Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

RNA Nanomed. Author manuscript; available in PMC 2026 November 07.

Published in final edited form as:

RNA Nanomed. 2026 November 5; 3(2): 1–12.

Viral RNAs as Dual Graphs: Extending the Motif Universe of RNAs

Jad Zbib[1], **Tamar Schlick**[1,2,3,4]

[1]Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012, USA

[2]Department of Chemistry, New York University, 100 Washington Square East, New York, NY 10003, USA

[3]NYU-ECNU Center for Computational Chemistry, NYU Shanghai, Shanghai 200062, China

[4]NYU Simons Center for Computational Physical Chemistry, New York University, 24 Waverly Place, New York, NY 10003, USA

Abstract

In the evolving landscape of RNA research, the classification and analysis of RNA motifs is necessary to uncover the intricate mechanisms governing cellular and viral processes. Here we apply the coarse-grained RAG (RNA-As-Graphs) framework to advance the classification and understanding of RNA motifs, with a focus on expanding the RNA Motif Atlas through the inclusion of novel viral RNA structures. By analyzing 273 experimentally resolved viral RNA structures from the Protein Data Bank (PDB) using RAG dual-graph representations, we identify 14 previously uncatalogued viral RNA motifs. These motifs, which include tRNA-mimicking domains, exoribonuclease-resistant domains, and internal ribosome entry sites, expand the diversity of RNA to a total number of 197 dual-graph motifs. We applied k-means, PAM, and Ward clustering and observed substantial overlap between viral and general RNAs. The expanded library of RNA motifs and submotifs provides a resource for motif discovery and RNA design.

Keywords

Viral RNA Structures; Coarse-Grained Graphs; Novel RNAs; RNA Design

INTRODUCTION

In recent years, RNA has enjoyed heightened focus from the scientific community, as evidenced by the expanding body of research on RNA structure, function, and interactions.

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

[*]**Please send correspondence to** Tamar Schlick, Schlick@nyu.edu.

AUTHOR CONTRIBUTIONS

J.Z. and T.S. designed the work; J.Z. performed the computations; J.Z. and T.S. wrote and edited the manuscript.

CONFLICTS OF INTEREST

None declared.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Extensive research has been dedicated to unraveling the complexities of coding and noncoding RNA, including interactions with DNA and proteins, and their role in gene expression and genome maintenance (1). The intricate involvement of different types of RNA in numerous cellular and biological processes has also drawn attention to RNAs as potential therapeutic targets.

Notable examples include the mRNA-based COVID-19 vaccines developed by Pfizer and Moderna. In addition, a range of siRNA-, miRNA-, and aptamer-RNA-based therapeutics are in clinical development, with some granted FDA approval (2; 3). RNA therapies have the advantage of lacking significant genotoxic effects compared to DNA therapies (4) and could represent safer, direct genome-based interventions.

RNA secondary structure refers to the pairing and unpairing of ribonucleotides within a single-stranded polynucleotide chain. These nucleotide pairings are driven by various canonical and non-canonical base-base interactions, such as Watson-Crick, Hoogsteen, and sugar-edge (5). Each type of pairing may exhibit stereoisomeric variations (cis and trans). RNA secondary structure often contains defined motifs, such as hairpin loops, pseudoknots, and quadruplexes, which are employed in many biological processes. A shared structural basis can be associated with different functions, while similar biological functions, such as transcription initiation and nuclease resistance, can also exploit a variety of RNA motifs (6; 7). This versatility underscores the complexity and variability in RNA behavior, making RNA an interesting subject for computational biology research.

Coarse-grained models, such as graph representations, are useful for representing RNAs because they simplify complex molecules while retaining key structural and functional characteristics of interest. Such models reduce computational complexity by grouping smaller molecular building blocks into larger units, thereby reducing the number of degrees of freedom and accelerating the study of larger RNA molecules.

In the 1970s and 1980s, pioneers such as Waterman, Shapiro, Nussinov, and others introduced various representations in which RNA elements such as stems, single strands, and loops were assigned to different graph objects. These early works used graphs to model and define RNA secondary structures, representing multiple base pairs as edges and single-stranded RNA as vertices (8; 9). These representations facilitated structural comparisons, and graph-derived objects invited novel mathematical approaches for analyzing and describing RNA structure, such as approximating the free energy of RNA structures using adjacency matrices (8; 10). Early graph representations fell short in capturing complex motifs, most notably pseudoknots, which are critical in understanding RNA’s biological functions. A pseudoknot is a super secondary structure motif formed when two base pairs intertwine.

In 2003 (11), dual graphs for representing RNA were introduced by the Schlick group, enabling the representation of pseudoknots. By representing RNA structures as graphs, where nodes denote stems, and edges denote single-stranded RNA, we can capture the essential topological and geometric features of RNA. This abstraction facilitates the analysis of RNA folding, dynamics, assembly, and interactions, making it easier to identify patterns

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and motifs critical to biological functions, as recently exploited to estimate the size of the RNA motif universe (21). Such reduced models proved especially useful for designing minimal mutations in the SARS-CoV-2 frameshifting element and for describing folding landscapes (19; 20; 43).

The development of the RNA-as-Graphs (RAG) Motif Atlas (11) began with the more intuitive tree graphs, where loops are vertices and stems are connecting edges. However, tree graphs cannot represent pseudoknots. Pseudoknots are recurring motifs in both coding and non-coding RNAs, and they play roles ranging from regulating translation initiation to controlling translation elongation and frameshifting (32; 35). To represent pseudoknots, dual graphs use the opposite convention, in which loops are edges and stems are vertices.

Figure 1 shows the dual graph of a three-way junction RNA, and Figure 2 shows the graph of a large, more complex pseudoknotted RNA. Using dual graphs, pseudoknots (loop-loop or loop-stem bonding between two hairpin structures) can be represented, as shown in Figure 2.

While RNA generally serves as a versatile agent for carrying genetic information and participating in various cellular processes, viral RNAs likely possess unique characteristics due to distinct strategies for maintaining and propagating a viral life cycle.

Here, we expand the RAG motif atlas by adding 14 previously uncatalogued viral motifs derived from 273 viral-associated RNAs, bringing the total to 197 motifs (24).

Next, to assess whether viral RNAs occupy a distinct region, we cluster all RNA motifs using k-means (25), PAM (26), and Ward (27) (See Methods). The resulting clustering shows that viral and general RNAs do not occupy distinct clusters in RNA motif space; rather, they overlap, indicating shared basic RNA submotifs. Finally, we highlight the shared motifs between viral and general RNAs that rely on common secondary-structure building blocks.

MATERIAL AND METHODS

Dual Graph Definitions

Dual graphs were defined by representing RNA double-stranded stems as vertices and intervening single-stranded regions as edges. A stem containing at least two consecutive base pairs is represented as a vertex, and isolated single base pairs are ignored. Hairpin loops are represented as self-edges. Junctions connecting two helices are represented as edges between the corresponding vertices. The dangling 5′ and 3′ ends are ignored.

Small interruptions between adjacent base pairs are handled during graph construction as described below. In this implementation, neighboring helices separated by bulges or internal loops of up to 5 nucleotides are merged into a single stem vertex. In contrast, larger interruptions are treated as separating adjacent helical stems into distinct vertices connected by the corresponding loop or junction edge (11; 24).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RAG Library

The RAG library can be accessed at http://www.biomath.nyu.edu/?q=rag/home. By iteratively connecting smaller graphs, 110,668 dual graphs of 1–9 vertices have been enumerated (15). The RAG library organizes dual graphs by vertex count V, which are then sorted in ascending order of their Fiedler values. Each graph is assigned an ID V_n, ordered by increasing compactness for each V.

RAG Framework

RAG defines 2D RNA structures as tree or dual graphs. The RAG framework includes tools for subgraph partitioning (12) and clustering and scoring models (15; 13). Partitioning is valuable because of the inherent modularity of RNA structures (13). The RAG partitioning (12) decomposes larger parent dual graphs into connected subgraphs, or irreducible graph modules, using spectral graph properties, including the Fiedler-vector framework.

Additionally, partitioning graphs can enhance the efficiency of computational algorithms for RNA structure prediction and comparison, thereby aiding the systematic exploration of RNA architecture and its biological implications. Graph partitioning was improved in 2019 (14) with the introduction of an algorithm that preserves pseudoknots and junctions, as illustrated by partitioning one large virus (Figure 2). The resulting modules are interpreted as recurrent RNA submotifs within larger RNA architectures.

The dual graph enumeration algorithm allows us to expand the RAG atlas to include 2288 tree graphs with 1–13 vertices and 110,668 dual graphs with 1–9 vertices (15). Graphs that have real corresponding (experimentally resolved) RNA structures are denoted as ‘existing’. Here we refer to the 183 previously existing RNAs as general RNAs, to distinguish them from the viral RNA motifs we explore. Here, graphs that have not yet been assigned to a real RNA structure are called ‘hypothetical’.

Graph Representation

Secondary structures in dot-bracket representation were extracted from PDB structures using 3DNA-DSSR (40) and converted into connectivity table (CT) files using dot2ct from the RNA-structure package (41). CT files were processed using dualGraphs.py, based on the Schlick laboratory’s Existing-Dual-Search repository. The script extracts CT base-pairing patterns, groups consecutive paired residues into helices, assigns helices as vertices, and assigns single-stranded regions connecting helices as edges.

For this analysis, we used a lightly modified version of the helix-merging step in dualGraphs.py, merging helices separated by interruptions of up to 5 nucleotides. Isolated single-base-pair helices are removed. The resulting adjacency matrix is matched against the RAG dual-graph library to assign the corresponding graph IDs (11, 24).

Because the dual graph is constructed from the annotated base-pairing pattern, alternative base-pairing annotation methods could alter the inferred helices, loops, and junctions, thereby changing the assigned dual graph for some RNAs. To extract dual subgraphs from newly assigned structures, we used subgraphs.py.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Starting from an assigned graph ID, subgraphs.py retrieves the corresponding adjacency matrix, partitions the graph into irreducible components (see below under Dual Subgraphs/Partitioning), and assigns graph IDs to the resulting subgraphs by matching them to the RAG dual-graph library.

Laplacian and Fiedler Eigenvector

Every graph with n vertices can be described by an $n \times n$ adjacency matrix A with entries a_{ij}, where a_{ij} is the number of edges between vertices i and j, and a_{ii} is set to 2 for self-connected vertices. There also exists, for every graph, a degree matrix D where the a_{ii} entry equals the number of incident edges at vertex i. The Laplacian $n \times n$ matrix $L = D - A$ is computed by subtracting the adjacency matrix A from the degree matrix D.

The Laplacian is a positive semidefinite matrix and, as such, has non-negative eigenvalues, with $\lambda_1 = 0$ being the smallest eigenvalue, with corresponding eigenvector $\mu_1 = (1, 1, \ldots, 1)$. The focus is on the second smallest eigenvalue λ_2, also known as the Fiedler value, which represents the algebraic connectivity of the graph, and its corresponding eigenvector, the Fiedler vector (42). A higher Fiedler value suggests greater algebraic connectivity.

Dual Subgraphs/Partitioning

An articulation point is a vertex whose removal, along with the associated edges, results in a disconnected graph. The dual graph partitioning algorithm preserves connected components by only partitioning graphs along articulation points. The partitioning algorithm leverages the principles of a gap-cut algorithm, which efficiently partitions graphs into weakly interconnected regions joined by an articulation point (44).

To partition a graph, its Fiedler vector is derived, and the gap-cut algorithm identifies the largest ‘gap’ between the Fiedler vector components, which correspond to vertices of the graph. This also preserves pseudoknots, junctions, and internal loops, which are often closely related to RNA function (12). A subgraph is the resulting graph after partitioning and can be thought of as a connected component or a recurrent structural submotif within the larger parent RNA graph. An example is given in Figure 2.

Fiedler Scoring and Clustering

Although the Fiedler value provides insight into a graph’s algebraic connectivity, which indicates how tightly connected the graph is, it is not always sufficient to distinguish between graphs with similar connectivity but differing structural composition. We extract additional features by analyzing the entries of the Fiedler vectors, which help capture more information about the graph’s structure. For a graph with n vertices, there is a one-to-one correspondence between vertices and the entries $\mu_{2,i}$ of the Fiedler vectors $(\mu_{2,1}, \mu_{2,2}, \ldots, \mu_{2,n})$, which is unique for each Laplacian up to a sign change. The arrangement of vertices in a graph is reflected in the distribution of the Fiedler vector components $\mu_{2,i}$. Common topological modules in dual graphs show distinct patterns in this distribution. Vertices in linear modules tend to have Fiedler vector components that increase monotonically (13). Similarly, vertices forming K-way junctions with similar free-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

end branches often share the same Fiedler vector component values. When taken as an ordered set, these components tend to increase monotonically in common RNA dual graph modules. Therefore, given that adjacent vertices have close $\mu_{2,i}$ values that increase from one end of the graph to the other, analyzing the distribution of the ordered components provides additional structural insights that are then used to cluster graphs (13).

Features (s, e)

The scoring and motif search algorithms follow the approach outlined in the 2021 paper (13). The clustering process itself relies on parameters (s, e) which are derived from the sorted and rescaled entries of each graph's Fiedler vector μ_2 (the second eigenvector of the graph's Laplacian). To calculate these parameters for a particular graph:

1. Calculate the normalized Fiedler vector μ_2 and arrange the components $\mu_{2,i}$ in ascending order.
2. Scale each μ_i by:

$$v_i = \mu_i \frac{n-1}{\mu_n - \mu_1}$$

3. Perform linear least squares regression on the points $(1, v_1), (2, v_2), \ldots, (n, v_n)$ to obtain the slope s and the mean square error e. For each graph i, the planar coordinates (s, e) are plotted.

Scoring Model

The scoring model is constructed by incorporating the extracted features (s, e), which are plotted on a plane, along with the graphs' associated weights from a training set that defines the clustering.

An initial score for graph i in the training set, ES_i is calculated based on the weight w_i of graph i, with σ and ϵ as freely chosen parameters that control the influence of the weight score.

$$ES_i = \sigma \, log(w_i) + \epsilon$$

Each graph i, represented as a point in the plane, contributes to the score of a graph j based on their Euclidean distance d_{ij}, via the pairwise score $S_{j,i}$:

$$S_{j,i} = ES_i \times \exp\left[-\left(\frac{rd_{ij}}{x_i}\right)\right]$$

where r controls the effect of weight on the pairwise score and x_i is a scale parameter.

For each graph j, the total score it receives is:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\sum_{i=1}^{k} S_{j,i}$$

where k is the total number of graphs in the training set.

Clustering

We clustered (s, e) points with three standard methods: K-means (25), PAM (26), and Ward's linkage (27). For all methods, we set $K = 2$; for Ward, we cut the dendrogram at $K = 2$. Clusters are summarized by the overall mean silhouette and by cluster (28). Specifically, K-means solves

$$\min_{\{C_k, \mu_k\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \qquad \mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i .$$

PAM selects medoids $m_k \in C_k$ to minimize

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, m_k)$$

for a chosen dissimilarity $d(x_i, x_j) = \|x_i - x_j\|_2$.

After fitting each method, we relabel clusters so that Cluster 1 (C_1) is the cluster containing the larger number of viral IDs among the known RNAs, and Cluster 2 (C_2) is the other cluster. We report the overall mean silhouette and the per-cluster mean silhouettes for C_1 and C_2 (28).

Clustering was performed in the (s, e) feature space, where (s, e) are derived from the sorted and scaled components of the graph Fiedler vector. Together, these features summarize the distribution of Fiedler-vector entries and thus encode coarse-grained properties of the assigned dual-graph topology (13). We did not assign separate weights to different structural features after graph construction. All features were analyzed only through the final dual-graph topology and the associated Fiedler-based descriptors.

Significance testing

To assess whether cluster membership and viral label are associated among known RNAs, we applied Fisher's exact test on the 2×2 table of label (viral, general) by cluster (C_1, C_2). For each cluster separately, we additionally tested whether its viral fraction exceeded the overall viral fraction using a right-tailed Fisher's exact test. We report p-values, odds ratios, and 95% confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Available Code

The dual-graph construction workflow used in this work was based on the Schlick laboratory's Existing-Dual-Search repository, available at https://github.com/Schlicklab/Existing-Dual-Search. The general dual-graph construction code is provided in the repository as dualGraphs.py.

For this work, we used a slightly modified version of the one provided in Existing-Dual-Search/Jad/dualGraphs_ILmax5.py, which merges neighboring helices separated by interruptions of up to 5 nucleotides. Graph partitioning and subgraph identification were performed using subgraphs.py. Clustering analyses were performed in MATLAB using standard implementations of K-means, PAM, and Ward's linkage.

RESULTS

Viral Graph Motifs and Submotifs

From Figure 4, we see that the most common structure is 1_1 (138 graphs), followed by 2_1, 2_2, 2_3, and 3_1. Graphs 1_1 and 2_1 represent some of the simplest secondary structures: a stem-loop (hairpin) and a stem-loop stem (double hairpin), respectively. A hairpin loop consists of a variable-length double-stranded RNA (possibly containing mismatched nucleotides) and a terminal apical loop. Hairpin loops appear to have evolved independently in many RNA pathways across a multitude of processes, including mRNA localization, protein-binding sites, ribozyme elements, and gene-silencing pathways (34; 14). In the 273 viral RNA structures, single or double hairpin loops were present in the trans-activating response (TAR) element of HIV-1 (PDB ID: 6XH2), Hepatitis C IRES (PDB ID: 2KTZ), reverse transcriptase initiation complexes (PDB ID: 6WB2), and many others.

The smallest pseudoknot-containing graph is 2_3 (Figure 2), and was found in various phylogenetic and functional contexts, such as in an Influenza A virus splicing site at the 3′ end (PDB ID: 7RQ5), a Luteoviral RNA ribosomal frameshifting element (PDB ID: 2A43), and a Poxvirus RNA in a transcription initiation complex (PDB ID: 7AOZ).

Of the 37 distinct graphs, 14 new motifs emerged with vertex counts ranging from 4 to 8, shown in Figure 5. Of these 14 new graphs, 8 appear as parent graphs, and 6 as subgraphs after partitioning (Graph ID: 4_3, 5_1, 6_125, 7_530, 7_747, 8_4987) (Figure 5). Two new graphs, 6_1 and 6_480, represent multiple non-contiguous RNA strands that form the backbone of a Satellite tobacco Mosaic virus capsid protein (33). While double-helical segments in nature link the RNA strands of the virus, it was modeled as disjoint RNA strands due to its association with a crystallographic asymmetric unit in the experimental data (33). Here, we do not distinguish between graphs and subgraphs because many PDB structures vary in length and are truncated from larger RNA structures at varying distances from functional sites.

Comparing Viral and General RNA Motifs using Fiedler Score Based Clustering

To assess the extent to which viral RNAs share features (submotifs) found in general RNA, we analyzed 196 known RNAs, excluding the single-hairpin structure 1_1. Here, s (slope)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and e (residual error) are summary features computed from the ordered, rescaled entries of each graph's Fiedler eigenvector (see Methods). Of these, 161 are general, and 35 are viral, giving a baseline viral RNA prevalence of 17.86% (35/196). We applied three two-cluster algorithms: K-means, k-medoids (PAM), and Ward's linkage.

For each method, we designated the cluster containing more viral IDs as viral (C_1) and the other as general (C_2). Hypothetical motifs were projected to the nearest centroid/mean for visualization only and were excluded from all statistical tests. We use a two-sided Fisher's exact test to assess whether cluster membership and viral label are associated overall. We also use a right-tailed Fisher's exact test to ask, for each cluster, whether its viral fraction exceeds the baseline viral prevalence of 17.86%. We also report silhouette scores of each cluster, which measure how well separated a cluster is, with values ranging from −1 (misclassified) to +1 (well separated) (28).

For K-means, cluster 1 contained 135 RNAs, including 29 viral RNAs (21.48%), with a one-sided Fisher $p = 0.0346$. Cluster 2 contained 61 RNAs, including 6 viral RNAs (9.84%), with *a p-value* of 0.988. The global two-sided Fisher test yielded $p = 0.0685$, with an odds ratio of 2.51 and a 95% confidence interval of [0.982, 6.404]. The mean silhouette was 0.584 ($C_1 = 0.756, C_2 = 0.202$).

For PAM, cluster 1 contained 96 RNAs, including 19 viral (19.79%), with $p = 0.306$. Cluster 2 contained 100 RNAs, including 16 viral (16.00%), with $p = 0.810$. The global two-sided Fisher test yielded $p = 0.5767$, with an odds ratio of 1.30 and a 95% confidence interval of [0.622, 2.697]. The mean silhouette was 0.452 ($C_1 = 0.127, C_2 = 0.763$).

For Ward's linkage ($K = 2$), cluster 1 contained 111 RNAs, with 19 viral (17.12%), and had $p = 0.692$. Cluster 2 contained 85 RNAs, including 16 viral (18.82%), with $p = 0.450$. The global two-sided Fisher test yielded $p = 0.851$, with an odds ratio of 0.891 and a 95% confidence interval of [0.427, 1.857]. The mean silhouette was 0.468 ($C_1 = 0.716, C_2 = 0.144$).

Across methods, we find no significant association between cluster assignment and viral label. Taken together, these analyses support the sharing of many common structural building blocks between viral and general RNAs.

Shared Motifs Across Viral and General RNAs

Consistent with the lack of separation after motif clustering, we observe many similar dual-graph motifs shared by viral and general RNAs, as shown in Figure 8. In several cases, the same dual graph appears across contexts. For example, graph 4_19 corresponds to cellular tRNA and is found in viral contexts, such as the bacteriophage pRNA (PDB ID: 6JXM) and the Coxsackie virus replication element (PDB ID: 8DP3).

Specific motif families, like hairpins and double hairpins with IDs 1_1 and 2_1, are prevalent in both sets; viral examples include HIV-1 TAR (PDB ID: 6XH2) and HCV IRES hairpins (PDB ID: 2KTZ). Multi-way junctions show a similar pattern. Viral IRES domains

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

are mapped to dual graphs 5_25 (HCV), 6_7 (HAV), and 9_38596 (CrPV). Together, these examples illustrate that a common translation-initiation role can be realized through distinct dual-graph topologies, including both pseudoknot-free and pseudoknot-containing motifs.

The junction-containing subgraph 4_10 occurs in a bacterial riboswitch aptamer and T. thermophilus tRNA and also in a viral RNA as a subgraph of the HAV IRES graph 6_7.

In addition, graph 6_236 (present in E. coli rRNA) lies near 9_38596 and is the closest non-viral graph to 9_38596 using our scoring model (See Methods). Exoribonuclease-resistant folds, such as those found in flaviviruses like Zika (5_45) and Tamana Bat Virus (5_67), exhibit junction-plus-stem layouts, and the same motif recurs in non-viruses, as reflected by 5_45 in the B. anthracis ribozyme (PDB ID: 3G9C).

Partitioning further reveals the reuse of motifs across RNAs. The CrPV IRES 9_38596 admits two alternative partitions, 7_747+3_3 and 8_4897+2_3, with 2_3 and 3_3 serving as common pseudoknot subgraphs in the general atlas (Fig. 2). The BMV tRNA-like domain 8_1333 partitions into 7_530 and the widely used double hairpin 2_1 (Fig. 8). These partitionings of viral RNA show that viruses utilize general RNA submotifs.

Across these examples, variation is primarily in local geometry, such as branch or stem lengths, rather than in the underlying dual-graph topology. Taken together, these qualitative comparisons align with the clustering result indicating that viral and general RNAs draw on a shared secondary-structure repertoire.

DISCUSSION

The development and expansion of the RNA-as-Graphs Motif Atlas, as well as the introduction of the dual-graph partitioning algorithm and Fiedler vector-derived features, have enabled analysis of the structural intricacies of RNA secondary structure. In this study, we described and analyzed existing viral RNA motifs using dual-graph representations that preserve pseudoknots and junctions.

We extended the RAG Motif Atlas (Figure 3) by adding 14 novel viral dual graphs (Figure 5) derived from 273 structures, bringing the library to 197 unique topological motifs. These new motifs span internal ribosome entry sites, tRNA-mimicking domains, exoribonuclease-resistant RNAs, frameshifting elements, reverse-transcription initiation elements, and packaging/core signals. The distribution of motifs is skewed toward simple hairpins (1_1, 2_1), with rarer junction and pseudoknot motifs, as shown in Figure 2 for the Cricket Paralysis Virus IRES (PDB: 4D5N).

Our clustering using Fiedler vector features with K-means, PAM, and Ward found a large intersection between viral and general RNA motifs. Thus, viral and general RNAs overlap in this feature space. Correspondingly, qualitative comparisons of recurrent motifs across different contexts, along with dual-graph partitioning, agree with this picture. Some viral structures break into submotifs that are already common in the atlas, and the same secondary structures recur often across viral and non-viral sources.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The atlas extension and clustering results suggest that new viral entries fit naturally within the existing dual graph universe. Motif families such as hairpins, double hairpins, and multi-way junctions are abundant in both viral and cellular contexts, and several viral cases use the same graph IDs seen in general RNAs. In some cases where IDs differ, partitioning reveals shared submotifs, and completely new viral graphs may be close in feature space to already existing graphs. This recurrence helps explain why the viral RNAs do not cluster separately.

As more RNA data are analyzed and cataloged as graphs, some of the 14 newly identified viral RNA motifs (Figure 5) may also be assigned to general RNA structures. This complexity not only highlights the challenges of predicting RNA function based solely on structural motifs but also opens new avenues for understanding how RNA diversity is leveraged across biological systems to fulfill a wide array of roles.

Dual graph partitioning highlights modular aspects of RNA structure by decomposing larger topologies into recurrent subgraphs while preserving junctions and pseudoknots. In this sense, the resulting subgraphs can be viewed as structural building blocks within larger RNAs. For example, in Figure 8, the 8_1333 graph of the BMV tRNA-mimicking domain partitions into subgraphs 7_530 and 2_1. Here, 7_530 contains the acceptor-stem-like pseudoknotted region, whereas 2_1 corresponds to the double-hairpin portion associated with the putative anticodon-stem analog. The larger viral motif can be understood in terms of smaller structural units with identifiable roles. More generally, partitioning helps place large viral RNA motifs within a broader repertoire of shared RNA substructures and recurrent modules. Likewise, RNA folding studies have shown that local secondary-structure rearrangements and dynamics shape the behavior of larger RNAs (39; 24; 45; 46).

Limitations include reliance on deposited structures in the Protein Data Bank, uneven coverage across viral families, and the use of two-cluster partitions chosen to assess separability rather than to discover multiple groups directly. Our analysis also focuses on secondary structure topology. It does not incorporate proteins, ion conditions, or long-range tertiary interactions that can shift function within a shared topology.

In addition, graph assignments can depend on the completeness of deposited structures and secondary-structure annotations, both of which can affect the resulting topology for some RNAs. More generally, dual graphs are flexible coarse-grained representations, and changes to base-pair annotation, minimum-stem definition, loop/bulge threshold, or helix-merging rules may alter the resulting topology.

In summary, this study contributes to the expanding field of RNA research and RNA structural determination by providing a mathematical analysis of RNA secondary structures as graphs for viral RNA. More generally, identifying recurrent RNA topologies and submotifs may be useful for RNA nanotechnology and for therapeutic efforts that rely on modular RNA structural elements. Examples include antiviral strategies based on targeted RNA mutations, in the same spirit as mutant designs for frameshifting elements of SARS-CoV-2, for which confirmatory experiments were also reported (19; 20; 43), HIV-1 (47), and Chikungunya (48).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the National Science Foundation Award DMS-2151777 and CHE-2330628 from the Division of Mathematical Sciences and the Division of Chemistry, the National Institutes of Health R35GM122562 Award from the National Institute of General Medical Sciences, and Philip Morris International to TS. Support from the Simons Foundation through the NYU Center for Computational Physical Chemistry Award MPS-T-MPS-00839534 to MET for the work of JZ and TS is also gratefully acknowledged.

REFERENCES

1. Cech Thomas R., Steitz Joan A. (2014). “The noncoding RNA revolution-trashing old rules to forge new ones.” Cell, 157(1), 77–94.
2. Zogg Heather, Singh Reeti, Ro Sun. (2022). “Current Advances in RNA Therapeutics for Human Diseases.” Int J Mol Sci, 23(5), 2736.
3. Zhu Ying, Zhu Li, Wang Xiaoyu, et al. (2022). “RNA-based therapeutics: an overview and prospectus.” Cell Death Dis, 13(644).
4. Kim Yong-Kyu. (2022). “RNA therapy: rich history, various applications and unlimited prospects.” Exp Mol Med, 54(4), 455–465.
5. Leontis Nicholas B., Stombaugh Jesse, Westhof Eric. (2002). “The non-Watson-Crick base pairs and their associated isostericity matrices.” Nucleic Acids Res, 30(16), 3497–531.
6. Wu Meiting, D’Souza V. (2020). “Alternate RNA Structures.” Cold Spring Harb Per-spect Biol, 12(1), a032425.
7. Gan HH, Perlow RA, Roy S, Ko J, Wu M, Huang J, Yan S, Nicoletta A, Vafai J, Sun D, Wang L, Noah JE, Pasquali S, Schlick T. (2002). “Analysis of Protein Sequence/Structure Similarity Relationships.” Biophysical Journal, 83(5), 2781–2791. ISSN 0006–3495.
8. Waterman Michael. (1978). “Secondary structure of single-stranded nucleic acids.” Adv Math Suppl Stud, 1, 167–212.
9. Le Shu, Nussinov Ruth, Maizel Jacob V. (1989). “Tree graphs of RNA secondary structures and their comparisons.” Comput Biomed Res, 22(5), 461–473.
10. Nussinov Ruth, Jacobson Ann B. (1980). “Fast algorithm for predicting the secondary structure of single-stranded RNA.” Proc Natl Acad Sci U S A, 77(11), 6309–6313.
11. Gan HH, Pasquali S, Schlick T. (2003). “Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design.” Nucleic Acids Res, 31(11), 2926–43.
12. Kim N, Zheng Z, Elmetwaly S, Schlick T. (2014). “RNA Graph Partitioning for the Discovery of RNA Modularity: A Novel Application of Graph Partition Algorithm to Biology.” PLoS ONE, 9(9), e106074.
13. Zhu Q, Schlick T. (2021). “A Fiedler Vector Scoring Approach for Novel RNA Motif Selection.” J Phys Chem B, 125(4), 1144–1155.
14. Chastain M, Tinoco I. (1991). “Structural elements in RNA.” Prog Nucleic Acid Res Mol Biol, 41, 131–177.
15. Jain S, Zhu Q, Paz ASP, Schlick T. (2020). “Identification of novel RNA design candidates by clustering the extended RNA-As-Graphs library.” Biochim Biophys Acta Gen Subj, 1864(6), 129534.
16. Yan S, Zhu Q, Jain S, et al. (2022). “Length-dependent motions of SARS-CoV-2 frameshifting RNA pseudoknot and alternative conformations suggest avenues for frameshifting suppression.” Nat Commun, 13, 4284.
17. Yan S, Zhu Q, Hohl J, Dong A, Schlick T. (2023). “Evolution of coronavirus frameshifting elements: Competing stem networks explain conservation and variability.” Proc Natl Acad Sci U S A, 120(20), e2221324120.
18. Reynolds A, Richards G, de la Iglesia B, Rayward-Smith V. (2006). “Clustering rules: a comparison of partitioning and hierarchical clustering algorithms.” Journal of Mathematical Modelling and Algorithms, 5, 475–504.
19. Schlick T, Zhu Q, Jain S, Yan S. (2021). “Structure-altering mutations of the SARS-CoV-2 frameshifting RNA element.” Biophys J, 120(6), 1040–1053.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

20. Schlick T, Zhu Q, Dey A, Jain S, Yan S, Laederach A. (2021). “To Knot or Not to Knot: Multiple Conformations of the SARS-CoV-2 Frameshifting RNA Element.” J. Am. Chem. Soc. 143(30):11404–11422.
21. Wang R, Schlick T. (2025). “How large is the universe of RNA-like motifs? A clustering analysis of RNA graph motifs using topological descriptors.” PLoS Comput. Biol. 21(7): e1013230.
22. Siddell SG, Smith DB, Adriaenssens E, et al. (2023). “Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV).” J Gen Virol, 104(5), 001840.
23. Poltronieri P, Sun B, Mallardo M. (2015). “RNA Viruses: RNA Roles in Pathogenesis, Coreplication and Viral Load.” Curr Genomics, 16(5), 327–335.
24. Zhu Q, Petingi L, Schlick T. (2022). “RNA-As-Graphs Motif Atlas-Dual Graph Library of RNA Modules and Viral Frameshifting-Element Applications.” Int J Mol Sci, 23(16), 9249.
25. Lloyd S. (1982). “Least squares quantization in PCM.” IEEE Trans. Inf. Theory, 28(2), 129–137.
26. Kaufman L, Rousseeuw PJ. (2009). “Finding Groups in Data: An Introduction to Cluster Analysis.” Wiley, Hoboken, NJ.
27. Ward JH. (1963). “Hierarchical grouping to optimize an objective function.” J. Am. Stat. Assoc, 58(301), 236–244.
28. Rousseeuw PJ. (1987). “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.” J. Comput. Appl. Math, 20, 53–65.
29. Lance GN, Williams WT. (1967). “A general theory of classificatory sorting strategies: 1. Hierarchical systems.” Comput. J, 9(4), 373–380.
30. Jain AK. (2010). “Data clustering: 50 years beyond K-means.” Pattern Recognit. Lett, 31(8), 651–666.
31. Research Collaboratory for Structural Bioinformatics Protein Database. https://www.rcsb.org.
32. Brierley I, Pennell S, Gilbert RJ. (2007). “Viral RNA pseudoknots: versatile motifs in gene expression and replication.” Nat Rev Microbiol, 5(8), 598–610.
33. Larson SB, Day J, Greenwood A, McPherson A. (1998). “Refined structure of satellite tobacco mosaic virus at 1.8 A resolution.” J Mol Biol, 277(1), 37–59.
34. Svoboda P, Cara AD. (2006). “Hairpin RNA: a secondary structure of primary importance.” Cell Mol Life Sci, 63, 901–908.
35. Sorokin II, Vassilenko KS, Terenin IM, et al. (2021). “Non-Canonical Translation Initiation Mechanisms Employed by Eukaryotic Viral mRNAs.” Biochemistry (Mosc), 86(9), 1060–1094.
36. Berg MD, Brandl CJ. (2021). “Transfer RNAs: diversity in form and function.” RNA Biol, 18(3), 316–339.
37. Akiyama BM, et al. (2016). “Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease.” Science, 354, 1148–1152.
38. Brown JD, et al. (2020). “Structural basis for transcriptional start site control of HIV-1 RNA fate.” Science, 368, 413–417.
39. Jain S, Bayrak C, Petingi L, Schlick T. (2018). “Dual graph partitioning highlights a small group of pseudoknot-containing RNA submotifs.” Genes, 9(371).
40. Lu X, Olson W. (2003). “3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.” Nucleic Acids Res, 31, 5108–5121.
41. Mathews DH. dot2ct [Software]. Department of Biochemistry & Biophysics, University of Rochester Medical Center. Retrieved from https://rna.urmc.rochester.edu/index.html.
42. Merris R. (1997). “Laplacian Graph Eigenvectors.” Department of Mathematics and Computer Science, California State University
43. Dey A, Yan S, Schlick T, Laederach A. (2024). “Abolished frameshifting for predicted structure-stabilizing SARS-CoV-2 mutants: implications to alternative conformations and their statistical structural analyses.” RNA. 2024 Oct 16;30(11):1437–1450. doi: 10.1261/rna.080035.124.
44. Fiedler M. (1973). “Algebraic connectivity of graphs.” Czechoslovak Math. J, 23, 298–305.
45. Xue Y, Gracia B, Herschlag D, Russell R, Al-Hashimi HM. (2016). “Visualizing the formation of an RNA folding intermediate through a fast highly modular secondary structure switch.” Nature Communications, 7, 11768. doi:10.1038/ncomms11768

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

46. Bose R, Saleem I, Mustoe AM. (2024). “Causes, functions, and therapeutic possibilities of RNA secondary structure ensembles and alternative states.” Cell Chemical Biology, 31(1), 17–36. doi:10.1016/j.chembiol.2023.12.010

47. Newton K, Yan S, Schlick T. (2025). “Conformational Analysis and Structure-Altering Mutations of the HIV-1 Frameshifting Element.” International Journal of Molecular Sciences, 26(13), 6297. doi: 10.3390/ijms26136297.

48. Lee S, Schlick T. (2026). “Kinetic Traps in RNA Folding: Targeted Design of Frameshifting Element Mutants by Thermodynamic and Kinetic Analysis of the Chikun-gunya Virus.” Journal of Physical Chemistry B. doi: 10.1021/acs.jpcb.5c08223.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

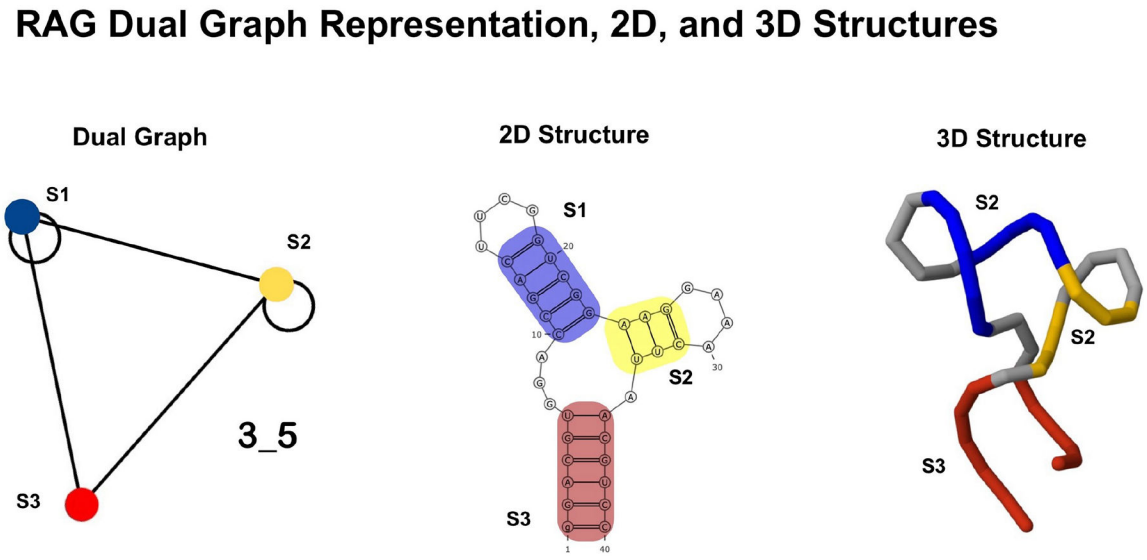


Figure 1:

A three-way junction-containing RNA from dengue fever virus RNA (PDB ID: 7UMD), shown in 3D, 2D, and dual-graph representation (Graph ID 3_5). The three stems are colored and labeled across all panels, and gray regions denote loops that match each stem in the RNA substructure to its corresponding vertex in the dual graph.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

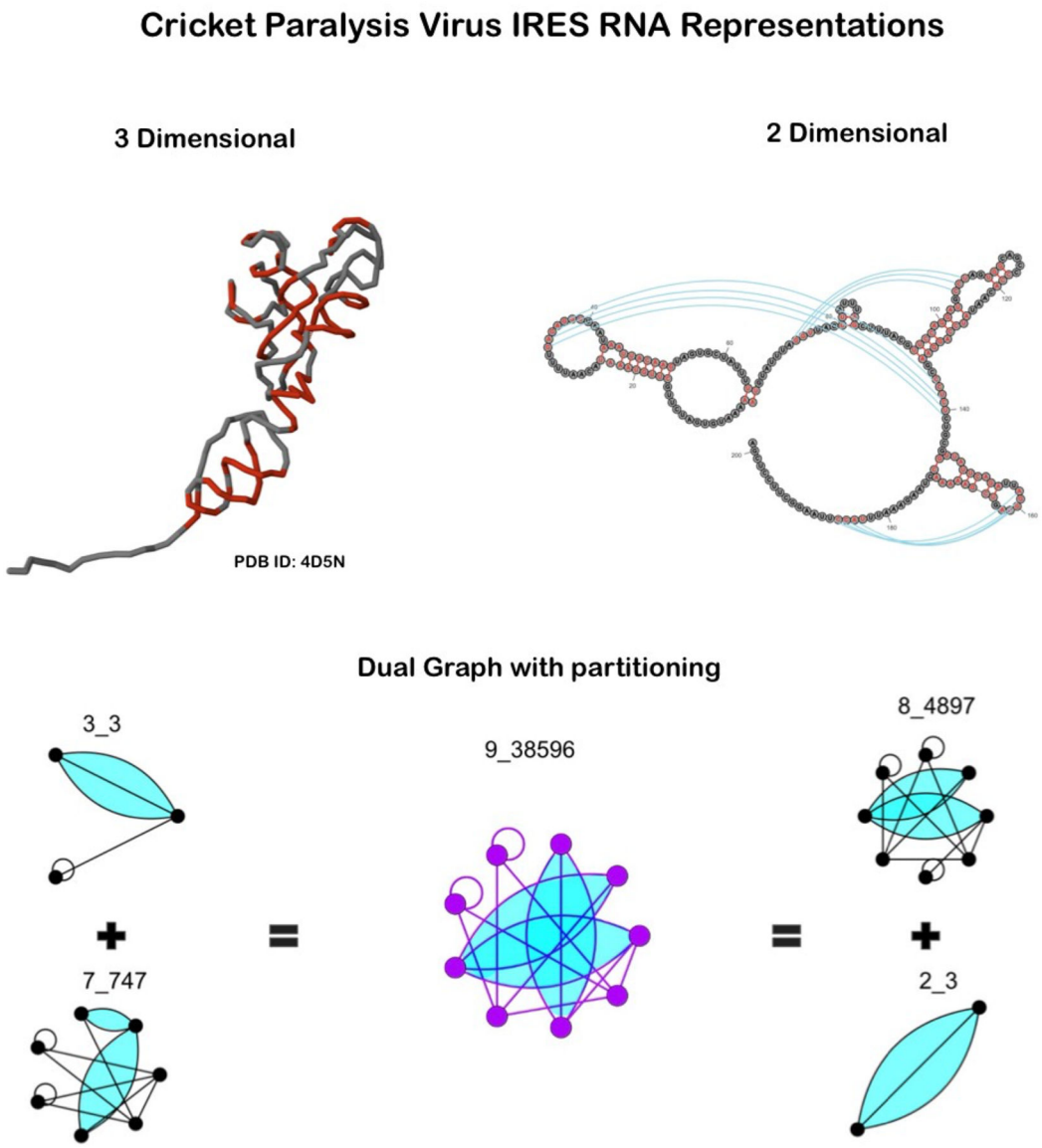


Figure 2:

3D, 2D, and dual graph 9_38596 representations of the Cricket Paralysis Virus IRES (PDB ID: 4D5N) and its partitioning. In the 3D and 2D structures, red nucleotides indicate base pairs forming a stem. Two different partitionings of dual graph 9_38596 are shown, with intact pseudoknots highlighted in light blue. The first partitioning yields graphs 7_747 and 3_3, while the second yields graphs 8_4897 and 2_3.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

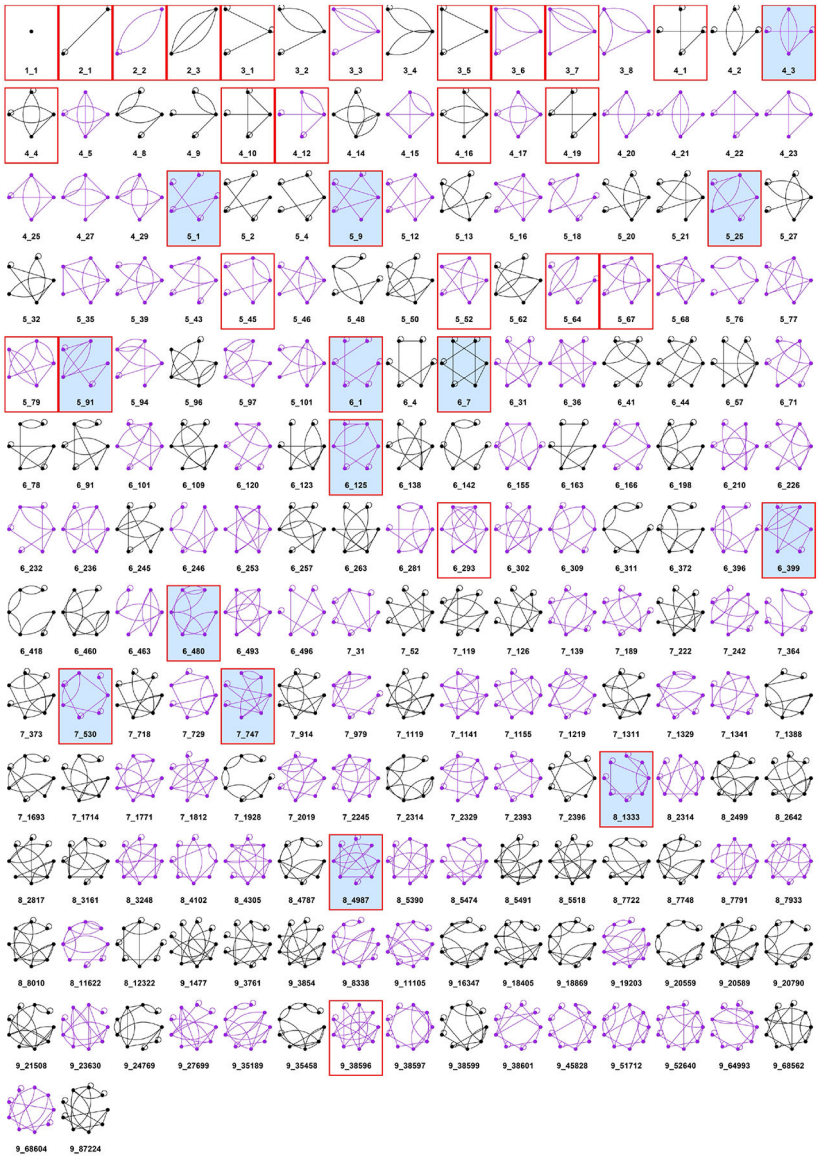


Figure 3:

The assembled library of 197 dual graphs in the updated RAG Atlas, comprising viral and general RNAs. Graphs containing at least one pseudoknot are highlighted in purple. All 37 distinct viral dual graphs identified in our dataset are boxed in red, while newly discovered viral graphs that were previously unlisted (non-existing) are highlighted in blue.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

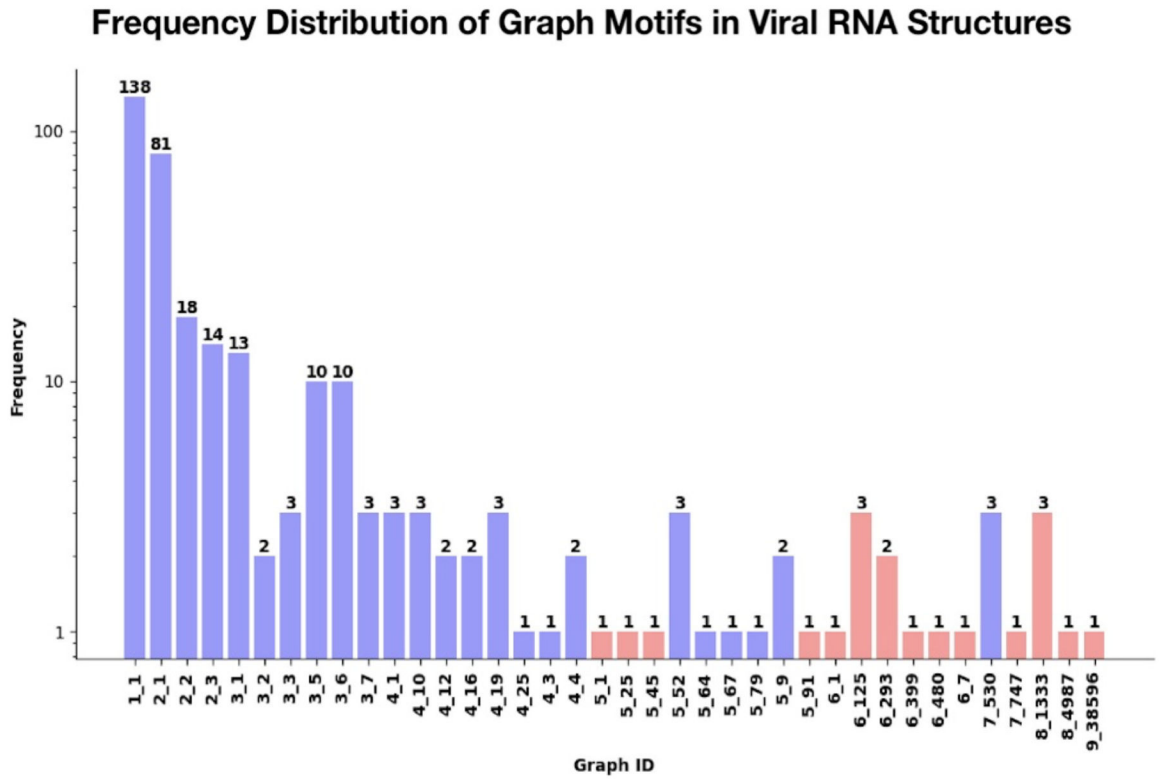


Figure 4:

Composition of dual graph motifs across all 339 structures extracted from viral RNAs in the PDB. The height of each bar indicates the frequency of the corresponding motif in the viral graph library. Newly discovered graphs are highlighted in red. Frequency is described on a logarithmic scale.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

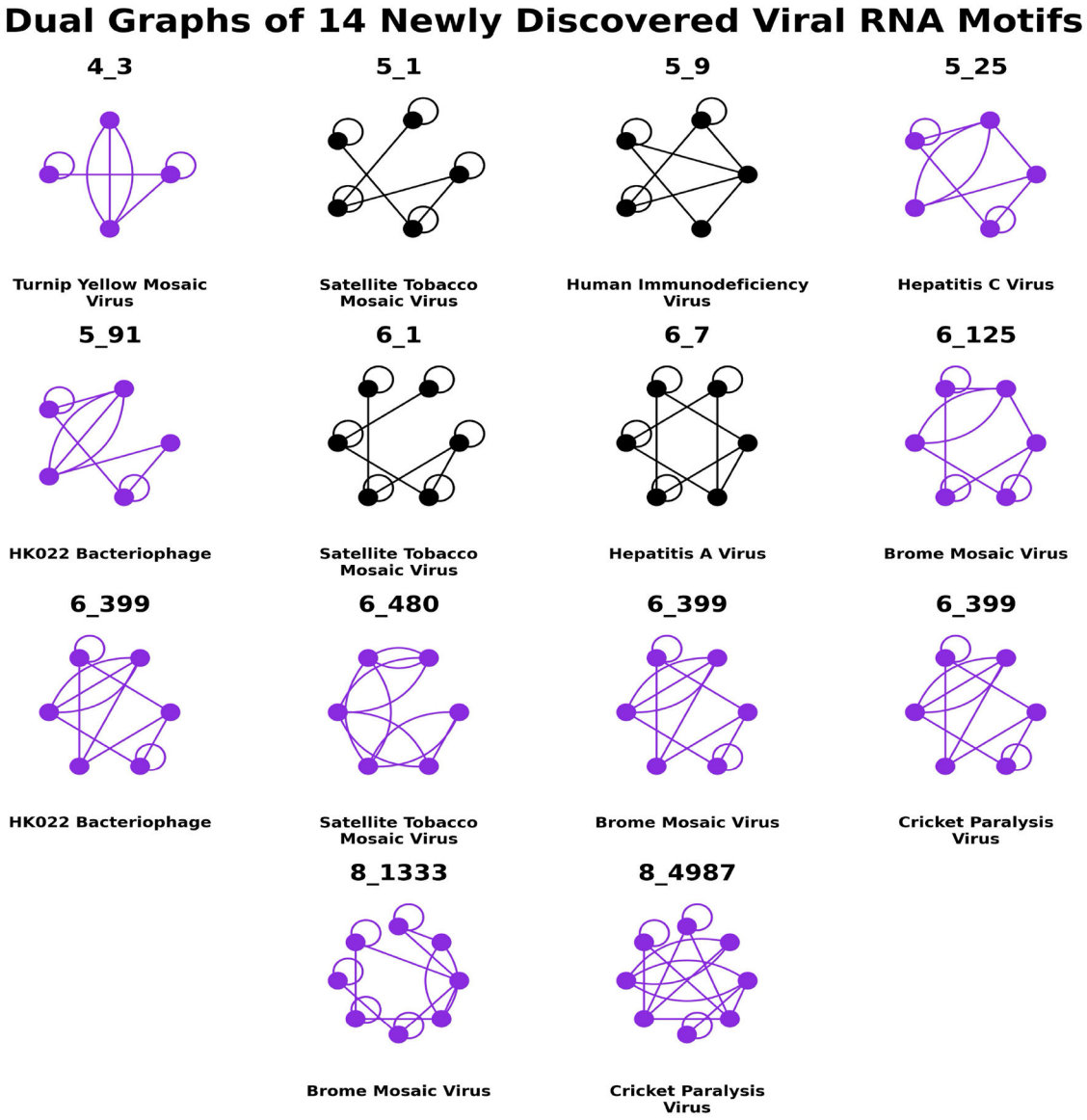


Figure 5:

Dual graphs of 14 newly discovered viral motifs that were not listed in the previous RAG atlas. Graphs containing pseudoknots are highlighted in purple.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

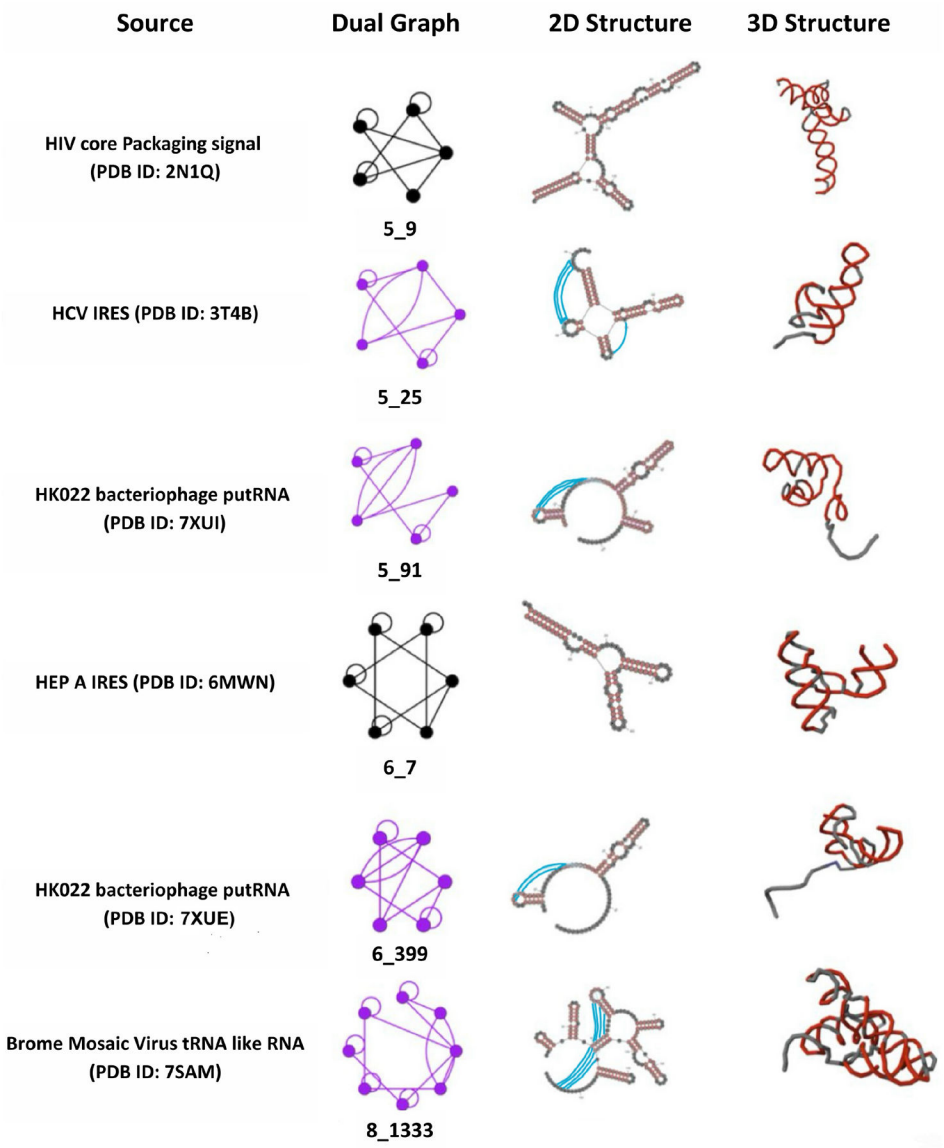


Figure 6:

Some new motifs of viral RNA elements from Figure 5, with their dual-graph, 2D, and 3D structures. Viral RNAs analyzed include trans-activation response elements, exonuclease-resistant domains at the 5′ and 3′ ends, reverse transcriptase binding domains, RNA polymerase interaction domain riboswitch elements, frameshifting elements, internal ribosome entry sites, core packaging signals, and self-cleaving ribozymes. The two HK022 putRNA examples correspond to distinct related deposited structures from the same study (PDB IDs 7XUE and 7XUI). Red nucleotides indicate base pairs forming stems.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

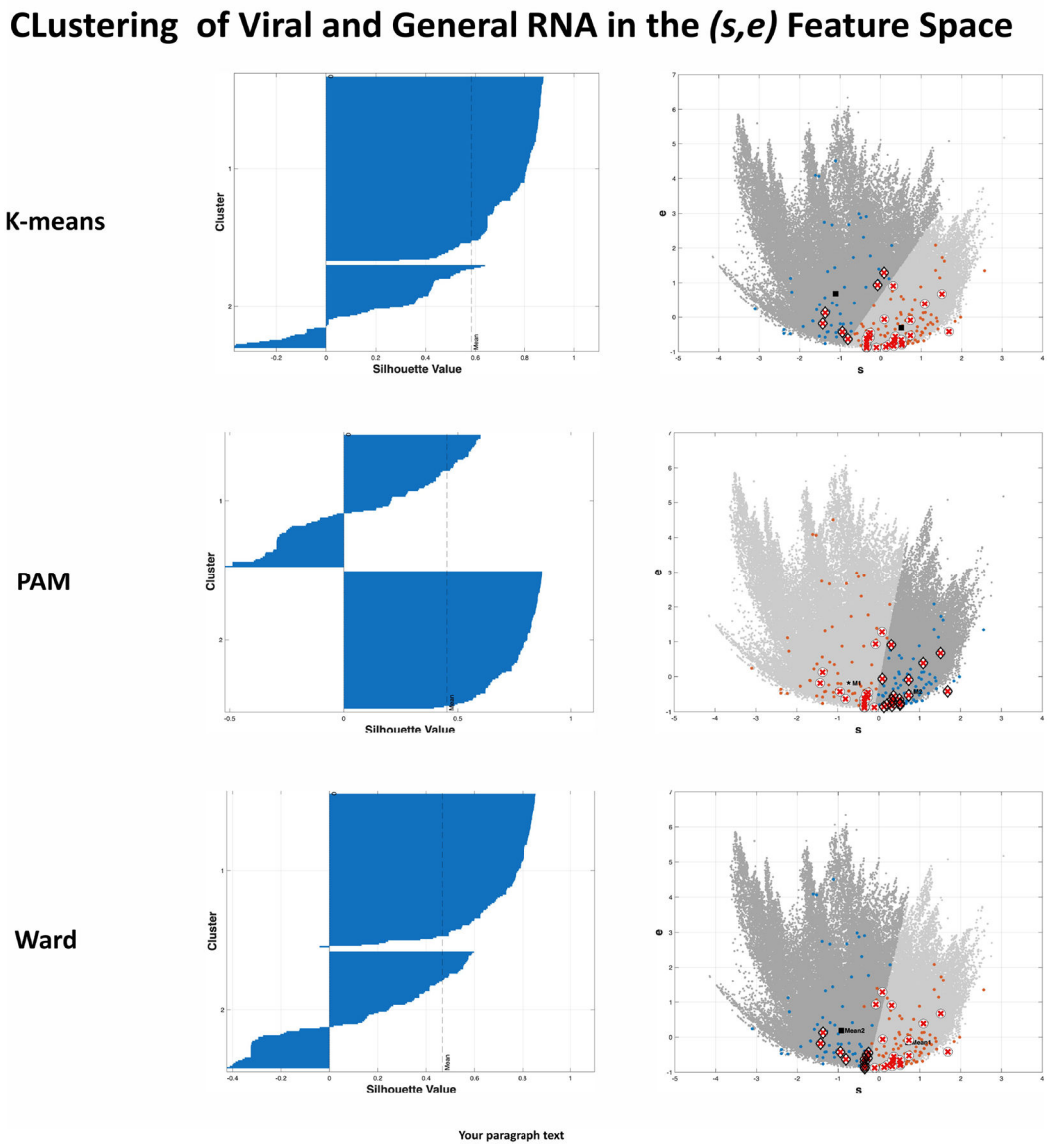


Figure 7:

Two cluster partitions of known RNAs in the (s, e) feature space (C_1 is the cluster with more viral IDs), with corresponding silhouette plots (bottom). Cluster centers are shown as black squares. General RNAs assigned to C_1 are shown as red circles, and those assigned to C_2 are shown as blue circles. Viral RNAs assigned to C_1 are red crosses, and those assigned to C_2 are black diamonds overlaid with black crosses. Hypothetical RNAs are projected to the nearest cluster centers in light or dark gray. Silhouette plots summarize overall and per–cluster cohesion for each method.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

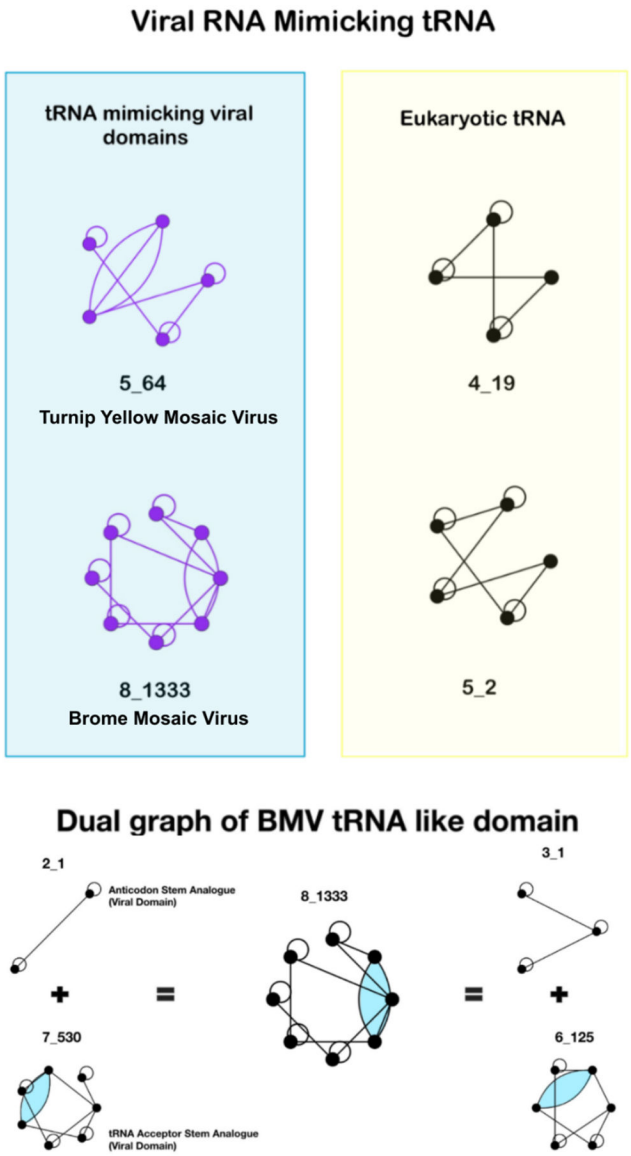


Figure 8:

Comparison of viral tRNA-mimicking domains and eukaryotic tRNA structures (top) and partitioning of the BMV tRNA-like domain (bottom). The top-left column shows viral tRNA-like motifs, with graph 5_64 representing the tRNA-like structure from Turnip Yellow Mosaic Virus (PDB ID: 4P5J) and graph 8_1333 representing the Brome Mosaic Virus tRNA-like domain (PDB ID: 7SAM). The right column displays eukaryotic tRNA motifs, such as 4_19, which has also been found in bacteriophage RNA. Graph 5_2 is also included, as it featured in our earlier update as representing a eukaryotic tRNA motif and has not yet been observed among viral graphs.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1:

Clustering of known RNAs in the (s, e) space. C_1 is the cluster with the most viral IDs.

Method	Mean sil.	C_1 sil.	C_2 sil.	C_1 viral frac	C_2 viral frac	Fisher p	OR [95% CI]
K-means	0.584	0.756	0.202	21.48%	9.84%	0.0685	2.51 [0.982, 6.404]
PAM	0.452	0.127	0.763	19.79%	16.00%	0.5767	1.30 [0.622, 2.697]
Ward	0.468	0.716	0.144	17.12%	18.82%	0.8510	0.891 [0.427, 1.857]

RNA Nanomed. Author manuscript; available in PMC 2026 November 07.