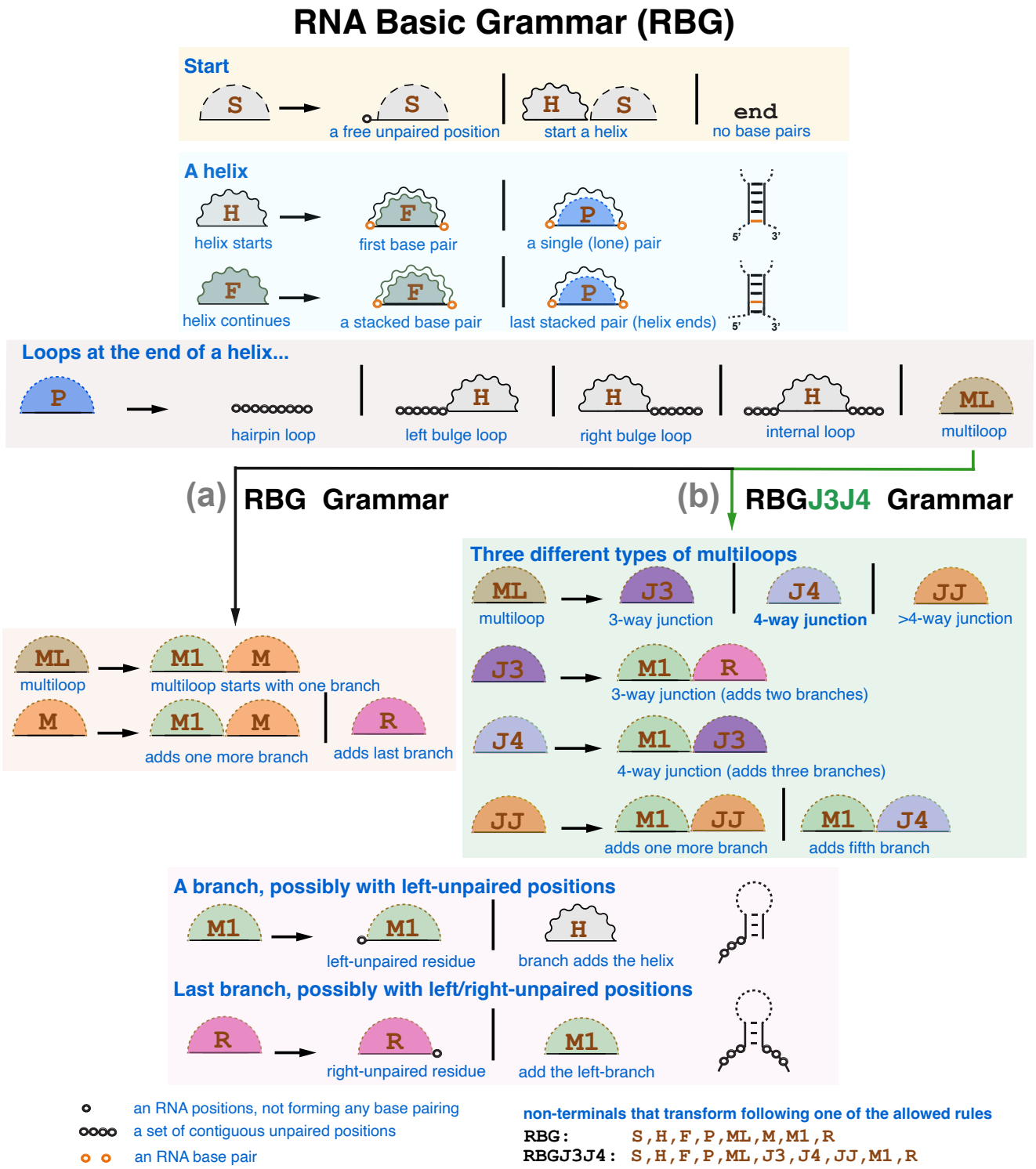


Figure S1: **The RBG and RBGJ3J4 generative grammars.** Inserts **(a)** and **(b)** describe the two distinct realizations of the multiloop non-terminal `ML`. RBGJ3J4 replaces the generic multiloop non-terminal `M` with non-terminals `J3`, `J4` and `JJ` in order to distinguish 3-way and 4-way junctions from other higher order multiloops. A solid line represent the RNA sequence, a curly line indicates that the two connecting residues are base paired, and a dashed line indicates that the relationship between the two residues is yet undetermined. Non-terminals are depicted in brown and actual residues/positions are depicted with circles (black for unpaired and orange for base paired positions). Each non-terminal describes a discrete random variable of events which are enumerated on the right-hand side of the arrows. The allowed events for a given non-terminal are separated by the | ("or") symbol. Starting from the **S** non-terminal, an RNA sequence/structure is produced by sampling from the discrete probability distributions (transitions and emissions) associated to each non-terminal.

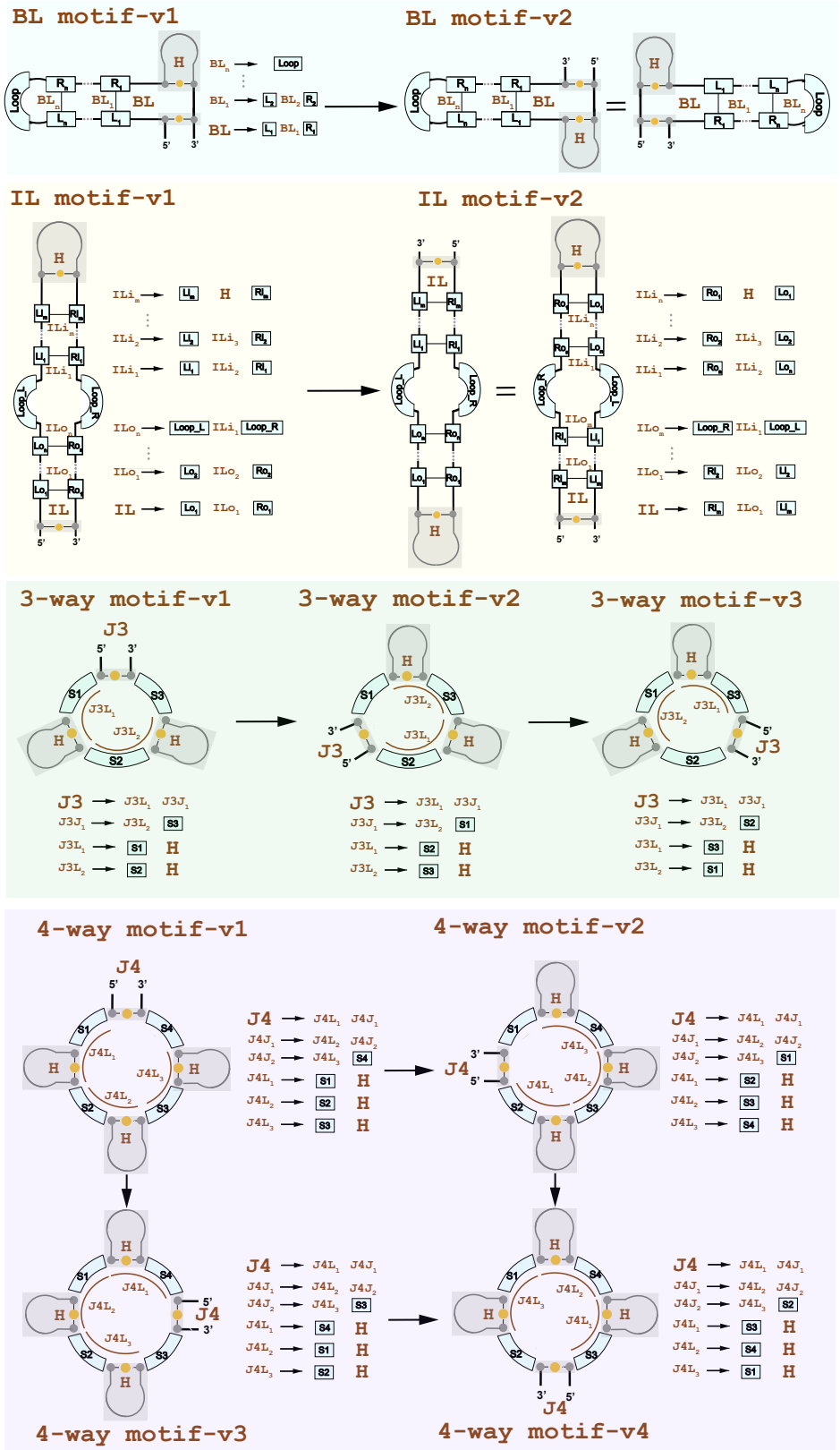


Figure S2: **RNA 3D motifs variants.** RNA 3D motifs bound by more than one helix (*i.e.* all except for hairpin loops) allow different topological variants depending on which $5'/3'$ ends are selected to integrate the motif into the rest of the structure. Bulge and Internal loop motifs have two variants, and 3-way and 4-way junctions have three and four variants respectively. For any 3D motif entry, CaCoFold-R3D internally models all possible variants of the motif.

```
#HL   Loop(5'-3')   L(5'-3')       R(5'-3')       name
#-----------------------------------------------------------------------------------------------------------
 HL   N             G              RA             GNRA-tetraloop
 HL   URA           -              -              U-turn
 HL   UNCG          -              -              UNCG-tetraloop
 HL   ANYA          -              -              ANYA-tetraloop
 HL   CUYG          -              -              CUYG-tetraloop
 HL   YGNN          -              -              YGNN-tetraloop
 HL   GANC          -              -              GANC-tetraloop
 HL   UNAC          -              -              UNAC-tetraloop
 HL   URRR          -              -              T-loop-tetraloop
 HL   UAACR         -              -              L8_RNaseP_bact_a
 HL   AG            UAGUACG        AGGACC         Sarcin-ricin_loop
 HL   AGGAY         -              -              CsrA_binding
 HL   GAGUA         -              -              GAGUA_pentaloop
 HL   AA            -              -              AA-5SrRNA
 HL   GCRYA         -              -              U6-loop

#BL   Loop(5'-3')   L(5'-3')       R(5'-3')       name
#-----------------------------------------------------------------------------------------------------------
 BL   UGRAA         -              -              Docking-elbow
 BL   AA            -              -              AA-bulge
 BL   G             -              -              bulged-G

#IL   Loop-L(5'-3') Loop-R(5'-3')  L-o(5'-3')     R-o(5'-3')     L-i(5'-3')   R-i(5'-3')    name
#-----------------------------------------------------------------------------------------------------------
 IL   -             RNN            N,G,A          N,A,R          -            -             K-turn
 IL   -             RNN            G,A            A,R            -            -             K-turn-b
 IL   G             N              G,A            A,U            U,A          A,G           Loop-E
 IL   UAAG          UAU            C,C            G,G            -            -             GAAA_Tetraloop-receptor
 IL   CVC           V              -              -              -            -             C-loop
 IL   G             -              G,A            A,A            U,A          A,G           G-bulge
 IL   G             -              U,A            C,C            U,A          A,G           G-bulge_Das
 IL   -             -              G,A            A,G            -            -             Tandem-GA
 IL   YCC           AAC            -              -              -            -             Twist-up
 IL   YAA           RAN            -              -              -            -             UAA_GAN
 IL   RR            YN             R              R              A            Y             J4a/4b
 IL   AA            AA             -              -              -            -             J4/5-IL
 IL   GUA           GG             -              -              -            -             GUA^GG_RRE
 IL   CAGG          AGCA           -              -              -            -             S_domain
 IL   AN            -              G,A            A,U            -            -             Hook-turn
 IL   UGRAA         -              -              -              -            -             Docking-elbow-IL
 IL   UU            AUU            -              -              -            -             pK-turn

#J3   S1(5'-3')     S2(5'-3')      S3(5'-3')      name
#-----------------------------------------------------------------------------------------------------------
J3    N             CUGA           A              J3_hammerhead
J3    U             YUCUAC         AC             J3_purine
J3    NN            NNNNNN         NN             J3_typeA
J3    NNNN          NNNN           NNNN           J3_typeB
J3    NN            NNN            NNNNNN         J3_typeC
J3    N             UGAGA          N              J3_TPP
J3    A             RAA            -              J3_groupII

#J4   S1(5'-3')     S2(5'-3')      S3(5'-3')      S4(5'-3')     name
#-----------------------------------------------------------------------------------------------------------
J4    -             AA             -              U            J4_HCV_IRES
J4    N             -              NNN            -            J4_tRNA
J4    N             N              NN             -            J4_manA
J4    R             -              -              R            J4_U1RNA

#BS  Loop(5'-3')      name
#-----------------------------------------------------------------------------------------------------------
BS    UKNRW            T-loop
BS    RRGU             LoopE-a
BS    RARR             LoopE-b
BS    AAYAARAACAANARR  CRC_binding
BS    AGGAY            CsrA_motif
```

Figure S3: **RNA 3D motifs descriptors.** The descriptor file includes 51 different distinct motifs. CaCoFold internally constructs SCFGS for a total of 96 motif variants. The 96 motif R3D SCFGs get integrated into the RBGJ3J4 grammar. The RBGJ3J4-R3D grammar folds and detect the motifs of the RNA simultaneously. HL = Hairpin Loop, BL = Bulge Loop, IL=Internal Loop, J3 = 3-way Junction, J4 = 4-way Junction, BS = Branch Segment.

RFAM SEED ALIGNMENTS
4,178 Families

Type	Motif	# Motif instances with cov support (all)	# Rfam families with motif with cov support (all)	SSU eukarya with support (all)	LSU eukarya with support (all)
HL	GNRA_tetraloop	170 (214)	101 (132)	4 (4)	6 (7)
IL	K_turn	68 (83)	54 (68)	2 (2)	2 (2)
BL	Docking_elbow	59 (94)	55 (83)	2 (2)	2 (5)
J3	J3_groupII	52 (57)	40 (43)	1 (1)	1 (1)
HL	UNCG_tetraloop	48 (78)	42 (69)	1 (1)	0 (0)
J3	J3_typeA	46 (49)	36 (39)	4 (4)	4 (4)
BL	AA_bulge	45 (101)	34 (81)	1 (1)	2 (2)
IL	pK_turn	45 (82)	41 (75)	0 (0)	0 (0)
J3	J3_typeC	43 (45)	42 (43)	0 (0)	1 (1)
J4	J4_U1RNA	40 (44)	29 (31)	0 (0)	3 (4)
IL	UAA_GAN	40 (50)	23 (33)	0 (0)	5 (5)
IL	G_bulge	37 (52)	30 (39)	0 (0)	1 (1)
IL	K_turn_b	35 (50)	32 (47)	0 (1)	1 (1)
J4	J4_HCV_IRES	33 (36)	25 (28)	3 (3)	1 (1)
IL	Hook_turn	32 (49)	28 (45)	0 (0)	3 (3)
IL	Tandem_GA	29 (37)	19 (25)	2 (3)	3 (3)
BS	LoopE_a	28 (33)	15 (20)	1 (1)	4 (4)
BS	T_loop	27 (33)	25 (30)	1 (1)	2 (2)
J4	J4_manA	26 (27)	24 (25)	0 (0)	2 (2)
IL	C_loop	26 (42)	25 (40)	1 (1)	0 (0)
BS	CsrA_motif.rev	26 (30)	23 (27)	2 (2)	1 (1)
HL	UNAC_tetraloop	26 (48)	26 (47)	0 (0)	1 (1)
BS	CRC_binding	26 (34)	18 (22)	5 (5)	0 (0)
J4	J4_tRNA	25 (26)	19 (20)	3 (3)	3 (3)
HL	T_loop_tetraloop	24 (43)	20 (34)	1 (3)	0 (0)
BS	LoopE_b.rev	24 (25)	18 (18)	1 (1)	1 (1)
BS	LoopE_b	21 (22)	15 (16)	1 (1)	2 (2)
BL	bulged_G	20 (43)	18 (37)	0 (0)	0 (0)
BS	T_loop.rev	20 (26)	17 (22)	1 (1)	3 (3)
HL	CUYG_tetraloop	20 (44)	20 (43)	0 (0)	0 (0)
HL	AA_5SrRNA	19 (19)	18 (18)	0 (0)	0 (0)
HL	ANYA_tetraloop	18 (39)	16 (35)	0 (0)	1 (1)
IL	J4a/4b	16 (37)	16 (35)	0 (0)	1 (1)
J3	J3_typeB	16 (16)	14 (14)	2 (2)	0 (0)
J3	J3_hammerhead	16 (27)	16 (26)	0 (0)	1 (1)
HL	L8_RNaseP_bact_a	16 (29)	14 (26)	0 (0)	0 (1)
BS	LoopE_a.rev	16 (21)	13 (17)	0 (0)	1 (1)
HL	U_turn	15 (24)	14 (23)	0 (0)	0 (1)
HL	GANC_tetraloop	15 (28)	14 (26)	0 (0)	2 (2)
BS	CRC_binding.rev	15 (18)	8 (11)	6 (6)	0 (0)
BS	CsrA_motif	15 (18)	13 (16)	0 (0)	0 (1)
J3	J3_TPP	14 (19)	12 (17)	0 (0)	0 (1)
HL	YGNN_tetraloop	13 (27)	12 (26)	0 (0)	0 (0)
IL	GUA_GG_RRE	10 (13)	10 (12)	0 (0)	0 (0)
HL	CsrA_binding	10 (29)	7 (14)	0 (1)	0 (0)
HL	GAGUA_pentaloop	10 (22)	9 (21)	0 (0)	0 (0)
IL	S_domain	10 (21)	10 (20)	0 (0)	0 (0)
IL	Loop_E	9 (11)	9 (10)	0 (0)	1 (2)
IL	Twisted_up	9 (31)	9 (30)	0 (0)	0 (0)
J3	J3_purine	7 (11)	7 (11)	0 (0)	0 (0)
IL	GAAA_Tetraloop_receptor	7 (17)	7 (17)	0 (0)	0 (0)
HL	U6_loop	7 (18)	7 (18)	0 (0)	0 (1)
IL	J4/5_IL	6 (13)	6 (13)	0 (1)	0 (0)
HL	Sarcin_ricin_loop	5 (7)	5 (7)	0 (0)	1 (1)
IL	G_bulge_Das	5 (11)	5 (11)	0 (0)	0 (1)
IL	Docking_elbow_IL	0 (1)	0 (1)	0 (0)	0 (0)

Rfam SEEDS
4,178 alignments

	# Motif instances with cov support (all)	# Rfam families with motif with cov support (all)	SSU eukarya with support (all)	LSU eukarya with support (all)
3D Motifs all	1460 (2124)	591 (822)	45 (51)	62 (74)
Nested helices	3877 (15395)	1721 (4168)	66 (89)	102 (152)
PK + higher order	504 (532)	246 (246)	1 (2)	2 (4)

Rfam CONTROLS
4,178 shuffled alignments

3D Motifs all	121 (290)	106 (208)	0 (0)	0 (1)
Nested helices	733 (14146)	676 (4149)	0 (34)	0 (75)
PK + higher order	175 (188)	102 (102)	0 (0)	0 (0)

Table S1. **RNA 3D motifs found in Rfam.** Results of running CaCoFold-R3D for all seed alignments in the database of structured RNAs Rfam. Results show RNA 3D motif occurrences with covariation support where at least one of the closing helices has at least one covarying base pair, and in parenthesis the total set of predictions. For the other structural elements (nested helices, pseudoknots and other higher order base pair interactions), having covariation support means that the element includes at least one significantly covarying pair of residues. In parenthesis we show the total number of predictions. The control shuffled alignments were obtained by randomizing the residues within each alignment column independently from each other. In these control alignments, covariation between columns is scrambled, but the base composition per column (thus the possible 3D motif identity) remains mostly intact. Results of running CaCoFold-R3D on the Rfam seed alignments were obtained using the command line `R-scape -s --cacofold --r3d Rfam.seed`. The control experiments were obtained using the command line `R-scape -s --cacofold --r3d --vshuffle Rfam.seed`.