Viral RNA as a branched polymer

Domen Vaupotič, Angelo Rosa, Rudolf Podgornik, Luca Tubiana, and Anže Božič

Abstract Myriad viruses use positive-strand RNA molecules as their genomes. Far from being only a repository of genetic material, viral RNA performs numerous other functions mediated by its physical structure and chemical properties. In this chapter, we focus on its structure and discuss how long RNA molecules can be treated as branched polymers through planar graphs. We describe the major results that can be obtained by this approach, in particular the observation that viral RNA genomes have a characteristic compactness that sets them aside from similar random RNAs. We also discuss how different parameters used in the current RNA folding software influence the resulting structures and how they can be related to experimentally observable quantities. Finally, we show how the connection to branched polymers can be extended to take advantage of known results from polymer physics and can be further moulded to include additional interactions, such as excluded volume or electrostatics.

Key words: +ssRNA viruses; RNA secondary structure; branched polymers; scaling exponents; graph theory

---

Domen Vaupotič
Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia

Angelo Rosa
Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy

Rudolf Podgornik
School of Physical Sciences and Kavli Institute of Theoretical Science, University of Chinese Academy of Sciences, Beijing, China

Luca Tubiana
Physics Department, University of Trento, Trento, Italy
INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, Trento, Italy

Anže Božič
Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia
e-mail: anze.bozic@ijs.si

1 Introduction

RNA is an incredibly versatile biological macromolecule: not only does it act as a messenger between the DNA genome and the protein product, but it also assumes various roles in the form of transfer RNA, ribosomal RNA, microRNA, guide RNA, and long non-coding RNA, to name just a few [1,2]. Its function is carried out not only on the level of its primary sequence of nucleotides, but also by the local and global structures that are formed when the constituent nucleotides form base pairs with each other [3, 4]. Many RNA structures are thus involved in translational control, RNA localization, gene regulation, RNA stability, and more [5]. RNA structure folding is hierarchical, with the formation of base pairs—described by *secondary* structure—dominating the contribution to the folding energy and leading into its embedding in three-dimensional space, described by *tertiary* structure [6,7]. In spite of recent improvements in the prediction of the tertiary structure of RNA molecules, it remains restricted to relatively short, individual sequences [8–10]. It is therefore of great advantage that RNA structure and its function can often be understood well by modelling it on the level of secondary structure, which can be further complemented by experimental methods such as SHAPE and its derivations [11–13].

In a large number of bacterial, plant, animal, and human *viruses*, positive-strand RNA (+ssRNA) takes on the role of their genomes [14]. Far from simply coding for the protein products, both local structural elements as well as long-range structural interactions in the genomes of +ssRNA viruses are involved in many fundamental viral processes such as virus disassembly, translation, genome replication, and packaging, and are thus in general important for viral fitness [15–18]. In particular, the genomes self-assemble together with capsid proteins to form a functional virion in an interplay of RNA sequence, length, and structure, further influenced by environmental variables such as pH and salt concentration [19–22]. For instance, in certain viruses, local structural elements called packaging signals—typically one or several hairpin loops with a more or less defined structure and nucleotide pattern—are responsible for specific interactions with the capsid proteins, initiating assembly through several possible pathways [22–24].

At the same time, non-specific electrostatic interactions between highly negatively charged RNA and positively charged domains of capsid proteins dominate the self-assembly of many +ssRNA viruses [25, 26]. Here, a number of experiments have demonstrated that viral capsids can assemble not only with their native RNA genomes but also with non-cognate RNA genomes of other viruses, other RNA molecules, and even linear polyelectrolytes [27, 28]. Success of the self-assembly and the resulting capsid(-like) structure, however, both depend on the *length and structure of the cargo* as well as on environmental variables [29–31]. Varying the salt concentration of the solution, for instance, changes the strength of RNA-protein interaction [32], and varying the strength of the interaction between RNA and an adsorbing substrate can change the latter's preference for adsorbing either single- or double-stranded RNA [33].

Branching structure of viral RNA, in particular, plays an important role in RNA-capsid interaction and virus assembly. Experiments have demonstrated that RNA

structure and topology influence both packaging efficiency and the resulting capsid size and shape [29, 30, 34], while theoretical studies have shown that the degree of branching can greatly increase the amount of RNA that can be packaged into a capsid [35, 36]. Moreover, branching patterns of different RNAs have been shown to influence their size [37, 38], with genomes of +ssRNA viruses with icosahedral capsids being significantly more compact compared to those with helical capsids [39, 40]—with the former capsid type providing more severe spatial restrictions than the latter. This characteristic compactness appears to be a global structural property, and while even ~ 5% of *synonymous* mutations were shown to destroy it [40], the question remains of where in the genome sequence these topological and structural properties are encoded [41]. Understanding the topological properties of the genomes of +ssRNA viruses is thus essential to understand their ability to self-assemble and consequently to design strategies to modify or interfere with their function [42].

In this chapter, we describe how the secondary structure of viral RNA can be mapped to a branched polymer, which properties can be extracted, what are some of the major results that can be obtained using this approach, and some pitfalls to be considered. To this purpose, we first introduce the main properties of branched polymers and demonstrate how RNA can be treated as one by being mapped onto a graph. We then describe some of the topological and structural properties that can be gleaned from this approach. Next, we illustrate this approach on random RNA sequences of different length and nucleotide composition, which provides a baseline for comparison of different biological RNAs. Focusing on the genomes of +ssRNA viruses, we explore the differences among them by comparing them to random RNAs as well as shuffled versions of themselves. We also show how model parameters used in the prediction of RNA secondary structure—specifically, multiloop energy and maximum base pair span—influence these predictions. Lastly, we briefly overview the field-theoretical description of RNA as a branched polymer, which makes use of the derived topological parameters and allows for a self-consistent inclusion of additional short- and long-range interactions in the analysis of interactions between the RNA genome and the capsid proteins.

2 RNA as a branched polymer

2.1 Secondary structure of RNA as a graph

Secondary structure prediction

Description of RNA structure on the intermediate level of its secondary structure forms a conceptually important step and explains the dominant part of the free energy of structure formation [43]. Modelling RNA on this level allows for analysis of large numbers of very long RNA sequences—which would be prohibitively expensive to model on the level of their tertiary structure—while retaining the majority of the pertinent information about its local and global structure resulting from base-pairing.

Numerous software packages exist for the prediction of RNA secondary structure, the most popular ones including ViennaRNA [44] and RNAstructure [45], based on energy models of base-pairing, and CONTRAfold [46] and EternaFold [47], learning model parameters using stochastic context-free grammar. All of these algorithms necessarily come with limitations [47–49], but due to the complexity of structure prediction for long RNA molecules, they remain the tool of choice for studies of +ssRNA viral genomes, which can range anywhere from ~ 1000–30000 nt in length (Sec. 3). While some of the uncertainty in the prediction of RNA secondary structure can be alleviated by taking into account experimental data [12], such data is not widely available for most viral genomes.

Since the energy landscape of RNA structures is very shallow, predicting only the minimum free energy structure is typically insufficient, as the RNA can sample different conformations and several functional structures can co-exist in vivo [50]. The benefit of using energy-based folding algorithms for the prediction of secondary structure is that they enable generation and sampling of *thermal ensembles* of representative structures at a given temperature [51]. In the examples presented in this chapter, we use ViennaRNA v2.4 [44] to predict thermal ensembles of 500 structures at $T = 37°$ C for each RNA sequence and denote any quantity O averaged over this thermal ensemble of structures by $\langle O \rangle$. As we show later on, this sample size produces sufficient statistics for each quantity we consider.

RNA as a graph

The idea that the complexity of base pairs and sequence-structure patterns in a folded RNA sequence can be reduced by mapping its secondary structure onto a graph is not new (Ref. [52] provides a detailed overview of the topic). In the absence of pseudoknots—a typical, and usually justified simplification [53]—the secondary structure can be described as a planar tree (Fig. 1). The simplest way to construct such a tree is by mapping double-stranded regions (base pairs) to *edges* with weights corresponding to the stem lengths, while single-stranded regions (unpaired nucleotides) are mapped to *nodes* connecting the edges.[1] In the rest of the chapter,

Fig. 1 Mapping (a part of) RNA secondary structure onto a planar tree. Double-stranded (base-paired) regions are mapped to graph edges, weighted with the stem length, while single-stranded regions are mapped to graph nodes.

---
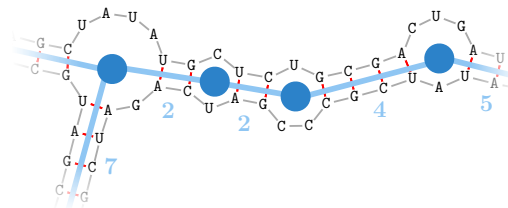
[1] Further modifications are possible, such as weighing the nodes with the size of single-stranded regions or excluding certain structures such as single-nucleotide bulges [54], but can simultaneously lead to peculiar behaviour, even resulting in disconnected graphs.

when we will refer to RNA trees, we will have in mind this procedure. This mapping is independent of the base-pairing model used to predict the structure (apart from the assumption of the absence of pseudoknots), and we discuss some differences that arise from using different model parameters in Sec. 4.

Once RNA secondary structure is mapped onto a tree composed of $N + 1$ nodes $v_i \in \mathbb{V}$, $i = 0, \ldots, N$, connecting N (undirected) edges $e_i = (v_j, v_k) \in \mathbb{E}$ with weights (stem lengths) b_i, it is possible to derive various parameters describing its topology and structure [55–57], including:

- The distribution of ladder distances $p(\ell)$, where the ladder distance $\ell(v_i, v_j)$ is defined as the shortest path between a pair of nodes v_i and v_j. The most important derived measures are the maximum ladder distance (MLD),

$$\text{MLD} = \max_{v_i, v_j \in \mathbb{V}} \ell(v_i, v_j), \tag{1}$$

corresponding to the diameter of the graph, and the average ladder distance (ALD),

$$\text{ALD} = \frac{1}{(N + 1)N} \sum_{v_i \neq v_j \in \mathbb{V}} \ell(v_i, v_j), \tag{2}$$

and its related quantity, the Wiener index $W = \sum_{v_i \neq v_j \in \mathbb{V}} \ell(v_i, v_j)$.
- The distribution of branch weights $p(N_{\text{br}})$, obtained by cutting the expanded tree at each edge and taking the smaller of the two total weights of the resulting trees.
- The distribution of node degrees $p(d_i)$, indicating the presence of multiloops (nodes of degree $d_i \geqslant 3$), with the total number of nodes with degree k given by D_k. Some derived quantities are, e.g., Zagreb indices $M_1 = \sum_{v_i \in \mathbb{V}} d_i^2$ and $M_2 = \sum_{(v_i, v_j) \in \mathbb{E}} d_i d_j$.
- The Laplacian spectrum, the eigenvalues λ_i of the Laplacian matrix $\mathsf{L} = \mathsf{D} - \mathsf{A}$, where D is the matrix of node degrees and A is the node adjacency matrix. The second smallest eigenvalue $\lambda_2 \leqslant 1$ describes the connectivity of the graph, with larger values indicating better connectivity or a more star-like structure.

Fig. 2 Illustration of some topological quantities described in the main text: maximum (MLD) and average ladder distance (ALD), second Laplacian eigenvalue λ_2, and Zagreb indices M_1 and M_2. The values of these quantities are shown for three different types of polymers—linear, branched, and star polymer—all of them having the same total number of monomers N and unit edge weight.

	linear	branched	star
MLD	19	5	2
ALD	7	2.91	1.90
λ_2	0.02	0.16	1
M_1	74	130	380
M_2	72	179	361

These quantities have the ability to distinguish, to various extents, between polymers with different types of tree topology, as illustrated in Fig. 2. Figure 3 further illustrates some of these quantities on an example of a (uniformly) random RNA sequence, N_{nt} = 2700 nt in length. From the distribution of node degrees (panel (b)), one can for instance determine the Zagreb indices of the RNA tree, and from the distribution of ladder distances (panel (e)), one can determine both the MLD and the ALD. Panels (f) and (g) further show thermal ensemble distributions of MLD and the total number of base pairs B, demonstrating that their averages are well-defined quantities. As we will see in the following, combining different topological properties of RNA graphs with statistical mechanics of branched polymers can be used to gain insight into their physical properties.
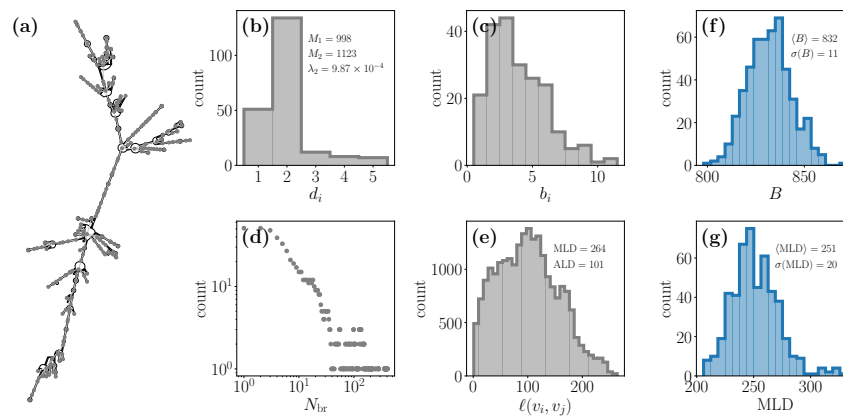
Fig. 3 Topological properties of RNA as a graph. **(a)** Representative secondary structure of a random RNA (N_{nt} = 2700 nt), overlaid with its graph representation. **(b)–(e)** Distributions of node degrees d_i, edge weights b_i, branch weights N_{br}, and path lengths $\ell(v_i, v_j)$ for the RNA structure shown in panel (a). Also shown are the values of M_1 and M_2, λ_2, MLD, and ALD. **(f)–(g)** Distributions of the total number of base pairs B and of MLD for 500 structures of the same RNA sequence drawn from a thermal ensemble. Also shown are the values of the thermal averages and their standard deviation.

2.2 Properties of branched polymers

Statistical mechanics of polymers is a very powerful theoretical tool with wide applications across biophysics, including, for instance, the scaling laws in the large-scale eukaryotic chromosome organization [58] and liquid-liquid phase separation in cells [59]. An important conceptual peculiarity of systems of *branched* polymers (trees)—such as RNA—concerns the necessity to distinguish between *annealed* (or, randomly *branching*) [60] and *quenched* (or, randomly *branched*) polymers [61].

Quenched trees are those whose topology of branches (tree connectivity) is fixed during the process of chemical synthesis and does not change afterwards. In contrast, the topology of annealed trees is not fixed at synthesis but instead can vary, typically in response to interactions (e.g., in the case of viral RNA, with capsid proteins) and/or changes in external conditions, and may fluctuate due to thermal motion. The class of annealed trees is particularly important as it is likely the most relevant for RNA molecules [62].

Scaling exponents

A physical description of polymer conformations has to be formulated by adopting the probabilistic language of statistical mechanics. In particular, since the total number of accessible conformations of a polymer chain increases exponentially with the number N of bonds (edges) [63], polymers are best described in terms of averages of corresponding observables.

The most distinct feature of a polymer conformation is its linear size, which can be expressed in terms of the radius of gyration, defined as

$$R_g^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{r}_i - \boldsymbol{r}_{\mathrm{cm}})^2, \tag{3}$$

where $\boldsymbol{r}_i$ is the spatial coordinate of the i-th monomer and $\boldsymbol{r}_{\mathrm{cm}} \equiv N^{-1} \sum_{i=1}^{N} \boldsymbol{r}_i$ is the centre-of-mass of the chain. The characteristic mean polymer size (i.e., its *mean gyration radius*) is given by the square root of the statistical average of Eq. (3) over the ensemble of all accessible conformations,

$$\langle R_g(N) \rangle \equiv \sqrt{\langle R_g^2 \rangle} \approx bN^{\nu}. \tag{4}$$

The quantity b is the mean bond length, while the *scaling exponent ν* is a fundamental parameter which—as we will shortly see—depends on several factors, particularly on monomer-monomer interactions [64]. While many fundamental works in polymer physics have dealt with determining the exponent ν for various polymer ensembles, exact values are known only for a very few cases [62]. In most—and often the most relevant—cases, though, only approximate (albeit accurate) results can be obtained by computationally extensive numerical methods or sophisticated mathematical tools [62]. Values of ν for polymer ensembles most relevant in the context of RNA are shown in Table 1; for other contexts, see the review by Everaers et al. [62].

While the exponent ν is sufficient to understand the physical properties of *linear* polymers (see Fig. 2), to completely understand an ensemble of *branching* polymers, such as viral RNAs, it is also necessary to characterize the topology of branching (or, equivalently, the *tree connectivity*) [62]. This is a particularly central problem for RNA secondary structure, since its mean gyration radius (Eq. (4)) and hence the exponent ν are not easily accessible.

Table 1 Best known values for scaling exponents of common polymer models in three dimensions. Here, the "ideal" and "self-avoiding" refer to either a complete neglect or inclusion of excluded-volume effects, respectively (cf. Flory theory in Sec. 2.2). The values shown as fractions are exact, while others are approximate (obtained either from numerical simulations or by analytical methods). For linear polymers, we have trivially $\rho = \varepsilon = 1$.

Polymer model	ν	ρ	ε	ν_{Flory}	ρ_{Flory}	Refs.
Ideal linear	1/2	1	1	1/2	1	[65]
Self-avoiding linear	0.5877	1	1	3/5	1	[66]
Ideal branching	1/4	1/2	1/2	1/4	1/2	[65]
Self-avoiding branching	1/2	0.654	0.651	7/13	9/13	[67, 68]

The problem of characterizing the connectivity of various ensembles of branching polymers has been theoretically addressed numerous times [68–70]. In the context of RNA, it is useful to introduce as a proper measure of chain connectivity the ensemble average of either the MLD or the ALD (Eqs. (1) and (2)) as a function of the number of monomers N,

$$\langle \text{MLD}(N) \rangle \sim \langle \text{ALD}(N) \rangle \sim b N^{\rho}, \tag{5}$$

both of which account for the average length of linear paths on the tree. While originally defined for characterizing the connectivity and introduced independently from ν, the exponent ρ in Eq. (5) is related to it, and consequently also provides a fundamental insight into RNA folding *in physical space*.

Last but not least, we can also consider the average branch weight [68]:

$$\langle N_{\text{br}}(N) \rangle \sim N^{\varepsilon}, \tag{6}$$

which is defined as the average weight of the smallest of the two sub-trees obtained by systematically removing—one at time—the edges connecting two neighbouring nodes of the original tree [69]. Note that while the two scaling exponents ρ and ϵ describe very different quantities, they are not independent from each other. In fact, the relation

$$\varepsilon = \rho \tag{7}$$

is expected to hold in general [68]. Equation (7) is particularly appealing because it can be used to validate *a posteriori* the initial hypothesis that RNA behaves as a randomly branching polymer: In fact, it is "sufficient" to measure $\langle \text{ALD} \rangle$ and $\langle N_{\text{br}} \rangle$ as a function of N and compare the estimates for the corresponding scaling exponents. Table 1 again summarizes the known values of ρ and ε for selected polymer ensembles.

Even the simplest theory of branching polymers thus has to deal with the three distinct observables introduced in Eqs. (4)–(6). Since we are primarily interested in the secondary structure of random and viral RNAs, we will focus on topological observables such as $\langle \text{MLD} \rangle$ and $\langle N_{\text{br}} \rangle$. Nonetheless, we will show that this has important consequences for how RNA molecules fold in space, i.e., on the average molecular size as given by $\langle R_g \rangle$.

Flory theory

Exact values for the scaling exponents ν and ρ are known only in few special polymer cases. In this respect, Flory theories of polymers [62, 64, 71] provide a simple framework for first—and yet remarkably accurate—approximations of both ν and ρ. Flory theory is formulated in terms of a balance between *(i)* an entropic (elastic) term, given by a sum of two contributions coming from the classical entropy of swelling (F_{sw}) and the entropy of reconfiguration of the tree architecture (F_{tree}) due to swelling and interaction, and *(ii)* an interaction term (F_{inter}) arising from monomer-monomer collisions. Taken together, the Flory free energy (in units of $\beta^{-1} = k_B T$) reads [62, 64, 71]:

$$F = F_{\text{sw}}(N, \langle R_g \rangle, \langle \text{ALD} \rangle) + F_{\text{tree}}(N, \langle \text{ALD} \rangle) + F_{\text{inter}}(N, \langle R_g \rangle)$$
$$= \frac{\langle R_g \rangle^2}{\langle \text{ALD} \rangle b} + \frac{\langle \text{ALD} \rangle^2}{Nb^2} + \upsilon_2 \frac{N^2}{\langle R_g \rangle^3}, \tag{8}$$

where $\upsilon_2 \sim b^3$ is on the order of the *second virial coefficient* [62], accounting for the excluded-volume interaction between any two monomers. Although physically appealing, this representation of the free energy is itself an approximation, since the terms in the free energy are not independent from one another. Nonetheless, Flory theories turn out to be quite accurate [62, 64, 71].

A key feature of the free energy in Eq. (8) is that the interaction term does not depend on $\langle \text{ALD} \rangle$, which likely remains valid even in other ensembles with different forms of interaction energy F_{inter} [62]. Consequently, we can balance the first two terms without worrying about the third, and connect $\langle R_g \rangle$, $\langle \text{ALD} \rangle$ and N:

$$\langle \text{ALD} \rangle \sim b^{1/3} N^{1/3} \langle R_g \rangle^{2/3}, \tag{9}$$

or, conversely (see Eqs. (4) and (5)),

$$\rho = \frac{1 + 2\nu}{3} \iff \nu = \frac{3\rho - 1}{2}. \tag{10}$$

By reinserting Eq. 10 into Eq. (8) and balancing the remaining terms, we finally get the estimates for ν and ρ shown in Table 1, which, when compared to the exact ones, are remarkably accurate. In general, the relation between ν and ρ in Eq. (10) has been compared in various ensembles of randomly branching polymers and has been found to be very accurate in all cases [62]. Its practical implications are quite remarkable, as it allows us to connect branching (ρ) to $3D$ conformations (ν) by determining either of the two exponents in terms of the other. In the context of RNA, this relation is particularly useful, since we can determine ρ from the topological properties of its structure and extract ν afterwards.

3 Branching properties of viral RNAs

3.1 Random RNAs

Unlike viral RNA genomes which have a well-defined sequence length, random RNAs can be used to generate sequences of (in principle) arbitrary length and nucleotide composition. This enables one to explore how their topological properties change with both length and composition and obtain different scaling relationships (Sec. 2.2). It is important to note here that RNA sequence length N_{nt} is, on average, directly proportional to the tree size N of its structure, and the two quantities can be used interchangeably.

Figure 4 shows some examples of scaling laws for uniformly random RNA sequences, $f(A) = f(C) = f(U) = f(G) = 0.25$, where $f(n)$ is the frequency of a nucleotide in the sequence. While some properties, such as $\langle MLD \rangle$ and $\langle N_{br} \rangle$, follow a scaling law with a well-defined exponent, others, such as the ratio of the number of degree 1 nodes (leaves of the tree—corresponding to hairpin configurations of RNA)
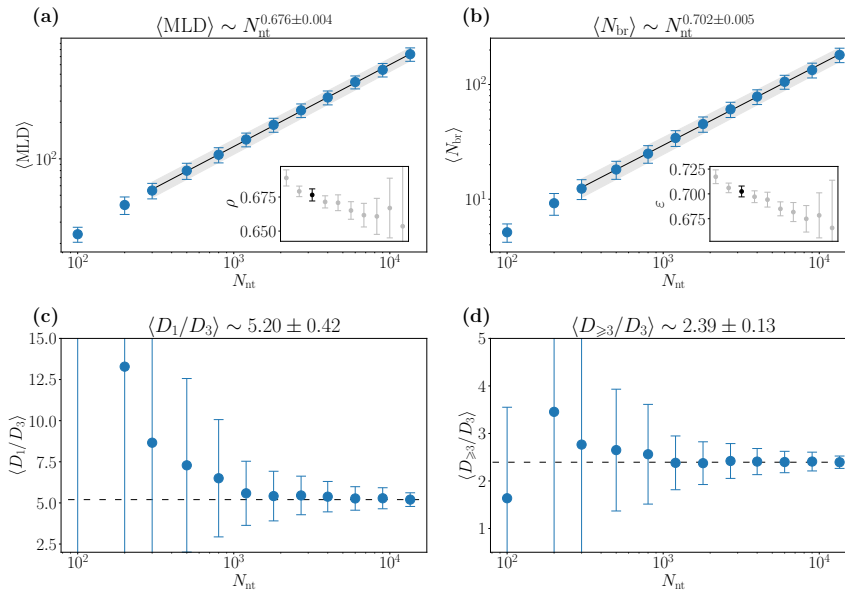
Fig. 4 Scaling of topological properties of (uniformly) random RNA sequences with their length: **(a)** maximum ladder distance $\langle MLD \rangle$, **(b)** branch weight $\langle N_{br} \rangle$, **(c)** ratio of the number of degree 1 and degree 3 nodes D_1/D_3, and **(d)** ratio of the number of degree 3 and all multiloop ($\geqslant 3$) nodes, $D_{\geqslant 3}/D_3$. Each point in the plots represents an average over 200 random sequences and the error bars show the standard deviation. Insets in panels (a) and (b) show how the scaling exponents change with the starting point of the fit, with the black point corresponding to a fit over the shaded region in the panel.

and degree 3 nodes, D_1/D_3, tend towards a constant value. The scaling exponents are of course *asymptotic* properties valid for large RNA structures, as seen in the insets in panels (a) and (b) of Fig. 4, which show how the fitted values of exponents change as shorter sequences are progressively removed from the fit.

Nucleotide composition can vary significantly between different viral species (and biological RNAs in general) [72, 73], affecting their properties. Still, changing the composition of random RNA sequences mainly influences the prefactor of the ⟨MLD⟩ scaling law (Fig. 5a) and only minimally its exponent (Fig. 5b), even when their composition deviates significantly from a uniformly random one, as evaluated by the Euclidean distance $\delta^2 = \sum_{n \in \{A,C,G,U\}} [f(n) - 0.25]^2$. The change in the prefactor appears to be related to a decrease in the base pair percentage $2B/N_{nt}$ (Fig. 5c), which is perhaps unsurprising, as this leads to a smaller size of the RNA graph N at the same sequence length N_{nt}.
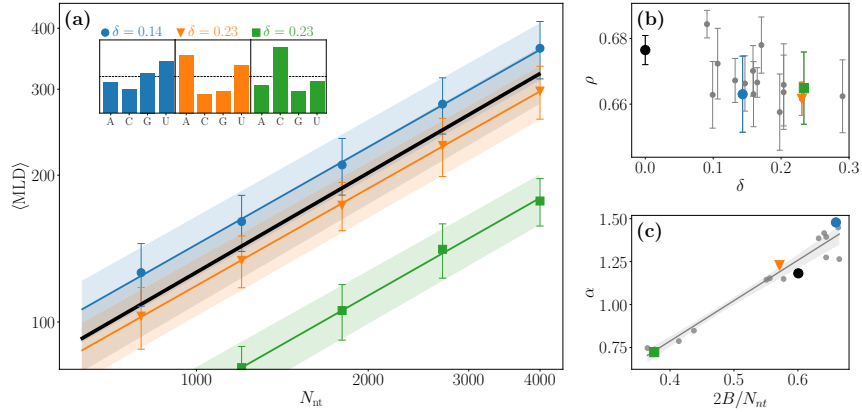
Fig. 5 (a) Scaling of ⟨MLD⟩ with sequence length, $\langle MLD \rangle = \alpha N_{nt}^{\rho}$, for random RNA sequences with different nucleotide compositions. Each point in the plots represents an average over 200 random sequences with 500 thermal ensemble folds for each. (b) Scaling exponent ρ for random RNAs as a function of the Euclidean distance from the uniform composition δ. (c) Scaling prefactor α for RNAs with different compositions as a function of the base pair percentage, $2B/N_{nt}$.

3.2 Viral RNAs

Random RNA sequences show what we can expect of biological and viral RNAs in general [74, 75]. An important example of this is the compactness of the viral RNA folds as captured by its proxy measure, the MLD (cf. Sec. 2.2). As already mentioned, the MLD of the genomes of +ssRNA viruses with icosahedral capsids, which need to pack the genome into a comparatively small capsid, was found to be significantly
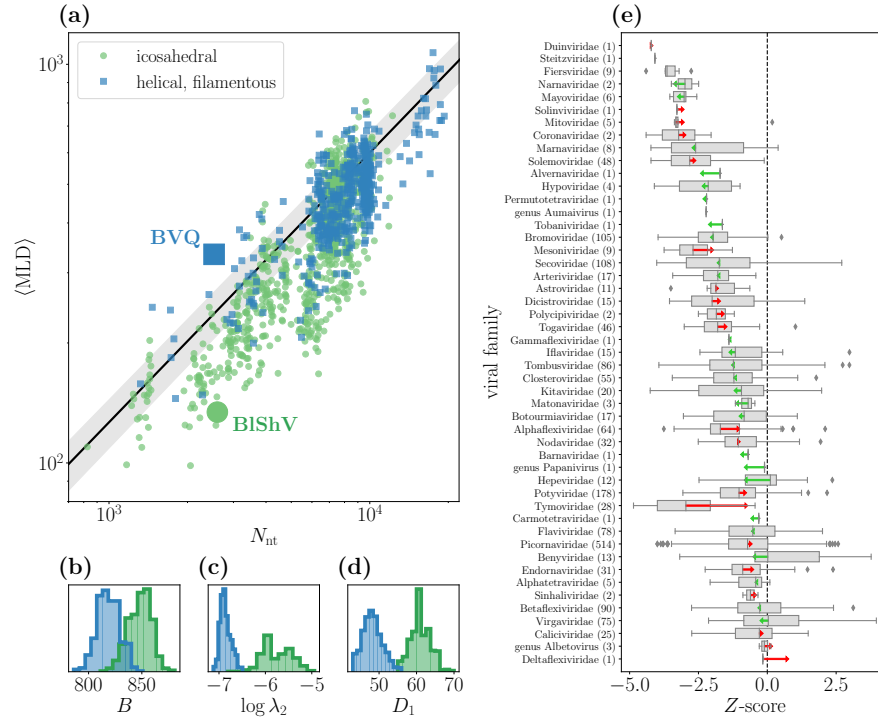
Fig. 6 **(a)** ⟨MLD⟩ of ∼ 1500 genomes of +ssRNA viruses of different lengths and capsid types. Black line shows the scaling for random RNA sequences with uniform composition, ⟨MLD⟩ ∼ $N_{\mathrm{nt}}^{0.676}$, and the shaded area shows the region where $|Z| \leqslant 1$. **(b)**–**(d)** Number of base-pairs B, logarithm of the second Laplacian eigenvalue $\log \lambda_2$, and number of degree 1 nodes D_1 of genomes of two viruses indicated in panel (a), BQV (large square) and BlShV (large circle). **(e)** Distribution of ⟨MLD⟩ Z-scores in different viral families, calculated with respect to random RNA sequences with uniform composition. Arrows indicate the shift in the median Z-score when the ⟨MLD⟩ is calculated with respect to random RNA with closest viral-like composition. The number of genomes included in each viral family is noted in parentheses next to its name.

smaller compared to random RNA sequences of viral-like composition [39, 40]. On the other hand, the MLD of genomes of viruses with helical capsids, which can in principle extend indefinitely, was indistinguishable from that of random RNAs. Figure 6a demonstrates these differences on an extended set of ∼ 1500 genomes of +ssRNA viruses from different families, obtained from the Virus Metadata Resource of ICTV [76], highlighted by capsid type.

Such an analysis opens up the possibility of comparing other topological properties of compact and non-compact viruses, as illustrated in panels (b)–(d) of Fig. 6 for genomes of beet virus Q (BVQ) and blueberry shock virus (BlShV), which are of comparable length but significantly less and more compact than random RNA, respectively (Fig. 6a). The second Laplacian eigenvalue, for instance, identifies the compact genome of BlShV as more star-like (cf. Fig. 2), but even more interesting is

that the genome of BlShV forms *more* base pairs compared to the genome of BVQ and is at the same time located *lower* than the scaling law for uniformly random RNAs. This is in direct contrast to observations in random RNAs with different nucleotide composition, where the scaling prefactor was reduced for those RNAs which form fewer base pairs (Fig. 5), and indicates that the compactness of viral RNAs goes beyond simple differences in nucleotide composition.

Difference in a quantity O between viral and random RNAs can also be evaluated through the Z-score,

$$Z = \frac{\langle O \rangle_{\text{viral}} - \langle O \rangle_{\text{random}}}{\sigma(O)_{\text{random}}}. \tag{11}$$

As Fig. 6e shows, this allows to study the properties of different sets of genomes, in this case grouped by viral family. It is immediately obvious that genomes in certain families are overall more compact than what would be expected of similar random RNAs, while the compactness of genomes in other viral families is indistinguishable from random RNAs. At the same time, there is also quite some degree of variation within viral families. Importantly, the difference in compactness typically persists no matter whether the genome $\langle \text{MLD} \rangle$ is compared to random RNA with a nucleotide composition similar to the one of the genome or to a uniformly random RNA, as indicated by the arrows in Fig. 6b. (Notable exception are Tymoviridae, which have a significantly different composition [39, 40].) This implies that the difference in nucleotide composition of various +ssRNA genomes does not suffice to explain the resulting differences in their compactness as measured by the $\langle \text{MLD} \rangle$.

k-let shuffle of viral genomes

Both mononucleotide and dinucleotide frequencies of viral RNA genomes exhibit biases among different viral species [72, 77], even if they share the same host [78]. These biases are reflected in other properties—for instance, dinucleotide frequencies at codon position 2-3 were shown to explain the majority of codon usage bias [79]. Studies have made it clear that nucleotide composition alone does not suffice to explain the observed differences in the $\langle \text{MLD} \rangle$ of viral genomes [39, 40] (see also Fig. 6), which is further supported by computational observations that *synonymous* mutations preserving both mononucleotide and dinucleotide frequencies easily erase their characteristic compactness [40, 41].

One can thus take a step further and compare instead the topological properties of viral RNAs with their shuffled versions which conserve higher-order nucleotide frequencies of the original genomes. This can be achieved by using a *k*-let preserving shuffling algorithm, such as implemented by uShuffle [80], to shuffle the original genome sequences while *exactly* preserving *k*-let nucleotide frequencies. This means that for $k = 1$ we preserve mononucleotide frequencies (nucleotide composition), for $k = 2$ we preserve dinucleotide frequencies, and so on. Of course, as k increases, the number of possible shuffled sequences decreases. In the examples of BVQ and BlShV, there are still $\sim 10^3$–10^4 possible shuffled sequences available for $k = 10$, while the count drastically drops for $k = 11$ where only $\lesssim 10$ different shuffled

sequences exist (Fig. 7). For the non-compact genome of BVQ, k-let shuffling does not produce drastic changes and the $\langle$MLD$\rangle$ of the shuffled sequences is comparable to that of random RNA for all possible values of k (Fig. 7b). On the other hand, for the compact genome of BlShV, there is a drastic change in the range of $k = 8$–10. For lower values of k, the shuffle completely destroys the compactness of the genome, as was previously seen for $k = 1$ and $k = 2$ [39–41]. For $k = 9$, however, the $\langle$MLD$\rangle$ remains close to the compactness of the original genome, and for $k = 10$ it is indistinguishable from it. There thus appears to be a particular length scale ($k \approx 10$) at which shuffling the compact viral genomes while preserving their k-let nucleotide frequency also preserves their compactness. This observation could hold important clues to the question of where at the sequence level the genome compactness is encoded, but of course needs to be explored more carefully.

Fig. 7 $\langle$MLD$\rangle$ of shuffled sequences of **(a)** BlShV and **(b)** BVQ genomes with preserved k-let frequencies. The genomes of these two viruses are more and less compact than comparable random RNA, respectively (Fig. 6a). The distributions represent sequence ensembles, with 200 shuffled sequences used where possible (i.e., except for $k \geqslant 11$ and the wild-type genome).
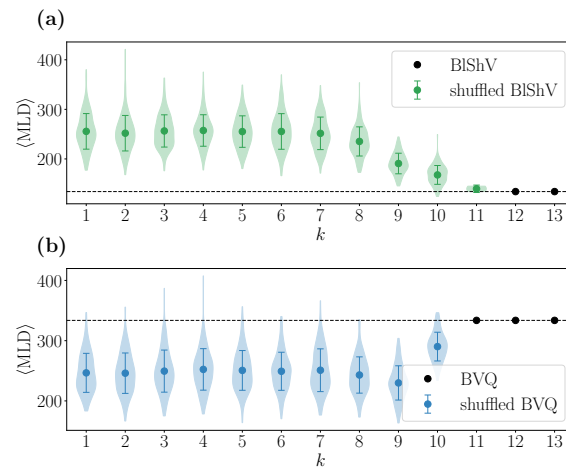
4 Influence of model parameters

RNA secondary structure treated as a branched polymer provides a lot of information about the topological and structural nature of +ssRNA viral genomes and random RNAs alike. The inability to exactly predict the base pair patterns in a folded RNA sequence, however, can lead to differences in the predicted topology of the secondary structure, with consequences also for the subsequent tertiary structure prediction [81]. While prediction of thermal structure ensembles necessitates the use of energy-based models, there is nonetheless a wide variety of model parameters that can influence the resulting predicted structures. We will briefly comment on the effect that two of the most important ones—namely, energy of multiloop formation

and the maximum allowed base pair span—have on the topological measures of secondary structure of viral RNA genomes.

4.1 Multiloop energy models

Energy-based folding algorithms predict both the minimum free energy fold of an RNA sequence as well as its pairing probability matrix, from which an ensemble of thermal folds can be obtained. These algorithms typically use a nearest neighbour energy model that breaks down the energy of an RNA structure into a sum of energies of its constituent loops. Commonly used sets of energy parameters are based on measurements provided by Turner [82], with two particular sets of parameters—Turner1999 and Turner2004—used as a basis by different versions of the most popular energy-based folding software such as ViennaRNA and RNAstructure. Several efforts have also been made to improve on these parameter sets by using various computational optimization techniques [83, 84].

Among the numerous energy parameters involved in structure prediction, *multiloop energies* are the least accurately known [85], even though occurrences of multiloops of degree 10 or higher are not uncommon in various RNAs [86]. Since allowing for an arbitrary size of a multiloop increases the computational complexity of structure prediction, earliest energy-based structure prediction models simply neglected multiloop contributions to the energy [87]. Most of the current structure prediction software assumes that the energy of a multiloop depends only on the amount of enclosed base pairs (number of branches) and the number of unpaired nucleotides in it, and uses a linear model of the form

$$E_{\text{multiloop}} = E_0 + E_{\text{br}} \times [\text{branches}] + E_{\text{un}} \times [\text{unpaired nucleotides}], \quad (12)$$

where E_0 is the energy contribution for multiloop initiation, and E_{br} and E_{un} are the energy contributions for each enclosed base pair and unpaired nucleotide, respectively. While this form has been chosen mostly for its computational simplicity, improved models of multiloop energy that have been proposed seemingly do not lead to improved multiloop predictions compared to the linear one [88, 89].

Table 2 Comparison of different energy parameters for the linear multiloop energy model [Eq. (12)] used in prediction of RNA secondary structure.

Energy model	E_0	E_{un}	E_{br}	Ref.
Turner1999	10.1	−0.3	−0.3	[82]
Turner2004	9.25	NA	0.63	[82]
ViennaRNA ($< $ v2.0)	3.4	0.0	0.4	[44]
ViennaRNA (v2.0+)	9.3	0.0	−0.9	[44]
RNAstructure	9.3	0.0	−0.6	[45]
Andronescu2007	4.4	0.04	0.03	[83]
Langdon2018	9.3	0.0	−0.8	[84]

Table 2 gives an overview of some of the most commonly used energy parameters for the linear multiloop model (Eq. (12)). A notable difference between the models lies not only in the magnitude but in *the sign* of the parameter E_{br} which controls the number of branches stemming from the multiloop. Earlier versions of ViennaRNA (until v2.0), for instance, used a positive value of this parameter, penalizing high-degree nodes, while the latest versions of the software use a negative value, promoting high-degree nodes.

These differences will of course reflect in the predicted structures of long RNAs and their topological properties. Since the different energy models in Table 2 also differ in other aspects, it is easiest to compare multiloop energy parameters by modifying *only* the multiloop parameters in the current parameter set used by ViennaRNA (v2.4) with the ones from older versions (< v2.0), resulting in a modified set of parameters ViennaRNA-mod. Differences in their predictions are illustrated in Fig. 8. The opposite signs of the parameter E_{br} clearly lead to very different distributions of
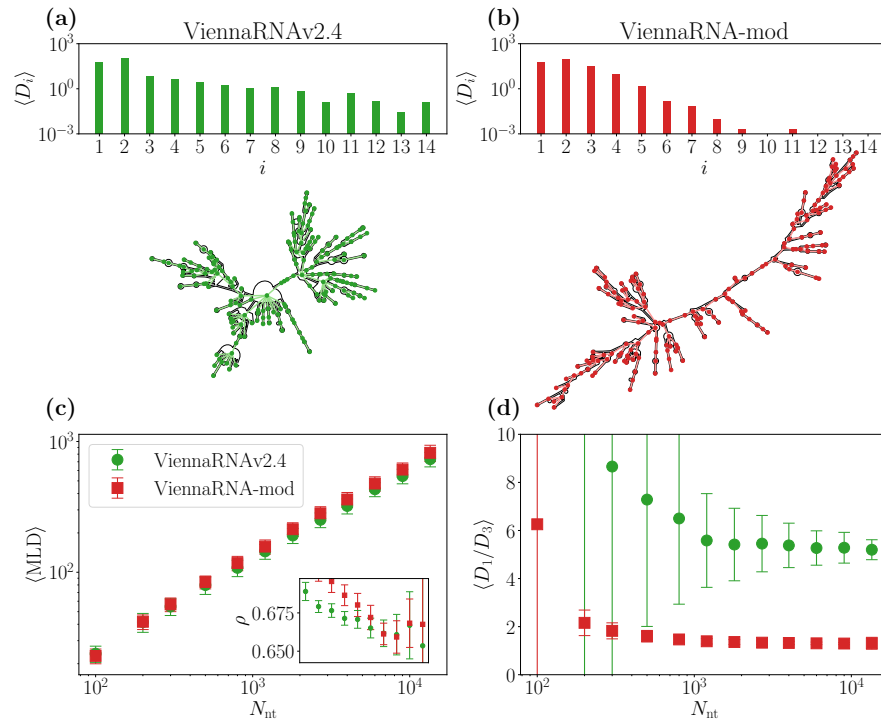
Fig. 8 Number of nodes of different degrees $\langle D_i \rangle$ in the BlShV genome obtained using the energy model in ViennaRNA (v2.4) with either **(a)** default multiloop energy parameters or **(b)** multiloop energy parameters from older versions of ViennaRNA (< v2.0). For the values of these parameters, see Table 2. Note the logarithmic scale in the histograms. Each panel also shows an example structure of the genome. Scaling of **(c)** $\langle MLD \rangle$ and **(d)** the ratio of the number of degree 1 and degree 3 nodes D_1/D_3 with the sequence length of uniformly random RNA as predicted by the two different sets of multiloop energy parameters.

node degrees in the genome of BlShV, with the modified set of parameters predicting far fewer multiloops ($D_{\geqslant 3}$). Interestingly enough, however, the different multiloop energy parameters do not seem to lead to a different scaling behaviour of the $\langle MLD \rangle$ of uniformly random RNA (Fig. 8c), as the exponent ρ becomes indistinguishable between the two cases in the asymptotic limit of long sequences. On the other hand, the ratio of the number of degree 1 and degree 3 nodes is completely different (Fig. 8d). The choice of the multiloop energy parameters can thus lead to important differences in the predicted RNA topology [88, 90], which needs to be taken into account when comparing results obtained by existing studies on the branching properties of viral RNAs [37–40] that use different versions of folding software and thus potentially different energy models.

4.2 Maximum base pair span

Parts of the folding process in very long RNA molecules (i.e., over several hundred nucleotides in length) are influenced by various factors such as co-transcriptional folding and the presence of other molecules in the cell [91, 92]. Consequently, the accuracy of RNA secondary structure prediction in general decreases with the span of a base pair—the length of the nucleotide sequence between two paired bases [93]. This effect can be incorporated in the folding prediction by restricting the maximum allowed base pair span [44, 93], which not only tends to yield more plausible local structure predictions but also drastically increases the computational efficiency. While restrictions on the maximum base pair span in the range of 200–600 nt are often made to improve the prediction of local structural elements [94, 95], this can neglect long-range base pairing and global structure which have been shown to be important in numerous +ssRNA genomes, including that of SARS-CoV-2 [17, 96, 97].

It is therefore natural to wonder to what extent restricting the maximum allowed base pair span influences the topological properties of the structure of viral RNAs. As panels (c) and (f) of Fig. 9 show, restricting the maximum base pair span to ~ 1000–2000 nt slightly changes the $\langle MLD \rangle$, regardless of whether the original genome belongs to a compact class or not, as long-range base pairs are progressively removed from the global structure (panels (b) and (e)). However, when the maximum base pair span is restricted to $\lesssim 1000$ nt, $\langle MLD \rangle$ starts to decrease drastically. Inspection of the resulting structures (panels (a) and (d)) shows that significant maximum base pair span restrictions eventually result in RNA topology becoming more star-like because the individual hairpins are effectively being "strung" on a backbone of single-stranded RNA. This leads to a decrease in $\langle MLD \rangle$ which by definition does not take into account single-stranded regions of RNA (Sec. 2.1), and could consequently affect the results showing that viral RNAs are more compact than random ones (Sec. 3). From a topological and structural perspective, imposing a drastic restriction on the maximum base pair span in RNA would thus not only require a redefinition of the MLD but perhaps even a different mapping of its secondary structure onto a graph.
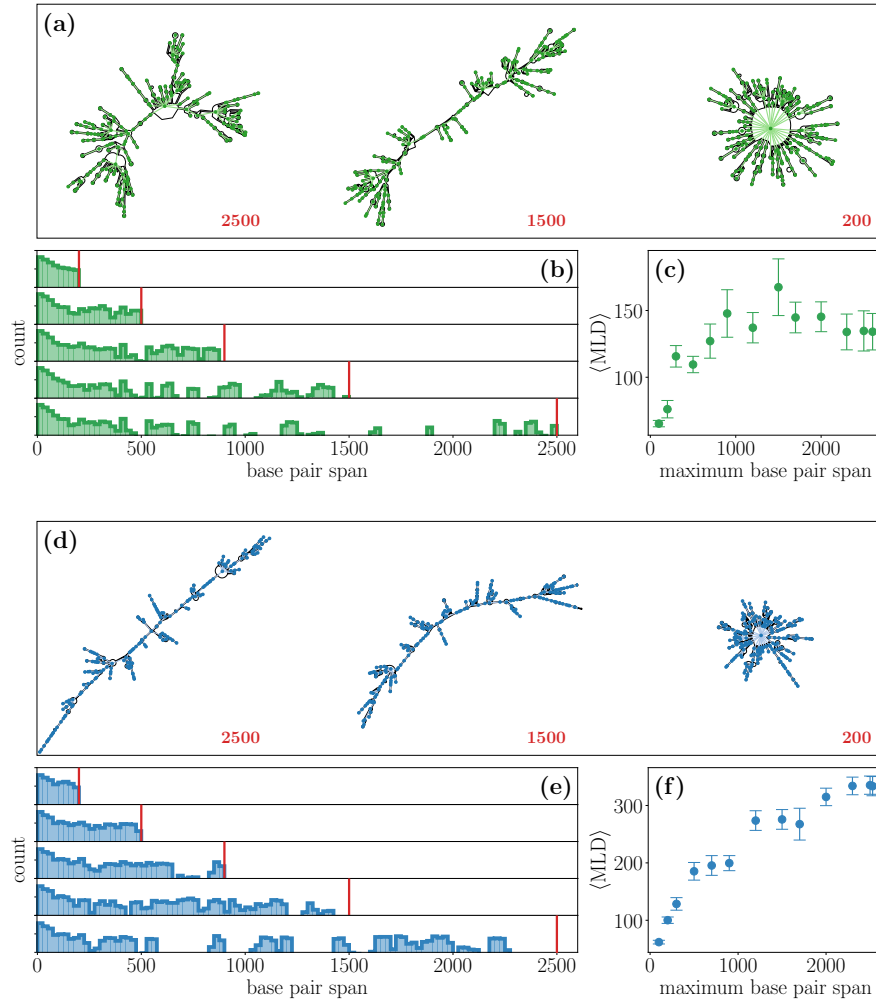
Fig. 9 **(a)** Representative structures of BlShV genome with different maximum base pair restrictions, indicated with each structure. **(b)** Distributions of base pair spans in the thermal ensemble of BlShV genome structures with different maximum base pair span restrictions (marked by vertical bars). **(c)** ⟨MLD⟩ of the BlShV genome as a function of the maximum base pair span restriction. **(d)–(f)** Same as panels (a)–(c) but for the BVQ genome.

5 Field-theoretical description of viral RNA as a branched polymer

RNA branching is also intimately connected with long-range interactions, such as electrostatic self-interaction and interactions between the RNA and capsid proteins. Since the strength of RNA self-interaction (base-pairing) is relatively weak and may

easily be affected by either thermal fluctuations or electrostatic interactions [25], *annealed branched polymers* present a viable *coarse-grained* model system. Here, one starts with the grand canonical partition function [98]

$$\Xi(K, f_e, f_b; V) = \sum_{N, N_e, N_b} K^N f_e^{N_e} f_b^{N_b} \Omega(N, N_e, N_b; V) \tag{13}$$

where K are bonds (edges), and f_e and f_b are end- and branch point fugacities of the annealed polymer (with hairpins counted as the end points, cf. Sec. 2.1). Branch points (nodes) of high degree ($d_i > 3$) can be considered as combinations of branch points of degree 3 and thus need not be treated separately in this description. The function $\Omega(N, N_e, N_b; V)$ is the number of ways to arrange N bonds, N_e end points, and N_b branch points on a lattice of volume V.

The grand canonical partition function can be obtained in the $n \to 0$ limit of the partition function of an $O(n)$ model of a magnet [99] as a functional integral over a continuous field Ψ [60]

$$\Xi(K, f_e, f_b; V) \simeq \int \mathcal{D}[\Psi] e^{-\beta F_0[\psi]}, \tag{14}$$

with the square of Ψ proportional to the monomer density. The saddle-point (mean-field) free energy F_0—in absence of electrostatic interactions—is then

$$\beta F_0[\Psi] = \int_V d^3 r \left[\frac{a^2}{6} |\nabla \Psi|^2 + \frac{1}{2} \upsilon \Psi^4 - \frac{1}{\sqrt{a^3}} \left(f_e \Psi + \frac{a^3}{6} f_b \Psi^3 \right) \right], \tag{15}$$

where a is the statistical step length (Kuhn length; for RNA, ~ 1 nm) and υ is the (repulsive) short-range excluded volume interaction term. Branch point and end point terms proportional to Ψ and Ψ^3 are negative (attractive), therefore increasing the local monomer density. The annealed numbers of end- (N_e) and branch points (N_b) of the RNA (corresponding to D_1 and $D_{\geqslant 3}$ in the graph representation, respectively) are related to the fugacities f_e and f_b by [60]

$$N_e = -f_e \frac{\partial \beta F_0}{\partial f_e} \qquad \text{and} \qquad N_b = -f_b \frac{\partial \beta F_0}{\partial f_b}. \tag{16}$$

If the total number of monomers is fixed, the number of end points for a single RNA molecule with no closed loops depends on the number of branch points as $N_e = N_b + 2$, meaning that f_e is not a free parameter. The polymer is linear if $f_b = 0$, and the number of branch points increases with f_b.

On this level of the description of RNA branching, *long-range interactions* can be straightforwardly implemented by adding additional terms to the free energy functional (Eq. (15)), depending on local interaction fields and couplings with the Ψ field [35, 36, 100]. The free energy for, e.g., electrostatic interactions described by electrostatic potential field Φ would read

$$\beta F = \beta F_0[\Psi] + \beta F[\Phi] + \beta\tau \int_V \mathrm{d}^3 r \; \Phi\Psi^2 \,, \qquad (17)$$

with τ being the charge per monomer and $\beta F[\Phi]$ the electrostatic interaction term [101], which can also describe the interactions between the genome and the capsid proteins. On the other hand, non-electrostatic interactions implied by *packaging signals* require a modified approach [102].

Modifying the topology of the genome by varying the fugacities f_e and f_b allows to use this methodology to simulate systems of viral RNAs with a fixed number of end- and branch points. This coarse-grained description of branching is complementary to the methodology based on explicit planar tree structure [42] and can, for instance, differentiate between the encapsulation behavior of RNA1 of brome mosaic virus (BMV) that has 65 branch points and of RNA1 of cowpea chlorotic mottle virus (CCMV) with 60.5 branch points—correctly confirming that BMV RNA1 is preferentially packaged over CCMV RNA1 by the CCMV capsid protein [35]. Furthermore, the straightforward implementation of electrostatic interactions on top of RNA topology is probably the most important forte of this methodology and allows one to assess the role of electrostatics in spontaneous co-assembly of the negatively charged genome and positively charged capsid proteins. Using this approach, it has been demonstrated that branching in fact allows viruses to maximize the amount of encapsulated genome and makes assembly more efficient [35], implies negative osmotic pressures across the capsid wall [36], and can explain the effect of number and location of charges in the capsid protein tails [103].

While the field-theoretical annealed-branching description is without doubt heavily coarse-grained, it not only provides an approximate implementation of topology, but also readily incorporates *short- and long-range interactions* on a level amenable to analytical calculations. Different approaches of standard polymer theory can then be transplanted into the statistical mechanics of RNA providing further insight into the coupling between topology and virion self-assembly.

6 Conclusions

Mapping RNA secondary structure onto a graph enables its description as a branched polymer and a subsequent study of its topological properties (e.g., MLD and node degree distribution; Sec. 2). This, in turn, can be connected to the physical properties of the RNA, such as its size as given by its radius of gyration. This approach also provides insight into how RNA structure compares to other types of branched polymers in terms of, for instance, their scaling exponents. Branching properties of RNA also allow comparison of +ssRNA genomes of different viral families, both among themselves and with random RNA of similar length and composition (Sec. 3). Such an analysis reveals the unusual compactness of genomes from certain viral families and may eventually provide an answer to the question of where in the sequence of viral RNA its physical compactness is encoded.

Describing RNA as a graph and in this way treating it as a branched polymer is, of course, an approximation. This description is made on the topological level of the RNA secondary structure, itself deriving from an energy-based base pair prediction, and thus depends on the model parameters used in it (Sec. 4). Secondary structure prediction furthermore remains agnostic to steric interactions between different parts of the RNA, long-range interactions such as electrostatics, and other tertiary interactions. Some of these effects can, however, be included on a coarse-grained level by treating the branching RNA structure with a field-theoretical description, enabling, for instance, a coupling between topological parameters (such as node degree distribution) and electrostatic interactions (Sec. 5).

Predictions obtained by treating viral RNA as a branched polymer can, to an extent, be verified experimentally, for instance by measuring its radius of gyration by gel electrophoresis [37] or by determining the distributions of node degrees and segment lengths from 2D projections of viral RNA molecules imaged by cryo-EM [37, 104]. Properly designed experiments on long RNA molecules of different lengths could thus, in principle, be compared with the predictions given by different multiloop energy models and in this way help determine the most appropriate model. Lastly, treating RNA as a graph and being able to understand how its sequence leads to its topological and structural properties can be beneficial not only in the ability to interfere with the function of viral genomes but also in the design of RNA molecules with specific topology for use in nanomedicine and synthetic biology [52, 105, 106].

References

1. S.R. Eddy, Nat. Rev. Genet. **2**, 919 (2001)
2. J.S. Mattick, I.V. Makunin, Human Mol. Genet. **15**, R17 (2006)
3. J. Gorodkin, W.L. Ruzzo, *RNA sequence, structure, and function: Computational and bioinformatic methods* (Springer, 2014)
4. X.W. Wang, C.X. Liu, L.L. Chen, Q.C. Zhang, Nat. Chem. Biol. **17**, 755 (2021)
5. S.A. Mortimer, M.A. Kidwell, J.A. Doudna, Nat. Rev. Genet. **15**, 469 (2014)
6. P. Brion, E. Westhof, Annu. Rev. Biophys. Biomol. Struct. **26**, 113 (1997)
7. A.M. Mustoe, C.L. Brooks, H.M. Al-Hashimi, Annu. Rev. Biochem. **83**, 441 (2014)
8. N. Leontis, E. Westhof (eds.), *RNA 3D structure analysis and prediction* (Springer, 2012)
9. Z. Miao, E. Westhof, Annu. Rev. Biophys. **46**, 483 (2017)
10. J. Li, S.J. Chen, Front. Mol. Biosci. **8** (2021)
11. J.T. Low, K.M. Weeks, Methods **52**, 150 (2010)
12. R. Lorenz, I.L. Hofacker, P.F. Stadler, Algorithms Mol. Biol. **11**, 1 (2016)
13. D. Mitchell III, S.M. Assmann, P.C. Bevilacqua, Curr. Op. Struct. Biol. **59**, 151 (2019)
14. E.C. Holmes, *The evolution and emergence of RNA viruses* (Oxford University Press, 2009)
15. Y. Liu, E. Wimmer, A.V. Paul, Biochimi. Biophys. Acta **1789**, 495 (2009)

16. L.R. Newburn, K.A. White, Virology **479**, 434 (2015)
17. B.L. Nicholson, K.A. White, Curr. Op. Virol. **12**, 66 (2015)
18. M.A. Boerneke, J.E. Ehrhardt, K.M. Weeks, Annu. Rev. Virol. **6**, 93 (2019)
19. A. Schneemann, Annu. Rev. Microbiol. **60**, 51 (2006)
20. A. Rao, Annu. Rev. Phytopathol. **44**, 61 (2006)
21. R.F. Garmann, M. Comas-Garcia, C.M. Knobler, W.M. Gelbart, Acc. Chem. Res. **49**, 48 (2016)
22. M. Comas-Garcia, Viruses **11**, 253 (2019)
23. R. Twarock, R.J. Bingham, E.C. Dykeman, P.G. Stockley, Curr. Op. Virol. **31**, 74 (2018)
24. P.G. Stockley, R. Twarock, S.E. Bakker, A.M. Barker, A. Borodavka, E. Dykeman, R.J. Ford, A.R. Pearson, S.E. Phillips, N.A. Ranson, et al., J. Biol. Phys. **39**, 277 (2013)
25. R. Zandi, B. Dragnea, A. Travesset, R. Podgornik, Phys. Rep. **847**, 1 (2020)
26. J.D. Perlmutter, M.F. Hagan, Annu. Rev. Phys. Chem. **66**, 217 (2015)
27. Y. Hu, R. Zandi, A. Anavitarte, C.M. Knobler, W.M. Gelbart, Biophys. J. **94**, 1428 (2008)
28. M. Comas-Garcia, R.D. Cadena-Nava, A. Rao, C.M. Knobler, W.M. Gelbart, J. Virol. **86**, 12271 (2012)
29. C. Beren, L.L. Dreesens, K.N. Liu, C.M. Knobler, W.M. Gelbart, Biophys. J. **113**, 339 (2017)
30. L. Marichal, L. Gargowitsch, R.L. Rubim, C. Sizun, K. Kra, S. Bressanelli, Y. Dong, S. Pana-handeh, R. Zandi, G. Tresset, Biophys. J. **120**, 3925 (2021)
31. J.D. Perlmutter, C. Qiao, M.F. Hagan, eLife **2** (2013)
32. R.F. Garmann, A.M. Goldfain, C.R. Tanimoto, C.E. Beren, F.F. Vasquez, D.A. Villarreal, C.M. Knobler, W.M. Gelbart, V.N. Manoharan, Proc. Natl. Acad. Sci. USA **119**, e2206292119 (2022)
33. S. Poblete, A. Božič, M. Kanduč, R. Podgornik, H.A. Vargas Guzmán, ACS Omega **6**, 32823 (2021)
34. S.W. Singaram, R.F. Garmann, C.M. Knobler, W.M. Gelbart, A. Ben-Shaul, J. Phys. Chem. B **119**, 13991 (2015)
35. G. Erdemci-Tandogan, J. Wagner, P. Van Der Schoot, R. Podgornik, R. Zandi, Phys. Rev. E **89**, 032707 (2014)
36. G. Erdemci-Tandogan, J. Wagner, P. van der Schoot, R. Podgornik, R. Zandi, Phys. Rev. E **94**, 022408 (2016)
37. A. Gopal, D.E. Egecioglu, A.M. Yoffe, A. Ben-Shaul, A.L. Rao, C.M. Knobler, W.M. Gelbart, PLoS One **9**, e105875 (2014)
38. A. Borodavka, S.W. Singaram, P.G. Stockley, W.M. Gelbart, A. Ben-Shaul, R. Tuma, Biophys. J. **111**, 2077 (2016)
39. A.M. Yoffe, P. Prinsen, A. Gopal, C.M. Knobler, W.M. Gelbart, A. Ben-Shaul, Proc. Natl. Acad. Sci. USA **105**, 16153 (2008)
40. L. Tubiana, A. Božič, C. Micheletti, R. Podgornik, Biophys. J. **108**, 194 (2015)
41. A. Božič, C. Micheletti, R. Podgornik, L. Tubiana, J. Phys. Condens. Matter **30**, 084006 (2018)
42. J. Farrell, J. Dobnikar, R. Podgornik, Submitted (2022)
43. J. Fallmann, S. Will, J. Engelhardt, B. Grüning, R. Backofen, P.F. Stadler, J. Biotechnol. **261**, 97 (2017)
44. R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, Algorithms Mol. Biol. **6**, 1 (2011)
45. J.S. Reuter, D.H. Mathews, BMC Bioinform. **11**, 1 (2010)
46. C.B. Do, D.A. Woods, S. Batzoglou, Bioinformatics **22**, e90 (2006)
47. H.K. Wayment-Steele, W. Kladwang, A.I. Strom, J. Lee, A. Treuille, Eterna Participants, R. Das, BioRxiv (2021). URL `https://www.biorxiv.org/content/10.1101/2020.05.29.124511v2`
48. R.V. Koodli, B. Rudolfs, H.K. Wayment-Steele, Eterna Structure Designers, R. Das, bioRxiv (2021). URL `https://www.biorxiv.org/content/10.1101/2021.08.26.457839v1`
49. M. Liu, E. Poppleton, G. Pedrielli, P. Šulc, D.P. Bertsekas, INFORMS J. Comput. (2022)
50. A. Spasic, S.M. Assmann, P.C. Bevilacqua, D.H. Mathews, Nucleic Acids Res. **46**, 314 (2018)

51. D.H. Mathews, D.H. Turner, Curr. Op. Struct. Biol. **16**, 270 (2006)
52. T. Schlick, Methods **143**, 16 (2018)
53. C. Micheletti, M. Di Stefano, H. Orland, Proc. Natl. Acad. Sci. USA **112**, 2052 (2015)
54. J.A. Izzo, N. Kim, S. Elmetwaly, T. Schlick, BMC Bioinform. **12**, 1 (2011)
55. J.L. Gross, J. Yellen, M. Anderson, *Graph theory and its applications* (Chapman and Hall/CRC, 2018)
56. R. Todeschini, V. Consonni, *Handbook of molecular descriptors* (John Wiley & Sons, 2008)
57. D.H. Rouvray, R.B. King, *Topology in chemistry: Discrete mathematics of molecules* (Elsevier, 2002)
58. S. Sazer, H. Schiessel, Traffic **19**, 87 (2018)
59. S.L. Perry, Curr. Op. Colloid Interface Sci. **39**, 86 (2019)
60. J. Wagner, G. Erdemci-Tandogan, R. Zandi, J. Phys.: Condens. Matter **27**, 495101 (2015)
61. A.M. Gutin, A.Y. Grosberg, E.I. Shakhnovich, Macromolecules **26**(6), 1293 (1993)
62. R. Everaers, A.Y. Grosberg, M. Rubinstein, A. Rosa, Soft Matter **13**(6), 1223 (2017)
63. Z.G. Wang, Macromolecules **50**(23), 9073 (2017)
64. S.M. Bhattacharjee, A. Giacometti, A. Maritan, J. Phys. Cond. Matter **25**(50), 503101 (2013)
65. M. Rubinstein, R.H. Colby, *Polymer Physics* (Oxford University Press, New York, 2003)
66. B. Li, N. Madras, A.D. Sokal, J. Stat. Phys. **80**(3), 661 (1995)
67. G. Parisi, N. Sourlas, Phys. Rev. Lett. **46**, 871 (1981)
68. E.J. Van Rensburg, N. Madras, J. Phys. A: Math. Theor. **25**, 303 (1992)
69. A. Rosa, R. Everaers, J. Phys. A. Math. Theor. **49**, 345001 (2016)
70. A. Rosa, R. Everaers, J. Chem. Phys. **145**, 164906 (2016)
71. P.J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca (NY), 1953)
72. D. Simón, J. Cristina, H. Musto, Front. Microbiol. **12** (2021)
73. E. Schultes, P.T. Hraber, T.H. LaBean, RNA **3**, 792 (1997)
74. P.G. Higgs, J. Phys., I **3**, 43 (1993)
75. P. Clote, F. Ferré, E. Kranakis, D. Krizanc, RNA **11**, 578 (2005)
76. E.J. Lefkowitz, D.M. Dempsey, R.C. Hendrickson, R.J. Orton, S.G. Siddell, D.B. Smith, Nucleic Acids Res. **46**, D708 (2018)
77. E.R. Gaunt, P. Digard, Wiley Interdiscip. Rev. RNA **13**, e1679 (2022)
78. F. Di Giallonardo, T.E. Schlub, M. Shi, E.C. Holmes, J. Virol. **91**, e02381 (2017)
79. I.S. Belalov, A.N. Lukashev, PLOS One **8**, e56642 (2013)
80. M. Jiang, J. Anderson, J. Gillespie, M. Mayne, BMC Bioinform. **9**, 1 (2008)
81. Y. Zhao, J. Wang, C. Zeng, Y. Xiao, Biophys. Rep. **4**, 123 (2018)
82. D.H. Turner, D.H. Mathews, Nucleic Acids Res. **38**, D280 (2010)
83. M. Andronescu, A. Condon, H.H. Hoos, D.H. Mathews, K.P. Murphy, RNA **16**, 2304 (2010)
84. W.B. Langdon, J. Petke, R. Lorenz, in *European Conference on Genetic Programming* (2018), pp. 220–236
85. S. Poznanović, C. Wood, M. Cloer, C. Heitsch, Genes **12**, 469 (2021)
86. J. Wiedemann, J. Kaczor, M. Milostan, T. Zok, J. Blazewicz, M. Szachniuk, M. Antczak, Bioinformatics (2022)
87. M. Zuker, P. Stiegler, Nucleic Acids Res. **9**, 133 (1981)
88. M. Ward, A. Datta, M. Wise, D.H. Mathews, Nucleic Acids Res. **45**, 8541 (2017)
89. M. Ward, H. Sun, A. Datta, M. Wise, D.H. Mathews, Bioinformatics **35**, 4298 (2019)
90. S. Poznanović, F. Barrera-Cruz, A. Kirkpatrick, M. Ielusic, C. Heitsch, J. Struct. Biol. **210**, 107475 (2020)
91. F. Amman, S.H. Bernhart, G. Doose, I.L. Hofacker, J. Qin, P.F. Stadler, S. Will, in *Brazilian Symposium on Bioinformatics* (Springer, 2013), pp. 1–11
92. A.M. Pyle, T. Schlick, J. Mol. Biol. **428**, 733 (2016)
93. R. Lorenz, P.F. Stadler, Genes **12**, 14 (2020)
94. E.J. Archer, M.A. Simpson, N.J. Watts, R. O'Kane, B. Wang, D.A. Erie, A. McPherson, K.M. Weeks, Biochemistry **52**, 3182 (2013)
95. T.C. Lan, M.F. Allan, L.E. Malsick, J.Z. Woo, C. Zhu, F. Zhang, S. Khandwala, S.S. Nyeo, Y. Sun, J.U. Guo, et al., Nat. Comm. **13**, 1 (2022)
96. P. Simmonds, A. Tuplin, D.J. Evans, RNA **10**, 1337 (2004)

97. C. Cao, Z. Cai, X. Xiao, J. Rao, J. Chen, N. Hu, M. Yang, X. Xing, Y. Wang, M. Li, et al., Nat. Comm. **12**, 1 (2021)
98. T. Lubensky, J. Isaacson, Phys. Rev. A **20**, 2130 (1979)
99. T. Lubensky, J. Isaacson, Phys. Lett. A **38**, 339 (1972)
100. S. Li, G. Erdemci-Tandogan, J. Wagner, P. van der Schoot, R. Zandi, Phys. Rev. E **96**, 22401 (2017)
101. A. Šiber, R. Podgornik, Phys. Rev. E **78**, 051915 (2008)
102. C. Huang, R. Podgornik, X. Man, Macromolecules **54**, 9602 (2021)
103. Y. Dong, S. Li, R. Zandi., Phys. Rev. E **102**, 062423 (2020)
104. R.F. Garmann, A. Gopal, S.S. Athavale, C.M. Knobler, W.M. Gelbart, S.C. Harvey, RNA **21**, 877 (2015)
105. S. Jain, Y. Tao, T. Schlick, J. Struct. Biol. **209**, 107438 (2020)
106. C. Geary, G. Grossi, E.K. McRae, P.W. Rothemund, E.S. Andersen, Nat. Chem. **13**, 549 (2021)