

# Scaling properties of RNA as a randomly branching polymer

Domen Vaupotič,<sup>1</sup> Angelo Rosa,<sup>2</sup> Luca Tubiana,<sup>3,4</sup> and Anže Božič<sup>1</sup>

<sup>1)</sup>*Department of Theoretical Physics, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*

<sup>2)</sup>*Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy*

<sup>3)</sup>*Department of Physics, University of Trento, via Sommarive 14, 38123 Trento, Italy*

<sup>4)</sup>*INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, via Sommarive 14, 38123 Trento, Italy*

(\*Electronic mail: [anze.bozic@ijs.si](mailto:anze.bozic@ijs.si))

(Dated: March 31, 2023)

Formation of base pairs between the nucleotides of an RNA sequence gives rise to a complex and often highly branched secondary and tertiary RNA structure. While numerous studies have demonstrated the functional importance of the high degree of RNA branching—for instance, for its spatial compactness or interaction with other biological macromolecules—the nature of RNA branching topology remains largely unexplored. Here, we use the theory of randomly branching polymers to explore the scaling properties of RNA sequences by mapping their secondary structures onto planar tree graphs. Focusing on random RNA sequences of varying lengths, we determine their scaling exponents  $\rho$  and  $\epsilon$ , related to the topology of branching. We find that  $\rho \simeq \epsilon \approx 0.67$ , indicating that the ensembles of RNA secondary structures are characterized by annealed random branching and scale similarly to branched self-avoiding walks in three dimensions. This is already a surprising result, as the secondary structure prediction software functions on the basis of nearest-neighbour energetic considerations, ignoring steric clashes and tertiary interactions. By combining the predicted values of  $\rho$  and  $\epsilon$  with Flory theory we show that the radius of gyration of RNA ensembles in good solvent scales as  $\nu = 0.5$ , as expected for self-avoiding randomly branching trees. We also examine how nucleotide composition, tree topology, and folding energy parameters influence the obtained scaling exponents and we show that both exponents are robust upon reasonable changes in these quantities. Finally, in order to apply the theory of branching polymers to biological RNAs, whose length cannot be arbitrarily varied, we demonstrate how both scaling exponents can be obtained from the distributions of the related topological quantities of individual RNA molecules with fixed length. In this way, we establish a framework to study the branching properties of RNA and compare them to other known classes of branched polymers, while also identifying the conditions under which this description becomes insufficient. By understanding the scaling properties of RNA related to its branching structure we aim to improve our understanding of the underlying principles and open up the possibility to design RNA sequences with desired topological properties.

## I. INTRODUCTION

Branching structures can be found in a wide range of polymeric materials, including networks, gels, and, importantly, biopolymers such as nucleic acids and polysaccharides<sup>1–3</sup>. Compared to their linear counterparts, branching imparts polymers with several favourable properties such as high surface functionality, globular conformation, high solubilities, and more<sup>4</sup>. A ubiquitous, functionally important biopolymer is ribonucleic acid (RNA), which can be seen to effectively behave as a branched polymer due to the arrangement of single-stranded (ss) regions (such as loops, bulges, and stems) and double-stranded (ds) regions formed by base-pairing which can lead to the formation of multi-loops with a large degree of branching. Indeed, a survey of PDB-deposited RNA structures has shown that while multi-loops having branching degree 3 and 4 are the most abundant, branched loops with degree  $\gtrsim 10$  are far from unusual<sup>5</sup>. This propensity for high branch-

ing degrees sets RNA aside from most other known classes of branched polymers<sup>1,2</sup>.

The branching structure of RNA also has a functional importance, which is the most apparent in the genomes of ssRNA viruses<sup>6,7</sup>. Recent experimental and computational studies on these viruses have shown that branching affects their assembly by both influencing the ability of RNA to bind to capsid proteins as well as making the size and the structure of the folded RNA genome comparable to the size of the capsid<sup>8–14</sup>. Furthermore, the RNA branching pattern when combined with electrostatic interactions impacts its osmotic pressure inside the capsid and in this way influences both its packaging efficiency and virion stability<sup>15,16</sup>. All of these properties are arguably the result of an evolutionary pressure. For instance, the ssRNA genomes of icosahedral viruses are significantly more compact than random RNA sequences of comparable length and composition<sup>8,17</sup>. At the same time, even a small percent of synonymous mutations suffices to destroy this compactness

and make the size of viral RNAs indistinguishable from that of random RNAs<sup>9,13</sup>.

While some attempts have been made to experimentally determine the branching patterns of long RNAs from two-dimensional (2D) projections<sup>18</sup>, for the most part such analysis still needs to be carried out using computational predictions of RNA secondary structure. In absence of pseudoknots, the secondary structure of RNA can be mapped to a planar tree graph by mapping its various structural elements (stems, bulges, hairpin loops, ...) to the tree vertices and edges<sup>19–21</sup>. Under this approximation, a basic measure of RNA size is its maximum ladder distance (MLD), which corresponds to the diameter of the tree and correlates well with its physical size as experimentally measured by, for instance, its hydrodynamic radius<sup>17,22</sup>. Such an approach has been previously used to extensively analyze the compactness of viral ssRNA genomes<sup>8,9,13,17</sup>. What is more, mapping RNA structures to planar tree graphs also opens up the possibility to apply the powerful framework of polymer physics—and Flory theory in particular<sup>23–25</sup>—to study the properties of long RNA molecules.

An ensemble of polymer conformations can be described in terms of a relatively small number of observables  $\mathcal{O}$ . Since the number of total accessible conformations of a polymer scales with its number of monomers (bonds)  $N$ , one can adopt the average value of an observable  $\langle \mathcal{O}(N) \rangle$  to classify the polymer ensemble under consideration (such as ideal polymers, self-avoiding polymers,  $\theta$ -polymers, ...) <sup>26</sup>. Due to the scale-invariant character of polymer chains in the large-chain limit<sup>26</sup>,  $\langle \mathcal{O}(N) \rangle$  behaves as a power law of the number of monomers  $N$  of the chain with some characteristic scaling exponent  $\gamma$ :

$$\langle \mathcal{O}(N) \rangle \sim N^\gamma. \quad (1)$$

In general,  $\gamma$  depends on the physical properties of the monomers, for instance on how the monomers interact with each other or with the solvent<sup>27</sup>. Scaling exponents obtained from such power laws allow one to distinguish between branched polymers that live in 2D or 3D, or between polymers that are self-avoiding or at the  $\theta$ -point. They furthermore allow one to distinguish between different kinds of branching patterns—in particular, between regular patterns, such as observed in dendrimers, and random patterns, where the size and disposition of the branches are probabilistic<sup>26</sup>.

In general, the aim of such a description is to obtain a scaling relationship for observables related to the molecule size, as these can be experimentally measured. This was the goal of Fang *et al.*<sup>28</sup>, who used secondary structure prediction and Kramers' formula<sup>29,30</sup> to obtain the scaling exponent  $\nu$  for the radius of gyration of long RNAs. In the case of branching polymers, however, two more exponents are required to get a complete picture of their properties. These two exponents,  $\rho$  and  $\varepsilon$ , describe the scaling of the average distance between two

nodes on the tree and the average weight of branches, respectively<sup>25,31</sup>. These properties are purely topological and their exponents are expected to capture the statistics of the branching structure. For example, when the topology of a random branching is annealed (i.e., when the branches can rearrange themselves), these two exponents coincide<sup>31</sup>. The scaling relation for the radius of gyration can then be obtained from these exponents and from the knowledge of the kind of solvent the polymer is immersed in.

Despite the importance of the branching structure of RNA, its exact nature when compared to other types of branched polymers remains unknown<sup>32,33</sup>, and the scaling properties relating to the topology of branching are largely unexplored<sup>8–10</sup>. In this work, we set to determine the scaling exponents  $\rho$  and  $\varepsilon$  of RNA secondary structures, compare them with known types of branching polymers, and demonstrate how these exponents connect to the radius of gyration of ensembles of RNA folds using Flory theory. Our approach allows us to decouple the assumptions made to obtain the scaling of the radius of gyration from those related to the topology of the folding. We focus on arbitrary (random) RNA sequences, which represent the baseline for RNA structure formation and show the general behaviour one can expect of biological RNAs<sup>34–38</sup>. Using ViennaRNA software to predict the thermal ensembles of their secondary structures, we explore the importance of nucleotide composition, folding energy parameters, and node degree distribution on the scaling behaviour of RNA. Importantly, we also demonstrate how scaling exponents can be determined for RNA sequences with fixed length, which is particularly relevant for biological RNAs, whose length cannot be varied arbitrarily. Our results, which yield  $\rho \simeq \varepsilon \approx 0.67$ , show that although ViennaRNA produces secondary structure folds only on the basis of energetic considerations, thus ignoring steric clashes and tertiary interactions, these folds scale as 3D self-avoiding annealed branched polymers, independently of the energy parameters. Interestingly, our results also show that, in a good solvent, the radius of gyration of RNA scales as  $N^{1/2}$  and not as  $N^{1/3}$  as estimated previously<sup>28</sup>.

The paper is organized as follows: In Sec. II we provide an overview of the theory of branching polymers and introduce the topological observables applicable to RNA secondary structure, their corresponding distribution functions, and their scaling exponents. In Sec. III, we describe the folding algorithm we use to predict RNA secondary structure and describe its tree representation. We also provide details on the important parameters, such as multiloop energy model and RNA sequence composition, that we make use of in our analysis. In Sec. IV, we describe and compare the results for the scaling exponents of random RNA sequences obtained in two different ways. We furthermore verify the robustness of the obtained results by varying the multiloop parameters of the energy model, sequence composition, and topology of branching. The same

section also frames RNA in terms of a randomly branching polymer and compares its scaling behaviour to that of other known types of branching polymers. Finally, in Sec. V we discuss our results in a wider context and show what the scaling exponents obtained from the topology of RNA structure tell us about the scaling of its size as given by the radius of gyration.

## II. SCALING THEORY OF BRANCHING IN RNA MOLECULES

### A. Average polymer behaviour

In this work, we build on a comparison between RNA secondary structure folds and randomly branched polymers. A large body of computational work<sup>25,31,39–41</sup> has demonstrated that in order to completely characterize an ensemble of branched polymers, it is necessary to consider the topology of branching, also known as *tree connectivity*<sup>25</sup>. This is a property that is accessible using simple RNA folding models (described in Sec. III).

As a proper measure of tree connectivity in branched RNA molecules, we introduce the ensemble average of either the maximum (MLD) or average ladder distance (ALD) as a function of the number of monomers  $N$ <sup>21,39</sup>,

$$\langle \text{MLD}(N) \rangle \sim \langle \text{ALD}(N) \rangle \sim N^\rho, \quad (2)$$

both of which account for the average length of linear paths on the tree. Here, the ladder distance (path length)  $\ell_{ij}$  between two nodes  $i$  and  $j$  is defined as the shortest path between them, and we thus have  $\text{MLD} = \max_{i,j} \ell_{ij}$  and  $\text{ALD} = [N(N+1)]^{-1} \sum_{i \neq j} \ell_{ij}$ <sup>21</sup>. As another topological observable, we consider the average branch weight<sup>31</sup>:

$$\langle N_{\text{br}}(N) \rangle \sim N^\varepsilon, \quad (3)$$

which is defined as the average weight of the smallest of the two sub-branches obtained by systematically removing—one at time—edges connecting two neighbouring nodes of the original tree<sup>39</sup> (see also Fig. 2 and Sec. III B).

While the two scaling exponents  $\rho$  and  $\varepsilon$  describe very different quantities, they are not independent from each other. In fact, by making very minimal assumptions on the randomly branching architecture of the polymers, the relation

$$\rho = \varepsilon \quad (4)$$

is expected to hold in general for annealed branching polymers<sup>31</sup>. Accurate numerical proofs of Eq. (4) are given in Refs. 31 and 39 for isolated self-avoiding trees in spatial dimensions from 2 to 9 and in Ref. 40 for melts of trees in 2D and 3D. Confirming the validity of Eq. 4 for RNA structures

would be a strong argument in support of the hypothesis that RNA molecules can also be modeled as randomly branching polymers.

Even the simplest theory for the characterization of secondary structure of generic branching polymers (and RNA molecules in particular) thus has to operate with these two distinct observables,  $\langle \text{ALD}(N) \rangle$  and  $\langle N_{\text{br}}(N) \rangle$ . Furthermore, obtaining the scaling exponents  $\rho$  and  $\varepsilon$  related to these two observables allows us to obtain the scaling exponent  $\nu$  for the radius of gyration as well. We elaborate on this connection in Sec. V, where we also explain in some detail the possible connection between RNA secondary and tertiary structure, in particular with regard to the mean radius of gyration of an ensemble of molecules.

### B. Distribution functions

Beyond the averages of observables introduced in the previous section, another fundamental part of information on the scaling exponents  $\rho$  and  $\varepsilon$  is contained in their corresponding distribution functions<sup>41</sup>. In particular, theoretical considerations show that the path length distribution functions  $p(\ell)$  in trees of size  $N$  collapse onto *universal* master curves when plotted as a function of the rescaled path length  $x = \ell / \langle \text{ALD}(N) \rangle$ <sup>41</sup>:

$$p(\ell) = \frac{1}{\langle \text{ALD}(N) \rangle} q\left(\frac{\ell}{\langle \text{ALD}(N) \rangle}\right). \quad (5)$$

These master curves are described well by the one-dimensional Redner-des Cloizeaux (RdC) function<sup>41</sup>:

$$q(x) = Cx^\theta \exp(-(Kx)^t), \quad (6)$$

whose numerical constants<sup>42</sup>

$$C = t \frac{\Gamma^{\theta+1}((\theta+2)/t)}{\Gamma^{\theta+2}((\theta+1)/t)}, \quad (7)$$

$$K = \frac{\Gamma((\theta+2)/t)}{\Gamma((\theta+1)/t)}, \quad (8)$$

follow from the conditions that  $p(\ell)$  is normalized to 1 and that its first moment,  $\langle \text{ALD}(N) \rangle$ , is the only relevant scaling variable. Here,  $\Gamma(x)$  denotes the gamma function. Importantly, the Pincus exponents  $\theta$  and  $t$  in Eq. (6) are both related to the scaling exponent  $\rho$  as:

$$\theta = \frac{1}{\rho} - 1, \quad (9)$$

$$t = \frac{1}{1 - \rho}. \quad (10)$$

In a similar fashion, it can be shown<sup>41</sup> that the probability distribution of branch sizes  $p(N_{\text{br}})$  in a randomly branching tree of total size  $N$  is accurately described by the Kramers-like<sup>29,30</sup> formula:

$$p(N_{\text{br}}) = \frac{\mathcal{Z}_{N_{\text{br}}} \mathcal{Z}_{N-N_{\text{br}}-1}}{\sum_{N_{\text{br}}=0}^{N-1} \mathcal{Z}_{N_{\text{br}}} \mathcal{Z}_{N-N_{\text{br}}-1}}, \quad (11)$$

where  $\mathcal{Z}_{N_{\text{br}}}$  is the total number (i.e., the partition function) of branched polymers with  $N_{\text{br}}$  edges. Next, we can take into account that the denominator in Eq. (11) is just  $\mathcal{Z}_N$ , which in general scales as  $\mathcal{Z}_N \sim c^N / N^{2-\varepsilon}$  for branching polymers<sup>41</sup>, where  $c$  is a numerical prefactor related to the tree branching probability per node (or, equivalently, the fraction of the tree branching points). In this way, Eq. (11) takes on the simple scaling form:

$$\frac{p(N_{\text{br}})}{N^{2-\varepsilon}} \simeq c^{-1} (N_{\text{br}} (N - N_{\text{br}} - 1))^{-(2-\varepsilon)}. \quad (12)$$

Expanding on the approach by Rosa and Everaers<sup>41</sup>, we tentatively set an ansatz for the partition function  $\mathcal{Z}_{N_{\text{br}}}$  to

$$\mathcal{Z}_{N_{\text{br}}} = \frac{I_\beta(2\lambda N_{\text{br}})}{(\lambda N_{\text{br}})^\beta}, \quad (13)$$

where  $I_\beta$  is the first modified Bessel function of order  $\beta = 3/2 - \varepsilon$ . This allows us to fit the analytical distribution  $p(N_{\text{br}})$  to the entire range  $0 \leq N_{\text{br}} \leq N/2$ . The justification of the ansatz in Eq. (13) is given in the supplementary material, where we demonstrate that it correctly describes the scaling behaviour of branched polymers as given by Eq. (12).

The *analytical* functions for the distributions of the path lengths [Eq. (5)] and branch sizes [Eq. (11)] have the advantage that we can use them to determine the scaling exponents  $\rho$  and  $\varepsilon$ , respectively, even for polymers whose length dependence of observables is not known or cannot be easily obtained—such as is the case of biological RNAs, whose length is often constrained to a narrow range within a particular type, function, or species. Due to the finiteness of the polymers, these distributions in general show finite-size effects which disappear in the long-chain limit, as we shall also see in Sec. IV C.

### III. METHODS

#### A. RNA folding

When an RNA molecule folds, it often does not adopt a single, well-defined conformation<sup>43</sup>. Therefore, it is more accurate to discuss an ensemble of RNA conformations instead of a

single structure. We generate thermal ensembles ( $T = 37^\circ \text{C}$ ) of RNA secondary structures using the RNAsubopt routine from the ViennaRNA software (v2.4.14)<sup>44</sup> with default settings. The secondary structure prediction does not include pseudoknots, which is a typical simplification that allows us to study the structures of even very long RNAs ( $\sim 10^4$  nt). For each RNA sequence, we generate 500 secondary structure folds, sampled from the thermal ensemble. We denote any quantity  $\mathcal{O}$  averaged over this thermal ensemble of structures by  $\langle \mathcal{O} \rangle$ .

#### B. Tree representation of RNA secondary structure

Each RNA secondary structure without pseudoknots can be mapped to a planar tree graph. The simplest way to construct such a tree is by mapping double-stranded regions (base pairs) to *edges* with weights  $w$  corresponding to the stem lengths, while single-stranded regions (unpaired nucleotides) are mapped to *nodes* connecting the edges<sup>21</sup>. (Note that this procedure differs slightly from the RNA-as-graph method<sup>20</sup> in order to avoid disconnected graphs which are sometimes produced by it.) The resulting number of edges  $N$  of the RNA tree corresponds to the number of bonds in a branched polymer.

We also produce what we term *expanded RNA trees*, where each edge of weight  $w$  is expanded to form  $w$  unit edges with

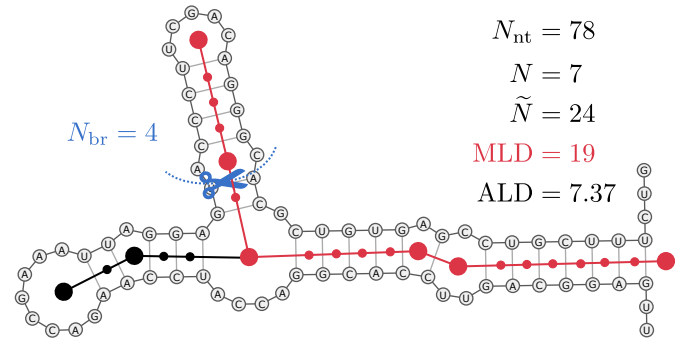


Figure 1. Tree representation of an example RNA secondary structure obtained for a random RNA sequence of length  $N_{\text{nt}} = 78$ . Single-stranded regions are mapped to nodes of a tree (large dots) while double-stranded regions are mapped to edges (lines), weighted with the length of the double-stranded region. Expanded tree produces additional nodes (small dots) between the nodes of the original tree corresponding to the weight of the edge, so that all the edges in the expanded tree have unit weight. Dotted line shows an example of removing one edge of the expanded tree to determine its branch weight, given by the weight of the smaller of the two resulting trees. The sketch also shows some essential properties of the resulting RNA tree, such as the number of edges in the original ( $N$ ) and the expanded ( $\tilde{N}$ ) tree, and the MLD and the ALD of the tree.



$w - 1$  nodes of degree 2 connecting them (Fig. 1). In this way, we obtain a tree with  $\tilde{N}$  edges where each edge has unit weight, which allows us to use the percolation algorithm from Ref. 41 to obtain the branch weight distributions. Furthermore, since it turns out that the RNA nucleotide sequence length  $N_{\text{nt}}$  perfectly correlates with both the number of nodes and the number of expanded nodes of the tree representation of its structure (Fig. S1 in the supplementary material), we use these quantities interchangeably throughout this work.

We also utilize Prüfer sequence representation of RNA trees, which gives a one-to-one correspondence between a tree with  $N$  edges and a sequence of  $N - 1$  integers (for details, see Ref. 45). Random permutations of the Prüfer sequence can then be used to yield trees with node degree distribution identical to the original tree but with different branching patterns.

### C. Multiloop energy model

While the occurrence of multiloops of degree 10 or higher is not uncommon in various RNAs<sup>5</sup>, multiloop energies are the least accurately known among the numerous energy parameters involved in RNA secondary structure prediction<sup>46</sup>. Most of the current energy-based structure prediction software—including ViennaRNA—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

$$E_{\text{multiloop}} = E_0 + E_{\text{br}} \times [\text{branches}] + E_{\text{un}} \times [\text{unpaired nucleotides}], \quad (14)$$

where  $E_0$  is the energy contribution for multiloop initiation, and  $E_{\text{br}}$  and  $E_{\text{un}}$  are the energy contributions for each enclosed base pair and unpaired nucleotide, respectively. Notable difference between different proposed energy parameters lies not only in the magnitude but in the *sign* of the parameter  $E_{\text{br}}$  which controls the number of branches stemming from the multiloop (see Ref. 21 for a comparison of multiloop energy parameters used in different energy models). Earlier versions of ViennaRNA (until v2.0), for instance, used a positive value of this parameter, penalizing high-degree nodes, while the later versions of the software use a negative value, promoting high-degree nodes.

Differences in the multiloop energy parameters can of course reflect in the predicted structures of long RNAs and their topological properties<sup>47,48</sup>. This needs to be taken into account when comparing results obtained by existing studies on the branching properties of RNA<sup>8,9,17,22</sup> that use different versions of folding software and thus potentially different

energy models. To verify how the choice of multiloop energy parameters influences the scaling properties of RNA in the cleanest fashion, we modify *solely* the multiloop parameters in the current parameter set used by ViennaRNA (v2.4:  $E_0 = 9.3$ ,  $E_{\text{un}} = 0.0$ ,  $E_{\text{br}} = -0.9$ ) with the ones from older versions ( $< v2.0$ :  $E_0 = 3.4$ ,  $E_{\text{un}} = 0.0$ ,  $E_{\text{br}} = 0.4$ ), resulting in a modified set of energy parameters which we denote by ViennaRNA-mod.

### D. Random RNA sequences

Nucleotide frequency  $f(b)$ ,  $b \in \{A, C, G, U\}$ , is the simplest statistical property of the primary sequence of RNA, which can nonetheless have a decisive influence on its secondary and tertiary structure. We define a uniformly random RNA sequence as one where all nucleotides are represented equally,  $f(b) = 1/4 \forall b$ . Deviations from this uniform composition can be then evaluated by the Euclidean distance  $\delta^2 = \sum_{b \in \{A, C, G, U\}} (f(b) - 1/4)^2$ , which correlates well with improved statistical measures such as the Jensen-Shannon divergence.

Since the nucleotide composition of biological RNAs often differs from the uniform one, we generate not only uniformly random RNA sequences but also random RNA sequences with different nucleotide frequencies. Specifically, we chose 16 different nucleotide compositions which cover the convex hull of the space of  $\sim 1800$  viral genomes of different positive single-stranded RNA viruses (see Table S1 in the supplementary material) and as such represent the most extreme cases within this dataset. Random RNA sequences are obtained by first generating a sequence with a desired nucleotide composition and length and then randomly shuffling it. In general, we consider random RNA sequences from 100 to 13500 nt in length, where we generate for each random RNA of a given length and composition 200 different random shuffles.

As an additional point of comparison, we also generate Prüfer-shuffled versions of random RNA trees, which preserve the node degree distribution of the original tree. For each RNA length, we select 10 different random sequences, and from each of their 200 secondary structures and resulting trees generate 500 random permutations of their Prüfer sequence.

### E. Scaling exponents

We determine the two scaling exponents  $\rho$  and  $\varepsilon$  introduced in Sec. II in two ways. Firstly, we determine the scaling of the ensemble-averaged topological properties (such as  $\langle \text{ALD} \rangle$ ,  $\langle \text{MLD} \rangle$ , and  $\langle N_{\text{br}} \rangle$ ) with sequence length  $N_{\text{nt}}$ . Coefficients of the scaling power laws of the form  $y = ax^\gamma$  can then be obtained with linear regression on logarithmically transformed

data. Where necessary to differentiate these exponents, we label them as  $\rho_{\langle \text{ALD} \rangle}$ ,  $\rho_{\langle \text{MLD} \rangle}$ , and  $\varepsilon_{\langle N_{\text{br}} \rangle}$ .

Secondly, we determine the scaling exponents  $\rho$  and  $\varepsilon$  also from the distributions of path lengths [Eq. (5)] and branch weights [Eq. (12)] in the tree representation of their secondary structure, respectively. Here, regression is inappropriate as it can in general give a biased estimate for the scaling exponents<sup>49</sup>. We therefore use the maximum likelihood estimator for the scaling exponents. We also note that the statistical error of this approach becomes negligibly small as the number of observations increases, as is also the case with our data. Where necessary, we label the exponents obtained in this way as  $\rho_\theta$ ,  $\rho_t$ , and  $\varepsilon_p$  to highlight that they were obtained from Eqs. (9), (10), or (12), respectively.

## IV. RESULTS

### A. Scaling exponents of random RNA sequences of varying lengths

First, we take a look at how the scaling exponents  $\rho$  and  $\varepsilon$  can be obtained from the scaling of  $\langle \text{ALD} \rangle$  and  $\langle N_{\text{br}} \rangle$  with RNA sequence length  $N_{\text{nt}}$ , respectively. The scaling over a range of sequence lengths for uniformly random RNAs ( $f(b) = 0.25$  for all  $b \in \{A, C, G, U\}$ ) is shown in Fig. 2, panels (a) and (c). While we consider a wide range of RNA sequence lengths ( $\sim 10^2$ – $10^4$  nt) to obtain the two scaling exponents, they are nonetheless *asymptotic* properties of the polymer (RNA tree) size. As such, their fitted values should become more reliable as the sequence length increases. This can be observed in the insets of the two panels, which show how the values of the exponents change as we change the length range of the sequences contributing to the fit, limiting the range to ever longer sequences. The exponent  $\rho_{\langle \text{ALD} \rangle}$  quickly assumes values of around 0.66, while the exponent  $\varepsilon_{\langle N_{\text{br}} \rangle}$  more gradually approaches similar values of around 0.68 as shorter sequences are removed from the fit. In Sec. IV D, we further compare these values with each other as well as with the values of scaling exponents obtained from individual distributions (Sec. IV C).

As the insets of panels (a) and (c) of Fig. 2 show, the range of the fit can clearly influence the value of the scaling exponent; in particular, fitting the exponents to small tree sizes ( $N_{\text{nt}} \sim 10^2$ ) is not necessarily warranted. However, the values of the scaling exponents tend, for the most part, to remain similar within the error of the fit regardless of the fit range, and any differences start to become negligible as shorter sequences are taken out of consideration. Sequences of length  $\gtrsim 1000$  nt should thus already suffice to determine the scaling exponents.

### B. Robustness of the scaling exponents

Now that we have determined the scaling exponents for the scaling behaviour of  $\langle \text{ALD} \rangle$  and  $\langle N_{\text{br}} \rangle$ , we wish to examine how robust these exponents are to changes in various quantities that might influence them. Specifically, we focus on the role of (1) nucleotide composition of random RNA sequences, (2) multiloop energy parameters of the secondary structure prediction software, and (3) node degree distribution of RNA trees.

#### 1. Nucleotide composition

Nucleotide content can vary significantly between different biological RNAs, and has been shown to both play an important role in some of their functions as well as influence the resulting RNA structures. While our analysis primarily focuses on uniformly random RNAs to make the interpretation of our results easier, we need to verify that our conclusions remain valid also for random RNAs of different nucleotide compositions. To this purpose, we have also obtained the scaling exponents  $\rho$  and  $\varepsilon$  from scaling relationships for  $\langle \text{ALD} \rangle$  and  $\langle N_{\text{br}} \rangle$  shown in panels (a) and (c) of Fig. 2 for 16 different nucleotide compositions (see Sec. III and the supplementary material). The results, shown in Figs. S2 and S3 in the supplementary material, demonstrate that any observed differences in the scaling relationship are only due to changes in the prefactor of the scaling law. The scaling exponents  $\rho$  and  $\varepsilon$ , on the other hand, remain essentially the same within the error of the fit, no matter what the nucleotide composition of random RNA is. This also means that the observations from our analysis of uniformly random RNAs can be generalized to random RNA sequences with different nucleotide compositions.

#### 2. Multiloop energy parameters

Energy-based RNA secondary structure prediction depends on the accuracy of the energy parameters that are provided as an input to the folding algorithm. While the two versions of energy parameters provided by Turner and Mathews<sup>50</sup>, Turner1999 and Turner2004, form the basis for the most commonly used secondary structure prediction software such as ViennaRNA<sup>44</sup> and RNAstructure<sup>51</sup>, several improvements have also been suggested<sup>52,53</sup>. These different sets of energy parameters noticeably differ in the energy of multiloop formation, something which is particularly relevant for large, highly branching RNA structures.

We have previously shown<sup>21</sup> that replacing the multiloop energy parameters of the current versions of ViennaRNA (v2.0 and higher) with those of its older versions results in a sig-

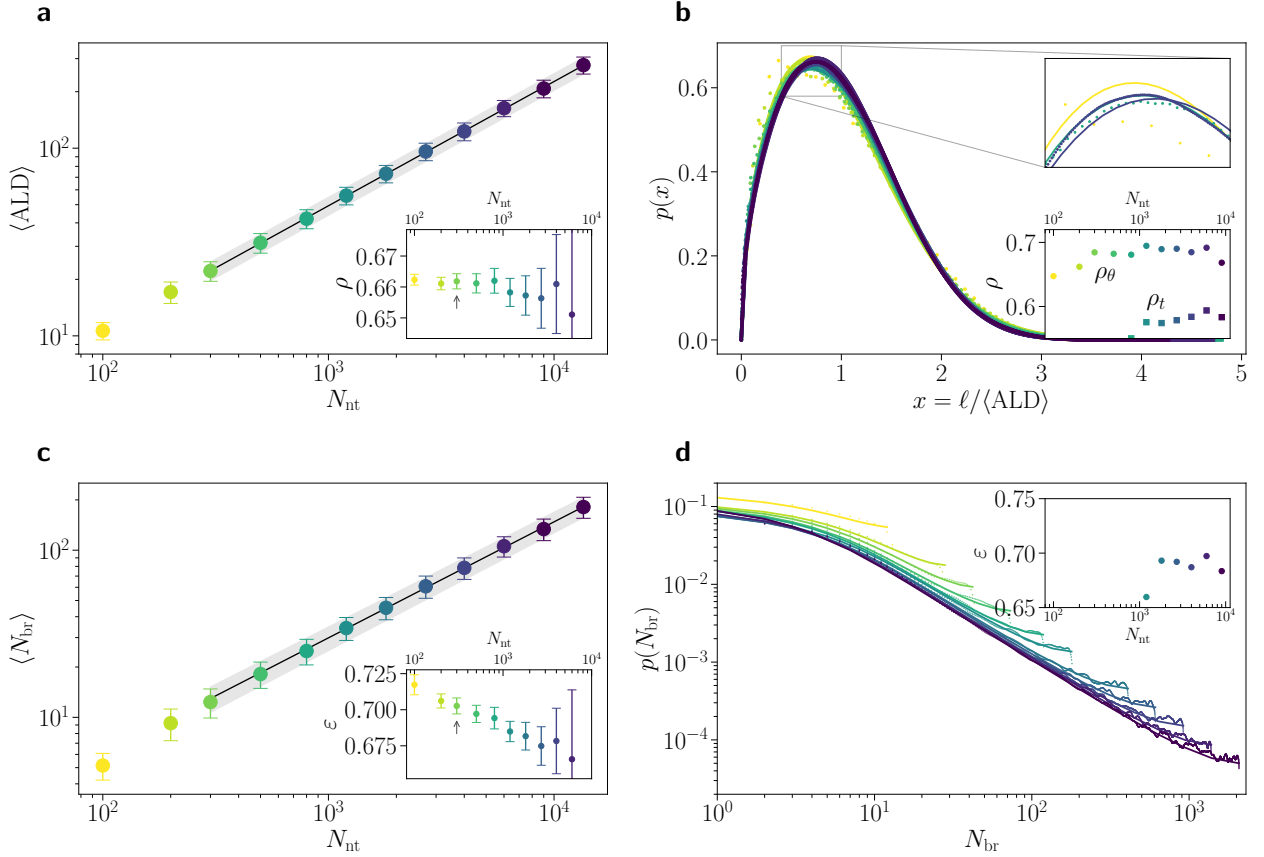


Figure 2. Scaling exponent  $\rho$  for random RNA sequences, as obtained (a) from the scaling of  $\langle \text{ALD} \rangle$  with sequence length  $N_{\text{nt}}$  and (b) from the distributions of scaled path lengths  $p(x)$ , where  $x = \ell / \langle \text{ALD} \rangle$ . Scaling exponent  $\epsilon$  for uniformly random RNA sequences, as obtained (c) from the scaling of  $\langle N_{\text{br}} \rangle$  with sequence length  $N_{\text{nt}}$  and (d) from the distributions of branch weights  $p(N_{\text{br}})$ . Insets in panels (a) and (c) show the scaling exponents obtained by fits to the data with different starting points, with the gray arrows denoting the fits over the shaded regions in the two panels. Insets in panels (b) and (c) show the scaling exponents obtained by fits to distributions at fixed RNA lengths. Colour gradients in panels (b) and (d) and their insets correspond to different sequence lengths from panels (a) and (c). Each point in panels (a) and (c) and their corresponding line in panels (b) and (d) correspond to the average over 200 random sequences with 500 secondary structure thermal ensemble folds per each random sequence.

nificantly different node degree distributions of the resulting RNA structures. In particular, predictions made with the older version of multiloop energy parameters result in RNA structures with a much lower amount of nodes with a high degree of branching ( $> 4$ ) compared to the newer version. This is due to the energy parameter of multiloop branch formation, disfavoured in the older version of energy parameters and promoted in the newer version (see also Table S2 in the supplementary material).

However, when we compare the scaling of the  $\langle \text{ALD} \rangle$  and  $\langle N_{\text{br}} \rangle$  with the RNA sequence length for structures predicted using two different sets of multiloop energy parameters

(Sec. III), we observe that the scaling exponents  $\rho$  and  $\epsilon$  remain the same (Fig. S4 in the supplementary material). Interestingly, this implies that the scaling of RNA as a branched polymer is not strongly dependent on the particularities of the node degree distribution of its structure. This is in line with previous observations, which have found that different RNA folding algorithms and parameter sets can have a strong influence on the details of the predicted structures, but that the structure statistics were much less sensitive to them<sup>54</sup>.

### 3. Randomly shuffled tree topologies

Planar tree graphs can be represented by a Prüfer sequence, which has a one-to-one correspondence to a particular tree topology. By (randomly) permuting this sequence, one obtains trees with a node degree distribution identical to the original one, but with in general a different branching pattern. This approach has previously been used by<sup>45</sup> to study the size of ideal randomly branching polymers, where they have observed that “RNA-like” trees exhibit different scaling than random Prüfer sequences. In Figs. S5 and S6 in the supplementary material we show that Prüfer-shuffling RNA trees originating from uniformly random RNA sequences indeed leads to a decrease in both scaling exponents  $\rho$  and  $\varepsilon$ . The exponents of Prüfer-shuffled RNA trees,  $\rho \approx 0.53$  and  $\varepsilon \approx 0.59$ , respectively, are indistinguishable from those of random Prüfer sequences corresponding to random trees of comparable size. This shows that random trees with the same node degree distribution as random RNAs but different branching patterns scale in a significantly different fashion. Consequently, the pattern of node degree distribution alone does not suffice to explain the observed scaling relationships of RNA secondary structures. What is more, the Prüfer-shuffled RNA trees appear to have even smaller size than what would be expected for a random Prüfer sequence or even a random unlabelled tree with the same number of nodes.

### C. Scaling exponents of random RNA sequences of fixed length

Since the length of biological RNAs—unlike that of random RNA sequences—in general cannot be varied arbitrarily, we next demonstrate how the two scaling exponents  $\rho$  and  $\varepsilon$  can be determined from the distributions of path lengths  $p(\ell)$  [Eq. (5)] and branch weights  $p(N_{\text{br}})$  [Eq. (12)] in uniformly random RNA sequences of *fixed length*. These two distributions are shown in panels (b) and (d) of Fig. 2, respectively, each line representing an average over 200 different random RNA sequences of the same length, which reduces the amount of noise in the data. In terms of biological RNAs, this could correspond to an average over a related group of RNAs of similar length, either by their function or by evolutionary relatedness (species, genus, ...). We remark that this approach does not necessarily require all of the RNAs to have exactly the same length  $N_{\text{nt}}$ , as during the mapping of RNA structure to a tree, there is already some variation in the resulting number of nodes  $N$  stemming from the thermal ensemble of structures.

In general, the fits of the distributions improve as the length of the RNA increases. This is very clearly observed in the case of the distribution of branch weights, where the scaling exponent  $\varepsilon$  acquires “physical” values ( $\varepsilon \geq 1/2$ ) only for RNA

lengths above  $\gtrsim 2000$  nt (Fig. 2d). In the case of the RdC distributions of path lengths, we can on the other hand observe that the scaling exponent  $\rho$  attains a different value depending on whether it is obtained from  $\theta$  or  $t$  [Eqs. (9) and (10)]—both parameters of the RdC distribution (Fig. 2b; see also Fig. S7 in the supplementary material). Under certain assumptions, to which we shall return in Sec. V, one would expect that the two coefficients would be connected through the relation  $\theta = 1/(1-t)$  and thus yield the same prediction for  $\rho$ . However, this is often not true even in the case of other branching polymers, and as Fig. S7c also shows, certainly does not hold for random RNA sequences. This is consequently reflected in the two different predictions of  $\rho$  as obtained from either  $t$  or  $\theta$  (Fig. 2b).

### D. RNA as a randomly branching polymer

In the previous sections, we have shown how to obtain the scaling exponents  $\rho$  and  $\varepsilon$  of the secondary structure of random RNA sequences in two different ways. When the sequences span a large range of lengths, the scaling exponents can be obtained from the scaling of their  $\langle \text{ALD} \rangle$  (or  $\langle \text{MLD} \rangle$ ) and  $\langle N_{\text{br}} \rangle$  with sequence length  $N_{\text{nt}}$  or, equivalently, number of nodes  $N$ . This leads to the scaling exponents  $\rho_{\langle \text{ALD} \rangle}$  (or  $\rho_{\langle \text{MLD} \rangle}$ ) and  $\varepsilon_{\langle N_{\text{br}} \rangle}$ , respectively (Sec. IV A). When the RNA sequence length cannot be varied arbitrarily, we have also shown how to obtain the same exponents through the distributions of path lengths  $p(\ell)$  and branch weights  $p(N_{\text{br}})$  (Sec. IV C). Since the distribution of path lengths is in general two-parametric [Eq. (5)], this leads to two separate estimates for  $\rho$ , namely  $\rho_t$  and  $\rho_\theta$ , while the distribution of branch

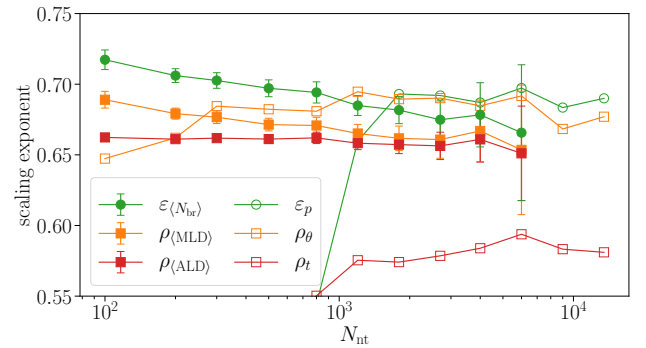


Figure 3. Comparison of the scaling exponents  $\varepsilon$  and  $\rho$  for uniformly random RNA sequences, obtained either from scaling properties (full symbols; panels (a) and (c) of Fig. 2) or through fits to distributions at fixed RNA length (empty symbols; panels (b) and (d) of Fig. 2).



weights gives us the exponent  $\varepsilon_p$ . Now we are left to examine what these scaling exponents obtained for random RNA sequences imply for the structural properties of RNA in the wider context of randomly branching polymers.

As described in Sec. II A, one would expect for randomly branching polymers that the two topological scaling exponents  $\rho$  and  $\varepsilon$  will be identical,  $\rho = \varepsilon$ . The scaling exponents for uniformly random RNA sequences, obtained either from scaling relationships or distributions at fixed sequence length (Fig. 2) are shown together in Fig. 3. We immediately see that the exponents  $\varepsilon$  and  $\rho$  obtained from the scaling relationships indeed seem to converge to the same value, at least within the error bars of the fit. (Note that the sequence length in this case refers to the *starting sequence length* of the fit range.) When we take a look at the scaling exponents obtained from the distributions of RNA sequences of fixed length, we see that the exponents  $\varepsilon$  and  $\rho_\theta$  also converge to approximately the same value. This value is, furthermore, very close to the one obtained from the scaling relationships, and we can claim that we have in general

$$\rho_{\text{RNA}} \simeq \varepsilon_{\text{RNA}} \approx 0.67. \quad (15)$$

(We remark again that for the fits to distributions, the statistical error is negligible, while the systematic error is difficult to estimate.)

A notable discrepancy, however, occurs with  $\rho_t$ , that is, the exponent  $\rho$  obtained from the fit parameter  $t$  of the RdC distribution. The value of this scaling exponent is significantly smaller than all the other scaling exponents, even though it already appears to have converged. While it is not clear why this particular exponent does not match the other ones, we can speculate that the reason behind it lies in the assumptions made during the derivation of the RdC curve. Specifically, to derive the relationship between  $\rho_\theta$  and  $\rho_t$ , one needs to assume that the branched polymer is self-similar (fractal). Our results indicate that while RNA can be treated as a randomly branching polymer, certain assumptions regarding its self-similarity might not hold. Apart from the exponent  $\rho_t$ , however, the scaling exponents  $\rho_{\langle \text{ALD} \rangle}$ ,  $\varepsilon_{\langle N_{\text{br}} \rangle}$ ,  $\rho_\theta$ , and  $\varepsilon$  appear to roughly obey the relationship  $\rho = \varepsilon$ . In this respect, RNA indeed behaves as a randomly branching polymer.

We can use the obtained scaling exponents to compare random RNA sequences to other types of branched polymers. Fig. 4 frames our result,  $\rho \simeq \varepsilon \approx 0.67$ , in the context of ideal linear and branched polymers, as well as branched self-avoiding walks (SAWs)<sup>31</sup> and melts of branching polymers<sup>40</sup> in 2D and 3D. While we have trivially  $\rho = \varepsilon = 1$  for ideal linear polymers and  $\rho = \varepsilon = 0.5$  for ideal branched polymers, the scaling exponents of random RNAs take on values which are very similar to those of branched SAWs in 3D. This is a particularly interesting observation since the RNA secondary structures we study are obtained using an energy-based prediction software which does not take into account any steric

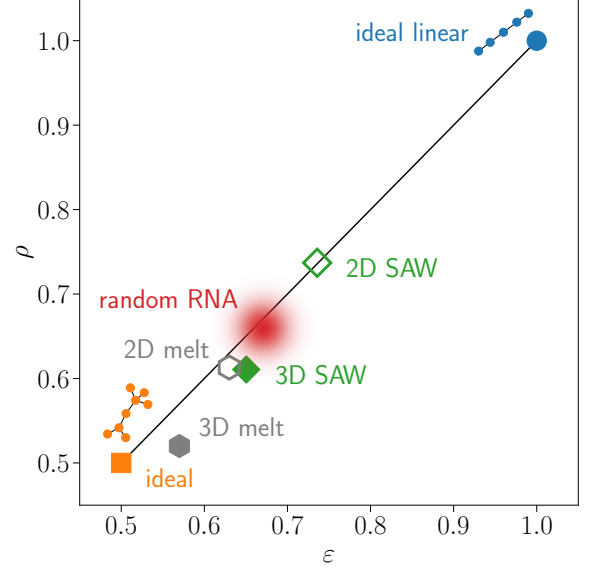


Figure 4. Scaling exponents  $\rho$  and  $\varepsilon$  for different types of branched polymers. Shown are the exact results for ideal branching and linear polymers, as well as the known computer simulation results for branching SAWs<sup>31</sup> and melts of randomly branching polymers<sup>40</sup> in both 2D and 3D. Positioned in this diagram is also our result on random RNA sequences, where the smeared region depicts the estimated value of  $\rho \simeq \varepsilon \approx 0.67$  and the corresponding uncertainty (cf. Fig. 3).

effects or tertiary interactions.

## V. DISCUSSION

As already briefly mentioned in Sec. II, the study of the connectivity of branched polymers provides an “incomplete” view of RNA conformations. The main reason is that—just as in the case of the more ordinary linear polymers<sup>26</sup>—how branched conformations fold *in space* is a matter of a non-trivial combination between branching and the specific interactions between different monomers.

In polymer physics<sup>26</sup>, the physical characterization of the spatial conformations of any polymer ensemble can be obtained in terms of the average linear size  $\langle R(N) \rangle$  defined through the root-mean-square gyration radius  $R_g$ :

$$\langle R(N) \rangle \equiv \sqrt{\langle R_g^2 \rangle} = \sqrt{\left\langle \frac{1}{N} \sum_{i=1}^N (\mathbf{r}_i - \mathbf{r}_{\text{cm}})^2 \right\rangle} \sim N^\nu, \quad (16)$$

where  $\mathbf{r}_i$  is the spatial coordinate of the  $i$ -th monomer and  $\mathbf{r}_{\text{cm}} \equiv N^{-1} \sum_{i=1}^N \mathbf{r}_i$  is the centre-of-mass of the chain. The scal-

ing exponent  $\nu$  introduced in Eq. (16) specifies the embedding of spatial conformations *in space* and depends on several factors, particularly on monomer-monomer interactions and solvent conditions<sup>21,24</sup>.

Contrary to what we have done for the scaling exponents  $\rho$  and  $\varepsilon$ , the determination of the scaling exponent  $\nu$  for RNA molecules is much less straightforward. The main reason, evident from Eq. (16), is that one needs to know with great accuracy the spatial positions of the atoms for an *ensemble* of RNA molecules at various given values of  $N$ , which is not an easy task. In general, the possible number of RNA conformations reconstructed from experiments (for instance, NMR experiments) and computational simulations is still limited. Fortunately, the theory of randomly branching polymers in the form of the classical Flory theory<sup>23</sup> comes to our aid. We now show that, under rather general assumptions, the knowledge of  $\rho$  *implies*  $\nu$ <sup>21</sup>, meaning that the branched architecture of RNA molecules and their average folding properties in space “determine” each other.

The Flory free energy of a randomly branching polymer or a tree with  $N$  monomers is a function of the polymer mean size  $\langle R \rangle$  and the average ladder distance  $\langle \text{ALD} \rangle$ <sup>25,55</sup>:

$$\frac{\mathcal{F}}{k_B T} \equiv \frac{\mathcal{F}(N; \langle R \rangle, \langle \text{ALD} \rangle)}{k_B T} \simeq \frac{\langle R \rangle^2}{\langle \text{ALD} \rangle^2} + \frac{\langle \text{ALD} \rangle^2}{b^2 N} + V(N; \langle R \rangle), \quad (17)$$

where  $k_B$  is the Boltzmann constant,  $T$  is the temperature,  $b$  is the mean bond length and  $V(N; \langle R \rangle)$  is the general interaction potential between monomers which, importantly, depends only on  $N$  and  $\langle R \rangle$  but *not* on  $\langle \text{ALD} \rangle$ . The sign “ $\simeq$ ” in Eq. (17) reminds us that the different terms in the expression are typically valid up to some numerical prefactor of  $o(1)$ <sup>23</sup>. When we minimize the Flory energy with respect to  $\langle \text{ALD} \rangle$ <sup>23–25</sup>, the exact form of  $V$  does not matter and we easily obtain the expression connecting  $\nu$  and  $\rho$ <sup>25</sup>:

$$\nu_{\text{Flory}} = \frac{3\rho - 1}{2}. \quad (18)$$

As demonstrated in Ref. 25, Eq. (18) is in excellent agreement with all available numerical data for  $\nu$  and  $\rho$  in different ensembles of randomly branching polymers, both in 2D and 3D.

A strong signature for annealed connectivity of the branching topology of a molecule is the equivalence between the exponents  $\rho$  and  $\varepsilon$  [Eq. (4)], for which we have demonstrated that it holds true for RNA molecules as well (Figs. 3 and 4). Taking all these considerations together, we propose that Eq. (18) is applicable to RNA as well, and therefore we get using our estimated value of  $\rho \simeq 2/3$  (Fig. 3)

$$\nu_{\text{RNA}} \simeq 1/2. \quad (19)$$

This is, interestingly, the same value of the scaling exponent for single randomly branching self-avoiding polymers<sup>56</sup> in 3D.

We also remark that Eq. (19) contradicts the proposal by Fang *et al.*<sup>28</sup>, which was based on the Kramers’ formula<sup>29,30</sup> for the radius of gyration of branching trees:

$$\langle R_g^2 \rangle = \frac{\langle b^2 \rangle}{N} \frac{\sum_{N_{\text{br}}=0}^{N-1} N_{\text{br}} (N - N_{\text{br}}) \mathcal{Z}_{N_{\text{br}}} \mathcal{Z}_{N-N_{\text{br}}}}{\sum_{N_{\text{br}}=0}^{N-1} \mathcal{Z}_{N_{\text{br}}} \mathcal{Z}_{N-N_{\text{br}}}} \simeq \langle b^2 \rangle N^\rho, \quad (20)$$

where the last expression can be derived by using Eqs. (11)–(13) with  $\beta = 3/2 - \varepsilon = 3/2 - \rho$  (see supplementary material). Here,  $\langle b^2 \rangle$  is the mean-square bond spatial distance of the branched RNA structure (Fig. 1). By neglecting excluded volume interactions,  $\langle b^2 \rangle$  does not depend on  $N$ <sup>57</sup> and one would get  $\nu = \nu_{\text{RNA}} = \rho/2 \simeq 1/3$  which corresponds to the prediction by Fang *et al.*<sup>28</sup>. However, this hypothesis appears inconsistent with the measured  $\rho \simeq 2/3$  as branching polymers with no volume interactions have  $\rho = \varepsilon = 1/2$ <sup>25,26</sup>. Since the Kramers’ formula holds for both RNA molecules [Eqs. (11)–(13)] and generic branching polymers<sup>25</sup> alike, Eq. (20) can be reconciled with it by noticing<sup>57</sup> that  $\langle b^2 \rangle \sim N^{2\nu-\rho}$ . However, since this makes Eq. (20) identically true, it means that the same equation cannot be used to derive  $\nu$  from  $\rho$ .

Finally, we note that our results for  $\rho$  and  $\varepsilon$  are not necessarily in contradiction with the works where  $\langle R_g^2 \rangle$  as a function of sequence length was determined from 3D structures of RNA molecules deposited in PDB and which led to the estimate  $\nu \approx 1/3$ <sup>58,59</sup>. As described in Ref. 25, branching polymers prepared under specific solvent conditions (for instance, in a good solvent) can be moved—by keeping the connectivity quenched!—to different solvent conditions (for instance, to a bad one) which changes the scaling exponent  $\nu$  but not, of course,  $\rho$ . A similar process may occur during the determination of the structure of RNA molecules as well, and could be a reason for the observed discrepancy in the value of the scaling exponent  $\nu$ . Other potential reasons are also that the connectivity of RNAs in the PDB dataset might be quenched, or that the size distributions of these molecules is skewed towards short RNAs to the extent that one cannot observe the proper scaling exponents, which should be attainable only in the asymptotic limit of very long RNAs.

## VI. CONCLUSIONS

In this manuscript, we used the scaling theory of branching polymers (Sec. II) to characterize the branching properties of ensembles of secondary structure folds of random RNAs of varying lengths and uniform composition. By ignoring pseudoknots we were able to map RNA folds to tree structures (Sec. III) and obtain the scaling exponents  $\rho$  and

$\varepsilon$  from the length dependence of  $\langle \text{ALD} \rangle$  and  $\langle N_{\text{br}} \rangle$ , respectively (Sec. IV A). Importantly, we also demonstrated how the two exponents can be determined from the distributions of path lengths and branch weights of RNA sequences with fixed length (Sec. IV C), which is particularly relevant for biological RNAs. We found that  $\rho \simeq \varepsilon \approx 0.67$ , indicating that the ensembles of RNA secondary structure folds are characterized by annealed random branching and behave as branched self-avoiding walks in 3D (Sec. IV D). Furthermore, we have shown that this result is robust irrespective of nucleotide composition, node degree distribution, as well as multiloop energy parameters (Sec. IV B). Our characterization of the branched topology of RNA folds, complemented by very general polymer arguments, also allowed us to determine the scaling exponent  $\nu$  for the mean spatial size of RNA molecules, where we obtained  $\nu_{\text{RNA}} = 1/2$  (Sec. V). Simultaneously, we observed that other assessments of this exponent (as done by, e.g., Fang *et al.*<sup>28</sup>) implicitly assume a supplement of non-trivial hypotheses which do not appear to be justified based on our analysis.

Our work firmly places the branching (secondary) structure of RNA in the wider context of randomly branching polymers. In the future, the methodology developed here should be applied to biological RNAs, with viral ssRNA genomes providing a particularly good example, as the scaling exponents can be compared at the level of the viral species, genus, and family. The robustness of the scaling exponents should also be further checked by incorporating experimental reactivity data (such as SHAPE and DMS) into the prediction of RNA secondary structures, and by expanding the tree representation of RNA and the subsequent analysis with the inclusion of pseudoknots. We believe that a better understanding of the principles which give rise to RNA branching properties from its sequence will open up the possibility for the design of RNA sequences with desired topological properties<sup>60–62</sup>, where the ability to design compact RNA sequences is particularly relevant in the design of vaccines<sup>63</sup>.

## SUPPLEMENTARY MATERIAL

See the supplementary material for more details on: the derivation of Eq. (13) in the main text; correlation between the quantities  $N_{\text{nt}}$ ,  $N$ , and  $\tilde{N}$ ; exploration of the connections between nucleotide composition, multiloop energy parameters, and Prüfer shuffling with the scaling exponents  $\rho$  and  $\varepsilon$ ; and the determination of the scaling exponents  $\rho_l$  and  $\rho_\theta$  from the distributions of path lengths.

## ACKNOWLEDGMENTS

We thank Luka Leskovec for helpful discussions. A.B. acknowledges support by Slovenian Research Agency (ARRS) under contract no. P1-0055. L.T. acknowledges financial support from ICSC—Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union—NextGenerationEU. A.R., L.T., and A.B. acknowledge networking support by the COST Action CA17139 (EUTOPIA).

## AUTHOR DECLARATIONS

### Conflict of interest

The authors have no conflicts to disclose.

### Author contributions

Domen Vaupotič: formal analysis; investigation; methodology; visualization; writing – review & editing. Angelo Rosa: conceptualization; investigation; methodology; writing – review & editing. Luca Tubiana: conceptualization; investigation; methodology; writing – review & editing. Anže Božič: conceptualization; investigation; methodology; writing – original draft; writing – review & editing.

## DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available within the article and its supplementary material.

## REFERENCES

- <sup>1</sup>A. Kulkarni and G. Beaucage, “Quantification of branching in disordered materials,” *J. Polym. Sci. B: Polym. Phys.* **44**, 1395–1405 (2006).
- <sup>2</sup>B. I. Voit and A. Lederer, “Hyperbranched and highly branched polymer architectures—synthetic strategies and major characterization aspects,” *Chem. Rev.* **109**, 5924–5973 (2009).
- <sup>3</sup>J. R. Van der Maarel, *Introduction to biopolymer physics* (World Scientific Publishing Company, 2007).
- <sup>4</sup>A. B. Cook and S. Perrier, “Branched and dendritic polymer architectures: functional nanomaterials for therapeutic delivery,” *Adv. Funct. Mater.* **30**, 1901001 (2020).
- <sup>5</sup>J. Wiedemann, J. Kaczor, M. Milostan, T. Zok, J. Blazewicz, M. Szachniuk, and M. Antczak, “Rnaloops: a database of RNA multiloops,” *Bioinformatics* (2022).
- <sup>6</sup>M. A. Boerneke, J. E. Ehrhardt, and K. M. Weeks, “Physical and functional analysis of viral RNA genomes by SHAPE,” *Annu. Rev. Virol.* **6**, 93 (2019).

- <sup>7</sup>H. Wan, R. L. Adams, B. D. Lindenbach, and A. M. Pyle, “The in vivo and in vitro architecture of the Hepatitis C virus RNA genome uncovers functional RNA secondary and tertiary structures,” *J. Virol.* **96**, e01946–21 (2022).
- <sup>8</sup>A. M. Yoffe, P. Prinsen, A. Gopal, C. M. Knobler, W. M. Gelbart, and A. Ben-Shaul, “Predicting the sizes of large RNA molecules,” *Proc. Natl. Acad. Sci. USA* **105**, 16153–16158 (2008).
- <sup>9</sup>L. Tubiana, A. Božič, C. Micheletti, and R. Podgornik, “Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses,” *Biophys. J.* **108**, 194–202 (2015).
- <sup>10</sup>S. W. Singaram, R. F. Garmann, C. M. Knobler, W. M. Gelbart, and A. Ben-Shaul, “Role of RNA branchedness in the competition for viral capsid proteins,” *J. Phys. Chem. B* **119**, 13991–14002 (2015).
- <sup>11</sup>R. F. Garmann, M. Comas-Garcia, C. M. Knobler, and W. M. Gelbart, “Physical principles in the self-assembly of a simple spherical virus,” *Acc. Chem. Res.* **49**, 48–55 (2016).
- <sup>12</sup>C. Beren, L. L. Dreesens, K. N. Liu, C. M. Knobler, and W. M. Gelbart, “The effect of RNA secondary structure on the self-assembly of viral capsids,” *Biophys. J.* **113**, 339–347 (2017).
- <sup>13</sup>A. Božič, C. Micheletti, R. Podgornik, and L. Tubiana, “Compactness of viral genomes: effect of disperse and localized random mutations,” *J. Phys. Condens. Matter* **30**, 084006 (2018).
- <sup>14</sup>L. Marichal, L. Gargowitsch, R. L. Rubim, C. Sizun, K. Kra, S. Bressanelli, Y. Dong, S. Panahandeh, R. Zandi, and G. Tresset, “Relationships between RNA topology and nucleocapsid structure in a model icosahedral virus,” *Biophys. J.* **120**, 3925–3936 (2021).
- <sup>15</sup>G. Erdemci-Tandogan, J. Wagner, P. Van Der Schoot, R. Podgornik, and R. Zandi, “RNA topology remodels electrostatic stabilization of viruses,” *Phys. Rev. E* **89**, 032707 (2014).
- <sup>16</sup>G. Erdemci-Tandogan, J. Wagner, P. van der Schoot, R. Podgornik, and R. Zandi, “Effects of RNA branching on the electrostatic stabilization of viruses,” *Phys. Rev. E* **94**, 022408 (2016).
- <sup>17</sup>A. Gopal, D. E. Egecioglu, A. M. Yoffe, A. Ben-Shaul, A. L. Rao, C. M. Knobler, and W. M. Gelbart, “Viral RNAs are unusually compact,” *PLoS One* **9**, e105875 (2014).
- <sup>18</sup>R. F. Garmann, A. Gopal, S. S. Athavale, C. M. Knobler, W. M. Gelbart, and S. C. Harvey, “Visualizing the global secondary structure of a viral RNA genome with cryo-electron microscopy,” *RNA* **21**, 877–886 (2015).
- <sup>19</sup>H. H. Gan, S. Pasquali, and T. Schlick, “Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design,” *Nucleic Acids Res.* **31**, 2926–2943 (2003).
- <sup>20</sup>T. Schlick, “Adventures with RNA graphs,” *Methods* **143**, 16–33 (2018).
- <sup>21</sup>D. Vaupotič, A. Rosa, R. Podgornik, L. Tubiana, and A. Božič, “Viral RNA as a branched polymer,” (2022), [arXiv:2212.00829 \[physics.bio-ph\]](https://arxiv.org/abs/2212.00829).
- <sup>22</sup>A. Borodavka, S. W. Singaram, P. G. Stockley, W. M. Gelbart, A. Ben-Shaul, and R. Tuma, “Sizes of long RNA molecules are determined by the branching patterns of their secondary structures,” *Biophys. J.* **111**, 2077–2085 (2016).
- <sup>23</sup>P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca (NY), 1953).
- <sup>24</sup>S. M. Bhattacharjee, A. Giacometti, and A. Maritan, “Flory theory for polymers,” *J. Phys. Cond. Matter* **25**, 503101 (2013).
- <sup>25</sup>R. Everaers, A. Y. Grosberg, M. Rubinstein, and A. Rosa, “Flory theory of randomly branched polymers,” *Soft Matter* **13**, 1223–1234 (2017).
- <sup>26</sup>M. Rubinstein and R. H. Colby, *Polymer Physics* (Oxford University Press, New York, 2003).
- <sup>27</sup>Z.-G. Wang, “50th anniversary perspective: Polymer conformation—a pedagogical review,” *Macromolecules* **50**, 9073–9114 (2017).
- <sup>28</sup>L. T. Fang, W. M. Gelbart, and A. Ben-Shaul, “The size of RNA as an ideal branched polymer,” *J. Chem. Phys.* **135**, 10B616 (2011).
- <sup>29</sup>H. A. Kramers, “The behavior of macromolecules in inhomogeneous flow,” *J. Chem. Phys.* **14**, 415–424 (1946).
- <sup>30</sup>M. Daoud and J. F. Joanny, “Conformation of branched polymers,” *J. Physique* **42**, 1359–1371 (1981).
- <sup>31</sup>E. J. Van Rensburg and N. Madras, “A nonlocal monte carlo algorithm for lattice trees,” *J. Phys. A: Math. Theor.* **25**, 303 (1992).
- <sup>32</sup>A. Y. Grosberg and R. Bruinsma, “Confining annealed branched polymers inside spherical capsids,” *J. Biol. Phys.* **44**, 133–145 (2018).
- <sup>33</sup>J. Kelly, A. Y. Grosberg, and R. Bruinsma, “Sequence dependence of viral RNA encapsidation,” *J. Phys. Chem. B* **120**, 6038–6050 (2016).
- <sup>34</sup>P. G. Higgs, “RNA secondary structure: a comparison of real and random sequences,” *J. Phys.*, **13**, 43–59 (1993).
- <sup>35</sup>R. Bundschuh and T. Hwa, “Statistical mechanics of secondary structures formed by random RNA sequences,” *Phys. Rev. E* **65**, 031903 (2002).
- <sup>36</sup>E. A. Schultes, A. Spasic, U. Mohanty, and D. P. Bartel, “Compact and ordered collapse of randomly generated RNA sequences,” *Nat. Struct. Mol. Biol.* **12**, 1130–1136 (2005).
- <sup>37</sup>P. Clote, F. Ferré, E. Kranakis, and D. Krizanc, “Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency,” *RNA* **11**, 578–591 (2005).
- <sup>38</sup>F. Chizzolini, L. F. Passalacqua, M. Oumais, A. I. Dingilian, J. W. Szostak, and A. Luptak, “Large phenotypic enhancement of structured random RNA pools,” *J. Am. Chem. Soc.* **142**, 1941–1951 (2019).
- <sup>39</sup>A. Rosa and R. Everaers, “Computer simulations of randomly branching polymers: annealed versus quenched branching structures,” *J. Phys. A Math. Theor.* **49**, 345001 (2016).
- <sup>40</sup>A. Rosa and R. Everaers, “Computer simulations of melts of randomly branching polymers,” *J. Chem. Phys.* **145**, 164906 (2016).
- <sup>41</sup>A. Rosa and R. Everaers, “Beyond Flory theory: Distribution functions for interacting lattice trees,” *Phys. Rev. E* **95**, 012117 (2017).
- <sup>42</sup>In the original work of Rosa and Everaers<sup>41</sup>, the exponents  $\theta$  and  $t$  and the related numerical constants  $C$  and  $K$  bear the subscript  $\ell$  to distinguish them from analogous quantities appearing in other distribution functions. Here, to lighten up the notation, we drop this subscript.
- <sup>43</sup>C. T. Woods, L. Lackey, B. Williams, N. V. Dokholyan, D. Gotz, and A. Laederach, “Comparative visualization of the RNA suboptimal conformational ensemble in vivo,” *Biophys. J.* **113**, 290–301 (2017).
- <sup>44</sup>R. Lorenz, S. H. Bernhart, C. Höner zu Siederdisen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, “ViennaRNA package 2.0,” *Algorithms Mol. Biol.* **6**, 1–14 (2011).
- <sup>45</sup>S. W. Singaram, A. Gopal, and A. Ben-Shaul, “A Prüfer-sequence based algorithm for calculating the size of ideal randomly branched polymers,” *J. Phys. Chem. B* **120**, 6231–6237 (2016).
- <sup>46</sup>S. Poznanović, C. Wood, M. Cloer, and C. Heitsch, “Improving RNA branching predictions: advances and limitations,” *Genes* **12**, 469 (2021).
- <sup>47</sup>M. Ward, A. Datta, M. Wise, and D. H. Mathews, “Advanced multi-loop algorithms for RNA secondary structure prediction reveal that the simplest model is best,” *Nucleic Acids Res.* **45**, 8541–8550 (2017).
- <sup>48</sup>S. Poznanović, F. Barrera-Cruz, A. Kirkpatrick, M. Ielusic, and C. Heitsch, “The challenge of RNA branching prediction: a parametric analysis of multi-loop initiation under thermodynamic optimization,” *J. Struct. Biol.* **210**, 107475 (2020).
- <sup>49</sup>A. Clauset, C. R. Shalizi, and M. E. Newman, “Power-law distributions in empirical data,” *SIAM Rev.* **51**, 661–703 (2009).
- <sup>50</sup>D. H. Turner and D. H. Mathews, “NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure,” *Nucleic Acids Res.* **38**, D280–D282 (2010).
- <sup>51</sup>D. H. Mathews and D. H. Turner, “Prediction of RNA secondary structure by free energy minimization,” *Curr. Op. Struct. Biol.* **16**, 270–278 (2006).
- <sup>52</sup>M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy, “Computational approaches for RNA energy parameter estimation,” *RNA* **16**, 2304–2318 (2010).
- <sup>53</sup>W. B. Langdon, J. Petke, and R. Lorenz, “Evolving better RNAfold structure prediction,” in *European Conference on Genetic Programming* (2018) pp. 220–236.



<sup>54</sup>M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster, “Algorithm independent properties of RNA secondary structure predictions,” *Eur. Biophys. J.* **25**, 115–130 (1996).

<sup>55</sup>A. Y. Grosberg, “Annealed lattice animal model and Flory theory for the melt of non-concatenated rings: towards the physics of crumpling,” *Soft Matter* **10**, 560–565 (2014).

<sup>56</sup>G. Parisi and N. Sourlas, “Critical behavior of branched polymers and the Lee-Yang edge singularity,” *Phys. Rev. Lett.* **46**, 871–874 (1981).

<sup>57</sup>According to the Kramers’ theorem<sup>26,29,30</sup>, each bond contributes  $\simeq \langle b^2 \rangle$  to the mean-square radius of gyration. In terms of the notation of this paper,  $\langle b^2 \rangle$  can be defined as the ratio between the mean-square end-to-end distance between tree paths with average ladder distance  $\langle \text{ALD} \rangle$  and the last one, i.e.,

$$\begin{aligned} \langle b^2 \rangle &\simeq \frac{\langle R^2(\langle \text{ALD} \rangle) \rangle}{\langle \text{ALD} \rangle / b} \simeq b^2 \frac{(\langle \text{ALD} \rangle / b)^{2\nu_{\text{path}}}}{\langle \text{ALD} \rangle / b} \\ &= b^2 (\langle \text{ALD} \rangle / b)^{2\nu_{\text{path}}-1}, \end{aligned} \quad (21)$$

where  $b$  is the mean bond length as in Eq. (17) and  $\nu_{\text{path}} = \nu/\rho$ <sup>25</sup> is the scaling exponent for the spatial structure of tree linear paths. For non-interacting trees with  $\nu = 1/4$  and  $\rho = 1/2$ <sup>25</sup>,  $\nu_{\text{path}} = 1/2$  and  $\langle b^2 \rangle \simeq b^2$  does not depend on  $N$ . Conversely, for interacting trees we do have  $\langle \text{ALD} \rangle / b \sim N^\rho$

[Eq. (2)] and Eq. (21) then implies that  $\langle b^2 \rangle \simeq b^2 N^{2\nu-\rho}$ .

<sup>58</sup>C. Hyeon, R. I. Dima, and D. Thirumalai, “Size, shape, and flexibility of RNA structures,” *J. Chem. Phys.* **125**, 194905 (2006).

<sup>59</sup>Z.-H. Guo, L. Yuan, Y.-L. Tan, B.-G. Zhang, and Y.-Z. Shi, “RNAStat: An integrated tool for statistical analysis of RNA 3D structures,” *Front. Bioinform.* **1**, 809082 (2022).

<sup>60</sup>S. Jain, A. Laederach, S. B. Ramos, and T. Schlick, “A pipeline for computational design of novel RNA-like topologies,” *Nucleic Acids Res.* **46**, 7040–7051 (2018).

<sup>61</sup>S. Jain, Y. Tao, and T. Schlick, “Inverse folding with RNA-As-Graphs produces a large pool of candidate sequences with target topologies,” *J. Struct. Biol.* **209**, 107438 (2020).

<sup>62</sup>L. Rolband, D. Beasock, Y. Wang, Y.-G. Shu, J. D. Dinman, T. Schlick, Y. Zhou, J. S. Kieft, S.-J. Chen, G. Bussi, *et al.*, “Biomotors, viral assembly, and RNA nanobiotechnology: Current achievements and future directions,” *Comput. Struct. Biotechnol. J.* (2022).

<sup>63</sup>J. M. Herrero, T. Stahl, S. Erbar, K. Maxeiner, A. Schlegel, T. Bacic, L. Cavalcanti, M. Schroer, D. Svergun, U. Sahin, *et al.*, “Ultra-compacted single self-amplifying RNA molecules as quintessential vaccines,” *Research Square* (2022), <https://doi.org/10.21203/rs.3.rs-2142761/v1>.

**SUPPLEMENTARY MATERIAL:**  
**Scaling properties of RNA as a randomly branching polymer**

Domen Vaupotič

*Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia*

Angelo Rosa

*Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy*

Luca Tubiana

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

Anže Božič

*Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia\**

(Dated: March 31, 2023)

---

\* [anze.bozic@ijs.si](mailto:anze.bozic@ijs.si)

# JUSTIFICATION OF THE ANSATZ IN EQUATION (13) IN THE MAIN TEXT

Equation (11) in the main text is exact only in the case of *ideal* branching polymers, i.e., polymers with no monomer-monomer interactions for which the only contribution to the free energy of the system comes from branching. For such a situation, Daoud and Joanny [1] have shown that the partition function of branching polymers with size  $N$  is given by:

$$\mathcal{Z}_N^{\text{ideal}} = \frac{I_1(2\lambda N)}{\lambda N}, \quad (\text{S1})$$

where  $I_1(x)$  is the modified Bessel function of the first kind and  $\lambda$  is the branching fugacity (related to the branching probability per node). From the known asymptotic behavior for Bessel functions, we can write down the large- $N$  limit of  $\mathcal{Z}_N^{\text{ideal}}$  as

$$\lim_{\lambda N \rightarrow \infty} \mathcal{Z}_N^{\text{ideal}} \simeq \frac{e^{2\lambda N}}{2\sqrt{\pi}(\lambda N)^{3/2}}. \quad (\text{S2})$$

Equations (11) and (12) in the main text then follow from Eq. S2.

Even though Eq. (11) in the main text holds only in the ideal case, it was shown by Rosa and Everaers [2] that it remains remarkably accurate even for interacting branching polymers and hence should hold for RNA molecules as well. In order to take advantage of this and employ Eq. (11) to extract the exponent  $\varepsilon$  from the branch weight distribution functions of RNA molecules, we can introduce the following ansatz [i.e., Eq. (13) in the main text] for the partition function of interacting polymers:

$$\mathcal{Z}_N = \frac{I_\beta(2\lambda N)}{(\lambda N)^\beta}. \quad (\text{S3})$$

In a similar fashion as for Eq. (S2), we can also determine the large- $N$  limit of Eq. (S3):

$$\lim_{\lambda N \rightarrow \infty} \mathcal{Z}_N \simeq \frac{e^{2\lambda N}}{2\sqrt{\pi}(\lambda N)^{\beta+1/2}}. \quad (\text{S4})$$

By using the asymptotic expansion in Eq. (S4), we can show that the distribution of branch weights  $p(N_{\text{br}})$  given by Eq. (11) in the main text follows the power-law-like behaviour in the mid-weight regime,  $1/(2\lambda) \ll N_{\text{br}} < N/2$ :

$$\begin{aligned} p(N_{\text{br}}) &= \frac{\mathcal{Z}_{N_{\text{br}}} \mathcal{Z}_{N-N_{\text{br}}-1}}{\sum_{N_{\text{br}}=0}^{N-1} \mathcal{Z}_{N_{\text{br}}} \mathcal{Z}_{N-N_{\text{br}}-1}} \propto \mathcal{Z}_{N_{\text{br}}} \mathcal{Z}_{N-N_{\text{br}}-1} \\ &\sim e^{2\lambda N_{\text{br}}} e^{2\lambda N_{\text{br}}(N-N_{\text{br}}-1)} (\lambda^2 N_{\text{br}} (N - N_{\text{br}} - 1))^{-(\beta+1/2)} \\ &= \frac{e^{2\lambda(N-1)}}{\lambda^{-2(\beta+1/2)}} (N_{\text{br}} (N - N_{\text{br}} - 1))^{-(\beta+1/2)} \propto (N_{\text{br}} (N - N_{\text{br}} - 1))^{-(\beta+1/2)}. \end{aligned} \quad (\text{S5})$$

By comparing Eq. (S5) to Eq. (12) in the main text, it follows that  $\beta = 3/2 - \varepsilon$ .

# RNA SEQUENCE LENGTH AND THE SIZE OF THE TREE REPRESENTATION OF ITS STRUCTURE

By mapping an RNA secondary structure of length  $N_{\text{nt}}$  to a tree, one obtains a planar tree graph with  $N + 1$  nodes and  $N$  weighted edges (see Sec. II in the main text). The RNA tree can be furthermore expanded in such a way that all edges have unit length, resulting in an expanded tree with  $\tilde{N}$  edges. The RNA sequence length and the number of nodes of both of its tree representations turn out to be almost completely linearly correlated (Fig. S1)—the more so the larger the tree—and we use them interchangeably throughout our work.

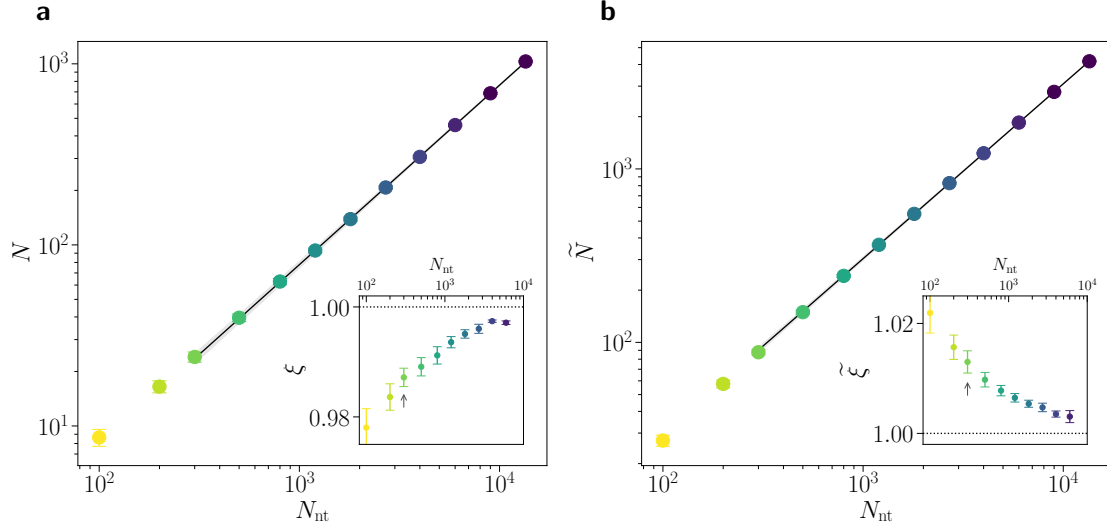


Figure S1. Scaling of the number of (a) tree edges  $N$  and (b) expanded tree edges  $\tilde{N}$  with RNA sequence length  $N_{\text{nt}}$ . The relationship between these quantities,  $N \sim N_{\text{nt}}^\xi$ , is almost completely linear ( $\xi \approx 1$ , insets of panels a and b).



# SCALING EXPONENTS AND NUCLEOTIDE COMPOSITION OF RNA

Most of our work focuses on uniformly random RNA sequences where the frequency of each nucleotide is identical,  $f(b) = 0.25 \forall b$ . To verify that the conclusions we obtain regarding their scaling exponents are also valid in general for an arbitrary nucleotide composition, we selected 16 different nucleotide compositions (Table S1) and evaluated their distance from the uniform random composition by a simple Euclidean metric,

$$\delta^2 = \sum_{b \in \{A, C, G, U\}} \left( f(b) - \frac{1}{4} \right)^2, \quad (\text{S6})$$

which correlates well with improved statistical measures such as Jensen-Shannon divergence. The 16 different nucleotide compositions cover a wide range of possibilities, as they were chosen in such a way that they cover the convex hull of the space of  $\sim 1800$  viral genomes of different positive single-stranded RNA viruses and as such represent the most extreme cases within this dataset.

Both the scaling of the  $\langle \text{ALD} \rangle$  (Fig. S2) and  $\langle N_{\text{br}} \rangle$  (Fig. S3) with sequence length for random RNAs with different nucleotide composition show only a small variation in the scaling exponents  $\rho$  and  $\varepsilon$ . This holds true even when the nucleotide composition differs significantly from that of a uniformly random RNA. The observed differences in scaling are thus solely a consequence of the prefactor of the scaling law, which has previously already been connected to the differences in the percentage of base pairs in the RNA structures at different compositions [3].

Table S1. Nucleotide composition of 16 different sets of RNA used for the comparison of the scaling behaviour of random RNA sequences with different nucleotide compositions.

Composition	$f(A)$	$f(C)$	$f(G)$	$f(U)$	$\delta$
uniform	0.2500	0.2500	0.2500	0.2500	0.0000
1	0.1789	0.2639	0.3048	0.2524	0.0909
2	0.3298	0.2336	0.2427	0.1939	0.0992
3	0.2109	0.2469	0.2042	0.3379	0.1066
4	0.2255	0.1560	0.2864	0.3321	0.1322
5	0.2095	0.1630	0.2735	0.3539	0.1433
6	0.3366	0.2920	0.1409	0.2304	0.1468
7	0.2249	0.3595	0.1408	0.2748	0.1587
8	0.2202	0.1624	0.2382	0.3792	0.1594
9	0.1550	0.3820	0.2352	0.2277	0.1648
10	0.2505	0.1427	0.2259	0.3809	0.1710
11	0.2252	0.4057	0.1311	0.2380	0.1978
12	0.2784	0.1302	0.1904	0.4010	0.2038
13	0.1494	0.3877	0.3082	0.1547	0.2039
14	0.3973	0.1297	0.1452	0.3277	0.2306
15	0.1877	0.4480	0.1486	0.2157	0.2336
16	0.1392	0.4989	0.1635	0.1984	0.2905

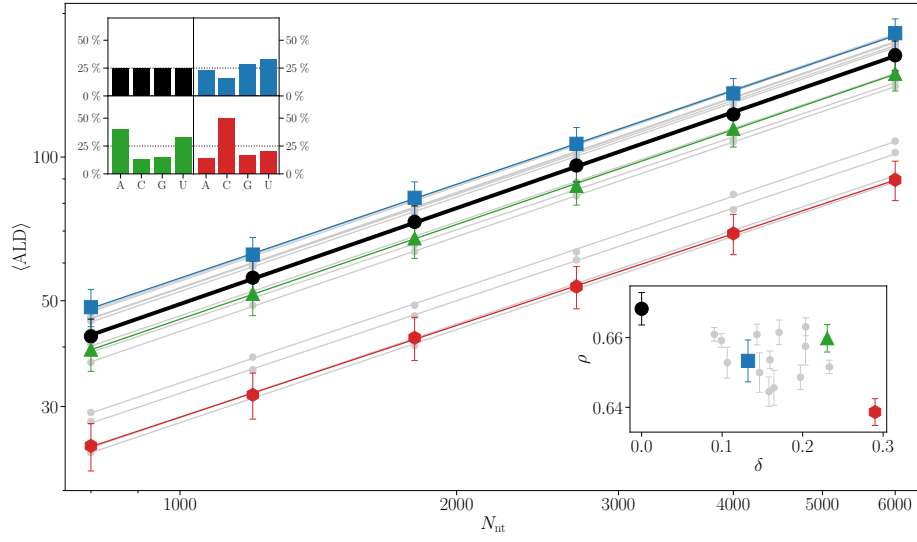


Figure S2. Scaling of  $\langle \text{ALD} \rangle$  with RNA sequence length  $N_{\text{nt}}$  for 16 different nucleotide compositions of random RNA sequences (Table S1). Each datapoint (for a given nucleotide composition and sequence length) contains 200 random sequences. Smaller insets show the nucleotide composition for four different examples. Larger inset shows the scaling exponent  $\rho$  obtained for random RNAs with different nucleotide composition, as measured by Euclidean distance  $\delta$  from uniformly random composition.

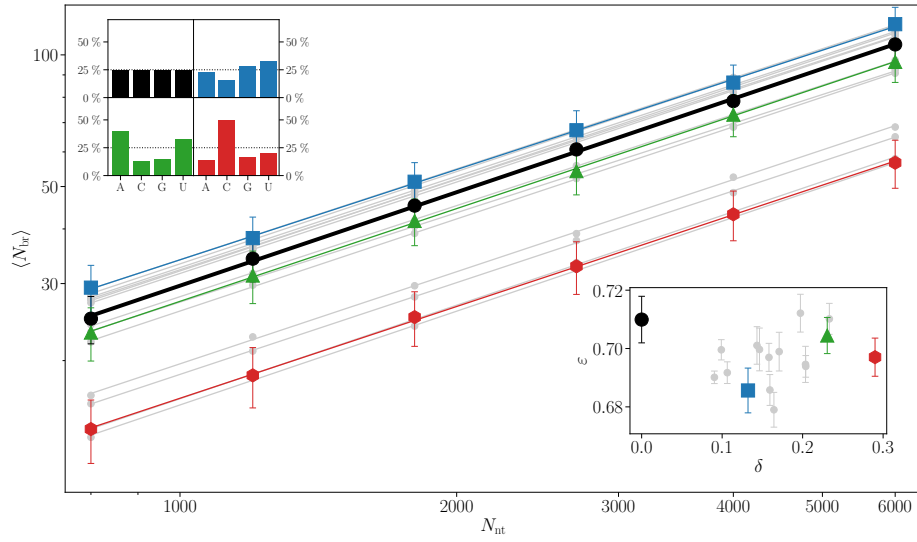


Figure S3. Scaling of  $N_{\text{br}}$  with RNA sequence length  $N_{\text{nt}}$  for 16 different nucleotide compositions of random RNA sequences (Table S1). Each datapoint (for a given nucleotide composition and sequence length) contains 200 random sequences. Smaller insets show the nucleotide composition for four different examples. Larger inset shows the scaling exponent  $\varepsilon$  obtained for random RNAs with different nucleotide composition, as measured by Euclidean distance  $\delta$  from uniformly random composition.

# SCALING EXPONENTS AND MULTILoop ENERGY PARAMETERS OF THE FOLDING SOFTWARE

In the commonly used linear model for the energy of multiloop formation,

$$E_{\text{multiloop}} = E_0 + E_{\text{br}} \times [\text{branches}] + E_{\text{un}} \times [\text{unpaired nucleotides}], \quad (\text{S7})$$

different versions of ViennaRNA software [4] (specifically, those before and after v2.0) use significantly different values of the multiloop energy parameters. In particular, the largest difference lies in the sign of the energy of branch formation  $E_{\text{br}}$  which changes sign in the two parameter sets, either inhibiting large multiloop formations in the older versions (positive value of the parameter) or promoting them in the newer versions (negative value of the parameter; see Table S2). To consistently compare the influence of the multiloop energy parameters on the scaling behaviour of RNA structures, we use the energy parameters from ViennaRNAv2.4 and use either its version multiloop energy parameters or we modify *solely* the multiloop energy parameters with those of older versions, resulting in an energy parameter set we term ViennaRNA-mod. In this way, we can be sure that any differences in scaling we might observe are due to the changes in the multiloop energies. The values for the two sets of multiloop energy parameters used are given in Table S2.

Figure S4 shows the scaling of both  $\langle \text{ALD} \rangle$  and  $\langle N_{\text{br}} \rangle$  with sequence length for uniformly random RNA folded using the two different multiloop energy parameters. Despite the fact that the two sets of parameters tend to result in quite different node degree distributions of the resulting RNA structures (see Ref. [3] for details), we can see that both scaling exponents  $\rho$  and  $\epsilon$  seem to converge to the same values.

Table S2. Multiloop energy parameters used in the current version of ViennaRNA (v2.4) and those from the older versions of the software ( $< 2.0$ ), which we use in a modified set of energy parameters termed ViennaRNA-mod.

Multiloop parameters	$E_0$	$E_{\text{un}}$	$E_{\text{br}}$
ViennaRNAv2.4	9.3	0.0	-0.9
ViennaRNA-mod	3.4	0.0	0.4

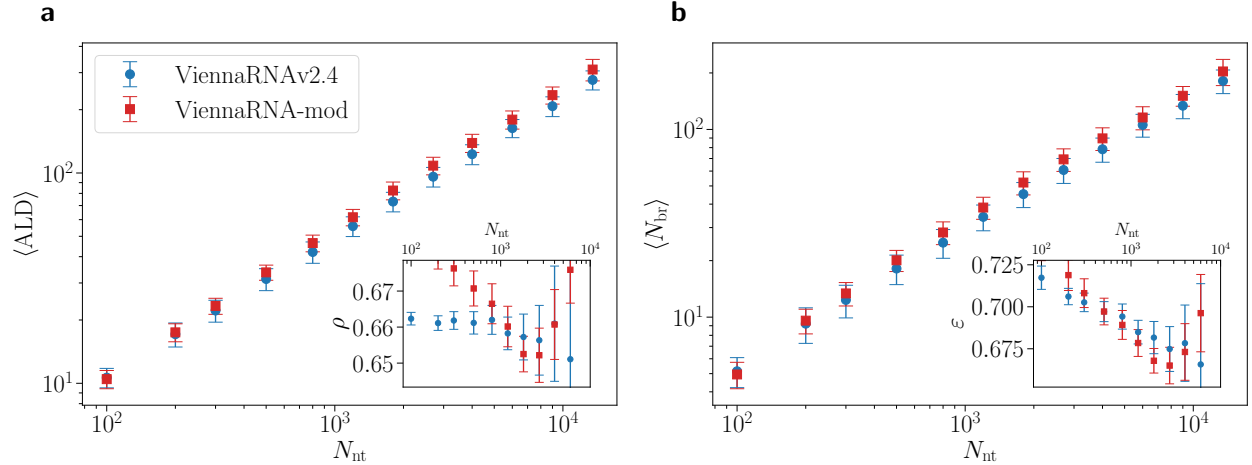


Figure S4. Scaling of (a)  $\langle \text{ALD} \rangle$  and (b)  $N_{\text{br}}$  with RNA sequence length  $N_{\text{nt}}$  using two different multiloop energy parameters, based either on the current version of ViennaRNA (ViennaRNAv2.4) or on its older versions ( $< 2.0$ , ViennaRNA-mod). The insets show how the scaling exponents change as we progressively narrow the fit region to longer sequence lengths.

# SCALING EXPONENTS AND PRÜFER-SHUFFLED RNA TREES

Every planar tree graph with  $N$  edges can be uniquely represented by its Prüfer sequence of length  $N - 1$  (see Ref. [5] for details). Permutation of this sequence (Prüfer shuffle) results in a different tree with an identical *distribution of node degrees*. Thus, by permuting the Prüfer sequences of RNA structures mapped to trees, we can obtain different trees whose node degree distribution exactly matches that of the original RNA and can be considered RNA-like. Figures S5 and S6 show the scaling of  $\langle \text{MLD} \rangle$  and  $\langle N_{\text{br}} \rangle$  with the number of edges  $N$  for random RNA and its Prüfer-shuffled versions (each sequence averaged over 500 Prüfer shuffles). For comparison, we also show the scaling of randomly generated Prüfer sequences, corresponding to random labelled trees with  $N$  edges.

We also note that while RNA trees can be considered unlabelled, Prüfer sequence representation in general operates in the space of labelled trees and is therefore sensitive to node labelling. To make sure that we can nonetheless compare the scaling of RNA trees and their Prüfer-shuffled versions, we also generated random unlabelled (non-isomorphic) trees with  $N$  edges as implemented in Giac/Xcas software [6]. The scaling of  $\langle \text{MLD} \rangle$  for both random labelled and unlabelled trees is essentially the same (Fig. S5), which justifies *post hoc* our initial comparison of RNA trees and their Prüfer-shuffled versions.

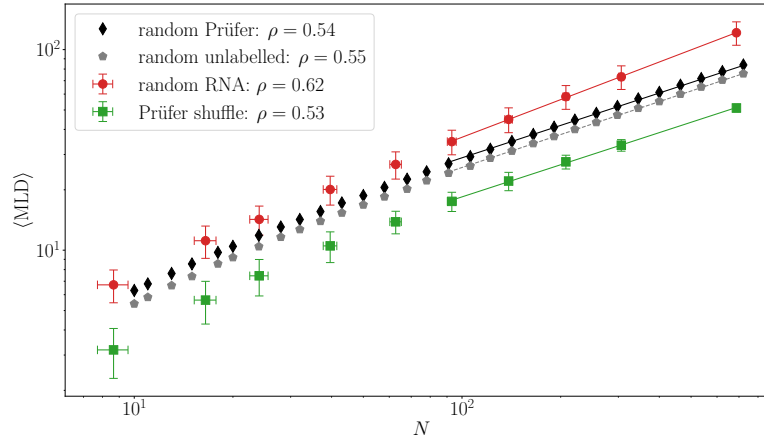


Figure S5.  $\langle \text{MLD} \rangle$  as a function of  $N$  for uniformly random RNA sequences of different lengths (red circles) and their Prüfer-shuffled versions (green squares). The error bars show standard deviations of averaging over 500 permutations of the corresponding Prüfer sequence of the original RNA graph. Black diamonds show the scaling of random Prüfer sequences, corresponding to a sampling of random labelled trees, and gray pentagons show the scaling behaviour of random unlabelled (non-isomorphic) trees. Lines show power law fits of the form  $\langle \text{MLD} \rangle = \alpha N^\rho$ .

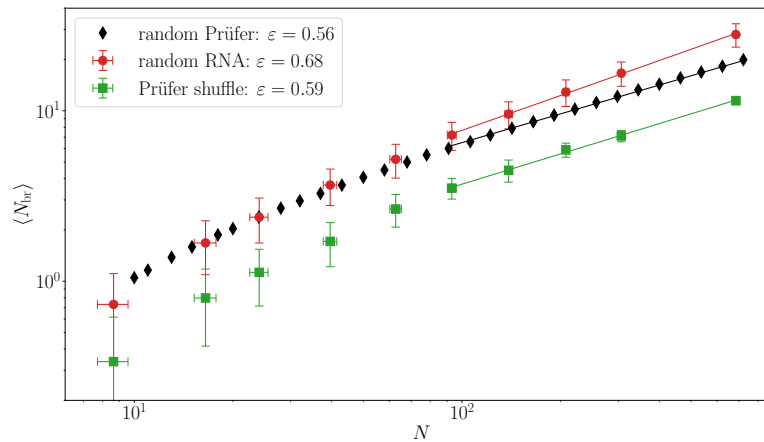


Figure S6.  $\langle N_{\text{br}} \rangle$  as a function of  $N$  for uniformly random RNA sequences of different lengths (red circles) and their Prüfer-shuffled versions (green squares). Each square represents an average over 500 permutations of the corresponding Prüfer sequence of the original RNA graph. Black diamonds show the scaling of random Prüfer sequences, corresponding to a sampling of random labelled trees. Lines show power law fits of the form  $\langle N_{\text{br}} \rangle = \alpha N^\epsilon$ .



# SCALING EXPONENTS $t$ AND $\theta$ OF THE RDC DISTRIBUTION OF PATH LENGTHS

To obtain the scaling exponent  $\rho$  from the distribution of path lengths  $p(\ell)$  of individual RNA structures (or, more precisely, sets of RNA structures *at fixed length*), we use the two-parametric RdC distribution [Eq. (8) in the main text] and fit it to the path length distribution. Typically, we do this for a set of random sequences and their structures at a fixed length, and obtain the two fit parameters  $t$  and  $\theta$  (panels a and b of Fig. S7). One can show that both fit parameters can be connected to the scaling exponent  $\rho$ . In this way, we can use the two parameters to obtain two independent estimates of the scaling exponent,  $\rho_t$  and  $\rho_\theta$ , respectively. Under certain assumptions (see the discussion in Sec. IV C in the main text), these two parameters are not independent but are related through

$$\theta = \frac{1}{t-1}. \quad (\text{S8})$$

Figure S7c shows how the RdC fit exponents for uniformly random RNAs of different lengths compare to Eq. (S8). For comparison, we also show the values of four different types of randomly branched polymers, taken from Ref. [2].

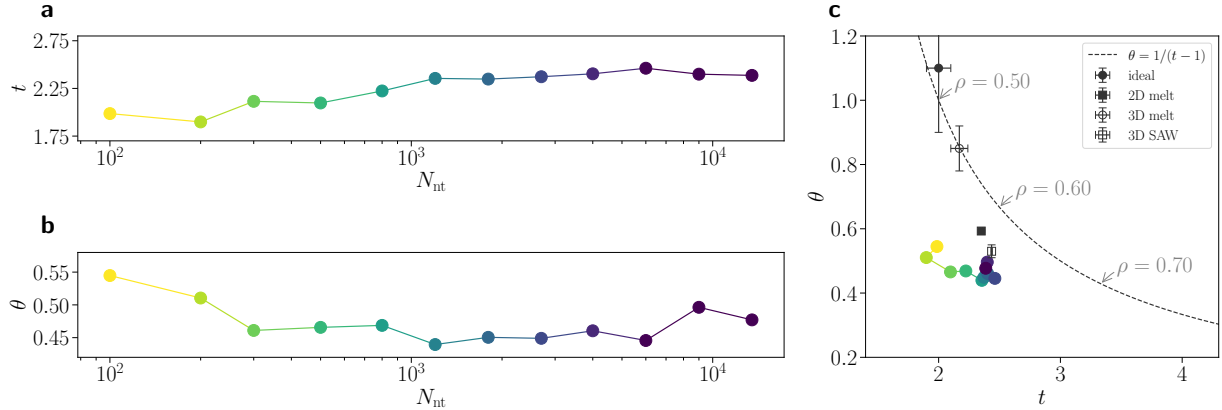


Figure S7. Parameters (a)  $t$  and (b)  $\theta$  of the RdC distribution obtained from path length distributions of uniformly random RNA sequences of different lengths. Each point represents a fit to the distribution of path lengths of 200 random sequences with 500 structures for each. Panel (c) shows the changes in the fitted exponents in the  $(t, \theta)$  plane, where one can observe how the relationship between them changes with increasing sequence length (colour-coded in panels a and b). Also shown are the values for four different types of branched polymers. Dashed line shows the theoretical relationship between the two fit exponents, Eq. (S8).

- 
- [1] M. Daoud and J. F. Joanny, *J. Physique* **42**, 1359 (1981).
  - [2] A. Rosa and R. Everaers, *Phys. Rev. E* **95**, 012117 (2017).
  - [3] D. Vaupotič, A. Rosa, R. Podgornik, L. Tubiana, and A. Božič, “Viral RNA as a branched polymer,” (2022), [arXiv:2212.00829 \[physics.bio-ph\]](https://arxiv.org/abs/2212.00829).
  - [4] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, *Algorithms Mol. Biol.* **6**, 1 (2011).
  - [5] S. W. Singaram, A. Gopal, and A. Ben-Shaul, *J. Phys. Chem. B* **120**, 6231 (2016).
  - [6] B. Parisse and R. De Graeve, “*Giac/xcas*, v1.9.0,” (2022), <https://www-fourier.univ-grenoble-alpes.fr/~parisse/giac.html>.