RNA Secondary Structures in a Polymer-Zeta Model
How Foldings Should be Shaped for Sparsification to Establish a Linear Speedup

Emma Yu Jin[1], Markus E. Nebel[2,3]
[1,2]University of Kaiserslautern, Germany
[3]University of Southern Denmark, Denmark
Email: {jin,nebel}@cs.uni-kl.de

Abstract

Various tools to predict the secondary structure for a given RNA sequence are based on dynamic programming used to compute a conformation of minimum free energy. For structures without pseudoknots, a worst-case runtime proportional to n^3, n the length of the sequence, results for a table of dimension n^2 has to be filled in while a single entry gives rise to a linear computational effort. However, recently it has been observed that reformulating the corresponding dynamic programming recursion together with bookkeeping of potential folding alternatives (a technique called sparsification) may reduce the runtime to n^2 on average, assuming that nucleotides of distance d form a hydrogen bond (i.e. are paired) with probability $\frac{b}{d^c}$ for some constants $b > 0$, $c > 1$. The latter is called the polymer-zeta model and plays a crucial role for the speedup of before mentioned algorithm. In this paper we first show, that the polymer-zeta property does not apply to the analysis of sparsification since there conditional probabilities need to used. Afterwards, we investigate the combinatorics of RNA secondary structures assuming that the needed conditional probabilities behave like in a polymer-zeta probability model. We show that even if many of the structural parameters behave almost realistic on average, the expected shape of a folding in that model must be assumed to highly differ from those observed in nature. More precisely, we prove our polymer-zeta model to be appropriate for mRNA molecules but to fail in connection with almost every other family of RNA. To this end we oppose our findings to a stochastic model of RNA proven to reflect the native shape of RNA and statistics derived from databases.

1. Introduction

The applications of RNA secondary structure prediction in computational biology are manifold and we refer the reader to [1] for an overview. Here we only discuss the rather special, nevertheless highly interesting application of identifying accessible motifs in mRNA on a genome wide level from [14] which motivated our research of this paper.

It is well known that, in order to respond to different stimuli, synthesis of proteins needs to be highly regulated [7]. One mechanism being in charge relies on cis-regulatory motifs within mRNAs to which trans-regulatory proteins and microRNAs bind. Since the chemical recognition is based on an interaction between amino acids residing in the protein and the corresponding nucleotides in the cis-regulatory motif residing in the mRNA (see [14] and the reference given there) it is of importance, that the nucleotides constituting the motif are accessible i.e. are not bonded for the mRNA. However, in order to decide this property it is no longer sufficient to work on sequence level, the 2D conformation of the mRNA, i.e. its secondary structure, needs to be taken into account. Thus, for the computational search for cis-regulatory motifs in RNA, structure prediction algorithms are needed. However, if one aims for identifying mRNA motifs on a genome wide level, the classical $\mathcal{O}(n^3)$ time algorithms (see e.g. [17, 5, 16]) are not appropriate. To this end, Wexler et al. came up with the following idea: Like for the common dynamic programming (DP) algorithms to minimize free energy, two recursions W and V are used. Processing input sequence $s_1s_2\cdots s_n$, $V(i,j)$ represents the minimal energy possible for a folding of subsequence $s_i\cdots s_j$ subject to the i-th and j-th nucleotide being paired to each other. $W(i,j)$ gives the corresponding minimum without that restriction. Then by distinguishing the cases of an optimal folding of sequence $s_i\cdots s_j$ to either

[1] The work of this author has been supported by the Alexander von Humboldt Foundation by a postdoctoral research fellowship.

[2] Author to whom correspondence should be addressed.

- be an optimal folding with s_i being paired to s_j,
- to result from unpaired s_i attached to an optimal folding for $s_{i+1} \cdots s_j$, or from unpaired s_j attached to an optimal folding for $s_i \cdots s_{j-1}$, or
- to result from combining two smaller optimal foldings with a *bifurcation* at position k,

we arrive at formula

$$W(i,j) = \min\{V(i,j),\, W(i+1,j),\, W(i,j-1),\, \min_{i \le k < j}\{W(i,k) + W(k+1,j)\}\}.$$

Recursion V is built along the same ideas, introducing different contributions e_x, $x \in \{h,s,b\}$, to the free energy depending on the kind of loop x the pairing of s_i and s_j closes, and a constant multi-branch penalty a:

$$\begin{aligned} V(i,j) \quad = \quad \min \Bigg\{ & \underbrace{e_h(i,j)}_{\text{hairpin}},\ \underbrace{e_s(i,j) + V(i+1,j-1)}_{\text{stem}},\ \underbrace{\min_{i<i'<j'<j}\{e_b(i,j,i',j') + V(i',j')\}}_{\text{bulge or interior loop}}, \\ & \underbrace{\min_{i \le k < j}\{W(i+1,k) + W(k+1,j-1)\} + a}_{\text{multibranch loop}} \Bigg\}. \end{aligned}$$

Since a quadratic number of different combinations of i and j have to be considered, each determining a minimum over a linear number of elements in expectation, using this representation implies a cubic running time to compute W even in the average-case. Now it is possible to rewrite recursion W (without affecting the computed optimum) such that it obeys the triangle inequality

$$\forall i < j' < j \quad W(i,j) \le W(i,j') + W(j'+1,j).$$

This property of the new recursion finally can be used to prove that for the computation of the optimal folding for subsequence $s_i \cdots s_j$ a pairing of s_i and s_k only needs to be considered if pairing of s_i and s_k already implied a minimum while considering $s_i \cdots s_{j'}$, $j' < j$ (see [14] for details). Furthermore, when computing W in the right ordering, this information is available whenever needed and the candidates s_k to be considered can be maintained in a list associated to index i. This idea gives rise to the following folding algorithm

Algorithm CandidateFold:

```
for each row $i := n$ to 1 do
    candidatelist:= $\varnothing$;
    for each column $j := i$ to $n$ do
        $W(i,j) := \min_{k \in \text{candidatelist}}\{V(i,k) + W(k+1,j)\}$;
        if $V(i,j) < W(i,j)$ then
            $W(i,j) := V(i,j)$
            Append $j$ to candidatelist;
```

It is obvious that the expected running time of CandidateFold depends on the expected length of the lists maintained during its execution. Wexler *et al.* have claimed in [14] that under the assumption of the so-called polymer-zeta property with parameter $c > 1$ this expected length is constant, implying a quadratic runtime in the average-case. As a result, one can assume the DP matrix to be sparse in case of the polymer-zeta property thus speaking of *sparsification* in connection with the technique used for CandidateFold. Here, *polymer-zeta property* means that the probability for the i-th and j-th nucleotides at distance (span) $d = j - i + 1$ to form a pair is given by $p_d = \frac{b}{d^c}$ (for some constants $b > 0, c > 0$). The theoretical choices for the parameters are $b = 1$ and $c = 1.5$ (see also the subsequent analysis). However, as we will point out in the sequel, their way of reasoning was faulty. As a consequence, we will show that sparsification cannot be assume to save a linear factor for above DP algorithm. Afterwards, we invert the question, asking how RNA structures would need to look like for sparsification to really be effective. To this end, we will assume a probability model for secondary structures that is in one-to-one connection to the way Wexler *et al.* argue in their paper. We find a rather unnatural appearance for some of the structural features.

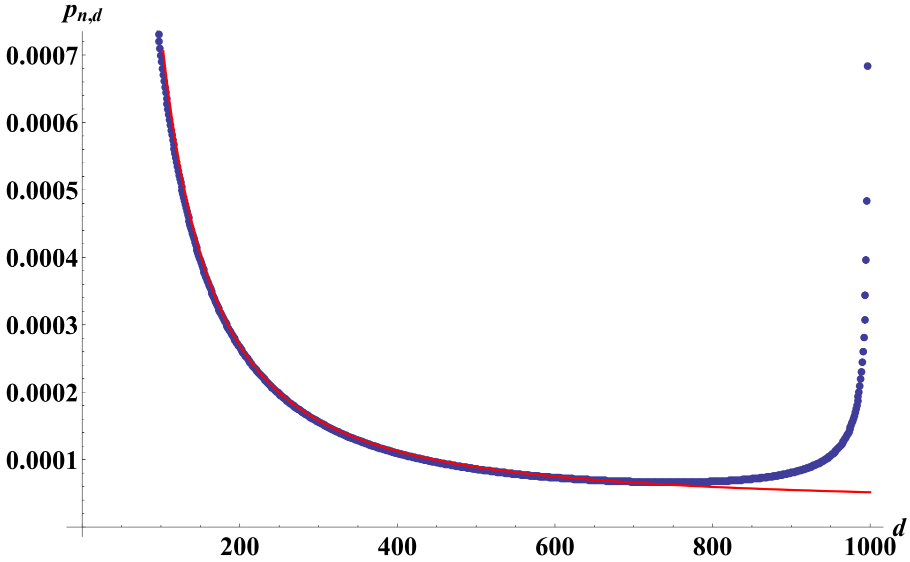

Figure 1. The probability $p_{n,d}$ (analytically determined formula as blue dots, fitted curve $\approx \frac{1}{d^{1.5}}$ as read line) for of span d to show up in a structure of size $n = 1000$ assuming a uniform distribution for secondary strutures.

But first let us show, why the arguments in [14] are wrong. We start with considering the classic combinatorial model for RNA secondary structure (see [9] and the references given there) assuming a minimal hairpin loop length of 1. Building on the generating functions from [2] we computed the (precise asymptotic) probability $p_{n,d}$ for having a distance d base pair in a secondary structure of size n (the related computations will be reported elsewhere):

$$p_{n,d} = \frac{\left(5 + 2\sqrt{5}\right)\sqrt{n}}{\sqrt{2\left(15 + 7\sqrt{5}\right)\pi\sqrt{(d-2)^3}\sqrt{n-d+2}}}.$$

Investigating this probability further shows, that even the combinatorial model behaves in a polymer-zeta style unless d is not to close to n. To make this point clear, take a look at Figure 1 where we depict the plot of $p_{n,d}$ for $n = 1000$ and d between 4 and 1000 (blue dots) together with a curve fitting (red line) we computed using Mathematica according to the pattern $\frac{b}{d^c} + k$ (finding $b \approx 1$, $c \approx 1.5$ and $k \approx 0$). Thus, at a first sight the probability for span d base pairings seems to behave as claimed by [14] with the only difference, that the native structures discussed in that paper gave rise to slightly different parameters – in [14] the authors have experimentally justified the polymer-zeta model for their mRNA data, finding constants $b = 2.11$ and $c = 1.47$. However, a careful look at their arguments yields a contradiction (see Observation 1 in [14]): "A new candidate j is added to the candidate list [...] iff the optimal predicted folding of substring $s_i \dots s_j$ forms a single structure from index i to index j", i.e. if and only if within the optimal folding for subsequence $s_i \dots s_j$ bases s_i and s_j are paired. But this kind of pairing is not equivalent to the assumptions of the polymer-zeta model since we must condition the base pair to connect the first and the last element of the considered (sub)structure instead of connecting any two bases (and thus disallowing many pairings which would cross the considered outermost one). We computed the corresponding asymptotic ($n \to \infty$) probability $\hat{p}_d$ again for the combinatorial model, finding

$$\hat{p}_d = \frac{\left(\sqrt{5}-3\right)^2\sqrt{d^3}}{4\sqrt{(d-2)^3}} = \left(\frac{7}{2} - \frac{3\sqrt{5}}{2}\right) + \frac{\frac{21}{2} - \frac{9\sqrt{5}}{2}}{d} + \frac{\frac{105}{4} - \frac{45\sqrt{5}}{4}}{d^2} + \mathcal{O}\left(\left(\frac{1}{d}\right)^{5/2}\right), \; d \to \infty.$$

This is close to a d^{-1} behavior and indeed, calculating the related expected length of a candidate list yields the linear behavior as shown in Figure 2. Thus, in accordance with the findings of [8]

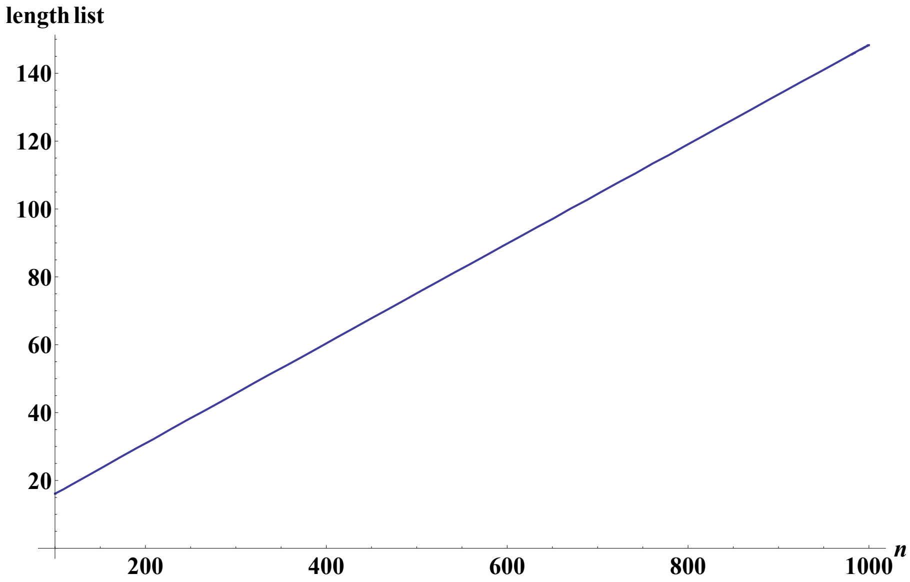

Figure 2. The expected length of a candidate list as a function of the input size n computed for the combinatorial model of RNA.

– there the authors also considered a combinatorial model to quantify the number of secondary structures to which sparsification applies – we find that a quadratic runtime is not possible on average but a cubic time with a huge reduction of the leading constants involved. Note, that due to the identical singular structure of the respective generating functions, the qualitative same behavior results for the so-called Bernoulli-model [10].

In conclusion, at least in the combinatorial regime we cannot assume that sparsification really gives rise to a linear speedup for structure prediction. On the other hand, it is hard to reason about native structures in a similar rigorous way for we do not have appropriate models that reflect the real world behavior of all the different classes of RNA sufficiently precise. Therefore, in this paper, we proceed in a different way. We study the expected shape of RNA secondary structures of size n assuming that the (conditional) probability of an outermost base pair of span d is given by $\frac{b}{d^c}$. Under this assumption we determine the distribution and the expected shape of structural motifs like hairpins, bulges, interior-loops, multiloops and the exterior loop of secondary structures under various (c, b)-polymer-zeta-models (various choices for b and c). Then, a comparison of the appearance of any RNA family to our average-case statistics can easily provide a first hint at whether or not the data shows a similar appearance as implied by a probability model that would allow for sparsification to save a linear factor indeed. Opposing our findings to a stochastic model of RNA assumed to reflect the native shape of RNA, we find that the *polymer-zeta world* only slightly differs from what is observed in nature for some structural motifs while we observe drastic differences for others.

2. The Expected Shape of RNA in a polymer-zeta Model

We assume the reader familiar with the definition of RNA secondary structures (considered as a combinatorial object) as well as their structural motifs like hairpins, bulges, etc. We refer to [9] for the terminology used here.

Let $\mathcal{R}_n$ be the class of all the RNA primary structures, i.e. sequences $s_1 \cdots s_n$ with $s_i = A, U, G, C$. Let $\mathcal{S}_n$ be the class of all RNA secondary structures of size n. Here we assume that $\mathsf{S}_n \in \mathcal{S}_n$ is represented as the set of all its base pairs (opposed to other equivalent representations like dot-bracket or planar graph). Our study of $\mathcal{S}_n$ will be conducted analogously to that from [12, 10] for the Bernoulli model. For the latter one considers the expected number of different RNA secondary structures (and related parameters) supposing that only structures compatible (with respect to complementarity) to a random sequence (this provides the model of randomness) are counted. Accordingly, since the entirely unpaired structure is compatible to any sequence, it has probability 1. For paired positions, the so-called stickiness p – defined as the probability of two (according to a Bernoulli experiment) random s_i (nucleotides) being complementary – comes into play; any two paired nucleotides are weighted p in the Bernoulli model. As a consequence, every secondary structure of size n and i pairs of paired nucleotides is considered with probability p^i. More precisely, if we aim at computing the number of different secondary structures $r(n)$ of size n on a sequence $s \in \mathcal{R}_n$ of length n we find [12] the following recursion (either attach an unpaired nucleotide at position $n+1$ to any of the $r(n)$ smaller structures or pair it with one of those at position 1 to $n-1$ which decomposes the structure accordingly):

$$r(n+1) = r(n) + \sum_{0 \le k \le n-2} r(k) r(n-k-1) \eta(k+1, n+1). \tag{2.1}$$

Here $r(0) = r(1) = r(2) = 1$ must be assumed and $\eta(i,j)$ is the indicator which is 1 iff s_i and s_j are complementary. Now taking expectations of (2.1) we arrive at $e(n+1) = e(n) + \sum_{0 \le k \le n-2} p \cdot e(k) e(n-k-1)$ where stickiness p obviously corresponds to the expectation of η. Thus, in order to compute the expected number of different secondary structures on a random sequence of size n, nucleotides which are paired have to be weighted p.

Here, we introduce a (c,b)-polymer-zeta-model for RNA structures in an analogous way. For any $\mathsf{S}_n \in \mathcal{S}_n$ we list all its base pairs indexed by their distances (length of enclosed subsequence $+1$), i.e. $\mathsf{S}_n = \{r_{d_1}, r_{d_2}, \cdots, r_{d_\ell}\}$ means, that the i-th base pair in S_n has distance d_i, $1 \le i \le \ell$. We say $\mathsf{S}_n = \varnothing$ if S_n has no base pairs. By weighting each base pair of distance d with probability $p_d = \frac{b}{d^c}$, we arrive at probability $\mathbb{P}(\mathsf{S}_n) = \prod_{i=1}^{\ell} \frac{b}{{d_i}^c}$, and – like for the Bernoulli model – the probability of having a completely unpaired structure is $\mathbb{P}(\varnothing) = 1$. We name this model as (c,b)-polymer-zeta-model. This model is consistent with the considerations on the expected length of candidate lists of the CandidateFold algorithm, since according to the underlying decomposition of secondary structures, no potentially crossing interactions contribute, i.e., we consider the case that the two connected nucleotides at distance d form a single structure with probability p_d. Accordingly, for $c > 1$, the expected length of a candidate list would indeed be constant in our model. To convince the reader, that our approach really behaves as claimed, we have adapted (2.1) to our (c,b)-polymer-zeta-model and (numerically) computed the probability p'_d of a span d base pair conditioned on the fact that it connects first and last elements within the folding of an entire (sub)sequence. The plot of Figure 3 shows the resulting behavior (blue dots) together with a curve (red line) fitted to the computed probabilities according to the pattern $\frac{b}{d^c} + k$ using Mathematica (finding $b \approx 0.7$, $c \approx 2$ and $k \approx 0$).

In what follows, we shall classify RNA secondary structures according to the number and the length of hairpin-loops, bulge-loops and interior-loops as well as the degree of multiloops, the number of unpaired nucleotides in the exterior loop and investigate the average behavior of these parameters in the context of the (c,b)-polymer-zeta-model. In contrast to the assumption of $c = 1.47$ motivated above, our methodology will only allow to deal with integer choices for c. However, assuming a continuous transition of parameters when changing c from 1 to 2 our findings pin the quantitative behavior of the various structural features down to a small interval. Accordingly, above motivated values like $c = 1.5$ (or $c = 1.47$ as reported by [14]) can be assumed to imply a behavior lying in those intervals. More importantly, it becomes possible by our results to describe the expected shape of RNA structures for which sparsification would provide a linear speedup – here only $c > 1$ is needed. An overview of some of the corresponding results to be derived can already be found in Table 1, where the findings of this paper are opposed to those from [11] which have been derived from a sophisticated stochastic context-free grammar model (proven to nicely reflect the native behavior of RNA).

For the reader's convenience, we have visualized in Figure 4 the results of Table 1 such that the respective values can easily be compared. A first glance at that figure already reveals a notable disapproval of our polymer-zeta model to the native behavior of RNA foldings for almost all parameters considered.

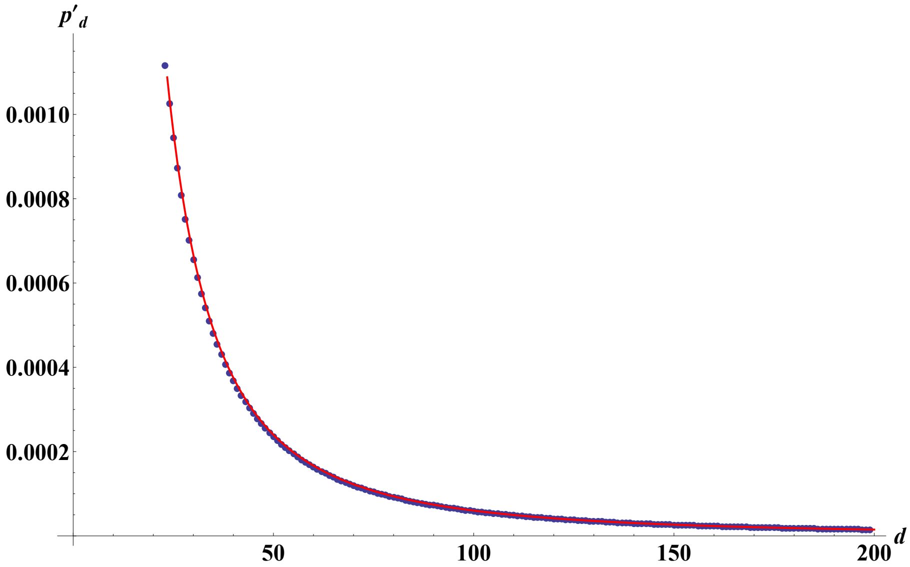

FIGURE 3. Probability p'_d of a span d base pair connecting the outermost nucleotides of a folded (sub)sequence for our $(2,1)$-polymer-zeta-model (blue dots) together with the fitted curve $\approx \frac{0.7}{d^2}$ (read line).

The expected behavior of various structural parameters as derived from a stochastic model:

parameter	expectation				
		$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
Number of hairpins	$0.0226n$	$0.1313n$	$0.1462n$	$0.1202n$	$0.1458n$
Length of a hairpin-loop	7.3766	1.7262	1.5291	1.7367	1.5467
Number of bulges	$0.0095n$	$0.0210n$	$0.0261n$	$0.0076n$	$0.0113n$
Length of a bulge	1.5949	2.0476	1.7625	2.4079	2.0354
Number of interior loops	$0.0164n$	$0.0110n$	$0.0099n$	$0.0055n$	$0.0059n$
Total Length of both loops within an interior loop	7.7870	4.2364	3.6162	5.3455	4.4068
Number of multiloops	$0.0106n$	$0.0330n$	$0.0390n$	$0.0112n$	$0.0220n$
Degree of a multiloop	4.1311	5.9848	5.5615	12.5536	8.3636

TABLE 1. The asymptotic expected behavior of various structural motifs in random RNA secondary structures of size n, $n \to \infty$, according to a stochastic model (second column) and our new polymer-zeta model for different choices for c and b as indicated by the column heading (c,b) (third to sixth column). All constants shown are rounded to the fourth decimal digit.

3. AVERAGE NUMBER OF HAIRPINS

Let $X_n^{c,b}$ denote the random variable counting the number of hairpins in a secondary structure of size n and $\mathbb{E}(X_n^{c,b})$ (resp. $\sigma(X_n^{c,b})$) be the expectation (resp. standard deviation) of $X_n^{c,b}$. We find:

Theorem 1. *Under the assumption of the (c,b)-polymer-zeta-model, $c \in \{1, 2\}$, the number of hairpins in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\mathbb{E}(X_n^{1,b}) = x_{1,b}^{(1)} n + x_{1,b}^{(2)} + \mathcal{O}(n^{-1}), \quad \mathbb{E}(X_n^{2,b}) = x_{2,b}^{(2)} n + x_{2,b}^{(2)} \frac{n}{\log n} + \mathcal{O}(\frac{n}{\log^2 n})$$

and standard deviation

$$\sigma(X_n^{1,b}) = x'_{1,b}\sqrt{n}(1 + \mathcal{O}(n^{-1})), \quad \sigma(X_n^{2,b}) = x'_{2,b}\sqrt{n}(1 + \mathcal{O}((\log n)^{-1}))$$

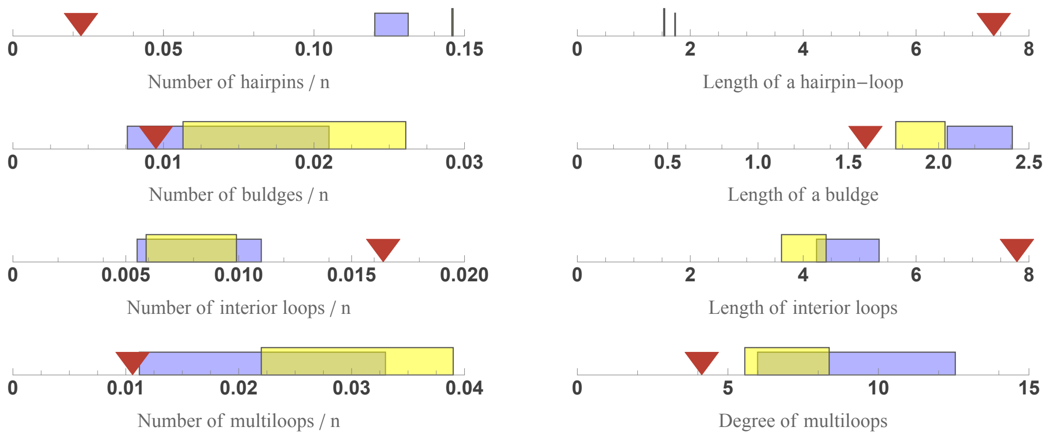

FIGURE 4. A visualization of the results shown in Table 1. For all diagrams, the red triangle represents the values according to a stochastic model (second column of table), the blue and marginally lower (resp. yellow and marginally higher) rectangle shows the range spanned by the choices $c = 1$ and $c = 2$ for $b = 1$ (resp. $b = 2$).

where $x_{c,b}$ and $x'_{c,b}$ are positive constants. For $b \in \{1, 2\}$ we have

$$(x^{(1)}_{1,1}, x^{(2)}_{1,1}, x'_{1,1}) \approx (0.1313, 0.1313, 0.1871) \qquad (x^{(1)}_{1,2}, x^{(2)}_{1,2}, x'_{1,2}) \approx (0.1462, 0.1460, 0.1857)$$
$$(x^{(1)}_{2,1}, x^{(2)}_{2,1}, x'_{2,1}) \approx (0.1202, 0.0376, 0.2047) \qquad (x^{(1)}_{2,2}, x^{(2)}_{2,2}, x'_{2,2}) \approx (0.1458, 0.0259, 0.2022)$$

Proof. Let $\mathcal{S}_{n,k}$ be the class of secondary structure of size n with k hairpins, and – as before – $\mathcal{S}_n$ the class of all secondary structures of size n. Accordingly, let $\mathbb{E}^{c,b}_{\#}(\mathcal{S}_{n,k})$ (resp. $\mathbb{E}^{c,b}_{\#}(\mathcal{S}_n)$) denote the corresponding expected (averaged) number[3] of structures in that class assuming the (c, b)-polymer-zeta-model. Then

$$\mathbb{E}(X^{c,b}_n) \quad = \quad \sum_{k\geq 1} k \cdot \frac{\mathbb{E}^{c,b}_{\#}(\mathcal{S}_{n,k})}{\mathbb{E}^{c,b}_{\#}(\mathcal{S}_n)}.$$

We turn to the bivariate generating function

$$S_c(z, w) = \sum_{n\geq 3}\sum_{k\geq 1} \mathbb{E}^{c,b}_{\#}(\mathcal{S}_{n,k}) w^k z^n + \sum_{n\geq 0} z^n$$

where the summation over size n assumes $n \geq 3$ since we consider a minimal length of hairpin-loops of 1. We derive a representation for $S_c(z, w)$ by regarding the class $\mathcal{T}_{n+2,k}$ of so-called irreducible structures (IS) which are given by those structures from $\mathcal{S}_{n+2,k}$ with the first and the last base paired by a hydrogen bond. We have for $k \geq 2$,

$$\mathbb{E}^{c,b}_{\#}(\mathcal{T}_{n+2,k}) = \frac{b}{(n+1)^c} \cdot \mathbb{E}^{c,b}_{\#}(\mathcal{S}_{n,k}), \tag{3.1}$$

and in the case $k = 1$,

$$\mathbb{E}^{c,b}_{\#}(\mathcal{T}_{n+2,1}) = \frac{b}{(n+1)^c}(1 + \mathbb{E}^{c,b}_{\#}(\mathcal{S}_{n,1}))$$

[3] Note that in our polymer-zeta model like for the Bernoulli model no fix numbers of structures but only expected numbers exist.

holds. Let $T_c(z,w)$ be the double generating function of $\mathbb{E}_{\#}^{c,b}(\mathcal{T}_{n+2,k})$ ($n \geq 3$, $k \geq 1$). Based on eq. (3.1), we find

$$T_c(z,w) \quad = \quad b\sum_{n\geq 3}\sum_{k\geq 1}\frac{1}{(n+1)^c}\cdot\mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k})w^k z^{n+2} + \sum_{n\geq 1}\frac{b}{(n+1)^c}wz^{n+2}. \tag{3.2}$$

On the other hand, each $\mathsf{S}_{n,k}$-structure can be considered a sequence of $\mathsf{T}_{i,j}$-structures with leading, intermediate and trailing run of unpaired bases. In terms of generating functions, we thus have

$$S_c(z,w) = \frac{1}{1-(T_c(z,w)+z)}\,. \tag{3.3}$$

We start our analysis from the case $c=1$. Dividing by z and taking the partial derivative in z (denoted by index z) on both sides of eq. (3.2), we obtain

$$(\frac{T_1(z,w)}{z})_z = b\sum_{n\geq 3}\sum_{k\geq 1}\mathbb{E}_{\#}^{1,b}(\mathcal{S}_{n,k})w^k z^n + \sum_{n\geq 1}bwz^n = bS_1(z,w) + \frac{b(wz-1)}{1-z} \tag{3.4}$$

and thus get rid of denominator $(n+1)$. In combination of eq. (3.3), we find the functional identity for $S_{1,b} = S_{1,b}(z,w)$, given by

$$\frac{\partial S_{1,b}}{\partial z} = -\frac{1}{z}S_{1,b} + \left[\frac{1}{z} + \frac{bz(wz-1)}{1-z}\right]S_{1,b}^2 + zbS_{1,b}^3 \tag{3.5}$$

with initial condition $S_{1,b}(0,w) = 1$. Our ultimate goal via eq. (3.5) is to derive the distribution of the number of hairpins for secondary structures of size n according to our polymer-zeta model. Here we shall first prove for any $w \in (1-\epsilon, 1+\epsilon)$ where $\epsilon > 0$ is sufficiently small, that eq. (3.5) has an algebraic solution with a unique dominant singularity at $z = \rho(w)$. The next step is to show that this is true uniformly for any $|w-1| < \epsilon$. As the third step we will apply a theorem of perturbation on singularity analysis to prove the Gaussian distribution of $X_n^{1,b}$ with both mean and variance linear in n asymptotically. The main tool used here is a transfer theorem from [6], which is phrased as follows:

Theorem 2. (Transfer theorem)[6] *Let $\mathcal{U} = \{(1-z)^{-\alpha}\lambda(z)^{\beta} : \alpha, \beta \in \mathbb{C}\}$ for $\lambda(z) = \frac{1}{z}\log\frac{1}{1-z}$. Assume that $f(z)$ is analytic at 0 with a singularity at z_0, such that $f(z)$ can be continued to some $\Delta_{z_0}(M,\phi)$ domain, and there exist two functions σ, τ from $\mathcal{U}$, such that*

$$f(z) = \sigma\left(\frac{z}{z_0}\right) + \mathcal{O}\left(\tau\left(\frac{z}{z_0}\right)\right).$$

Then we have $[z^n]f(z) = z_0^{-n}\sigma_n + \mathcal{O}(z_0^{-n}\tau_n)$ where $\rho_n = [z^n]\rho(z)$, $\rho(z) = (1-z)^{-a}\lambda(z)^b$ for $a \notin \mathbb{Z}_{\leq 0}$, is given by $\frac{n^{a-1}}{\Gamma(a)}(\log n)^b$, $\rho \in \{\sigma, \tau\}$.

We set

$$\begin{aligned} P(z,S_{1,b}) &= bz^2(1-z)S_{1,b}^3 + [(1-z) + bz^2(wz-1)]S_{1,b}^2 - (1-z)S_{1,b} \\ Q(z,S_{1,b}) &= z(1-z) \end{aligned}$$

and accordingly $\frac{\partial S_{1,b}}{\partial z} = \frac{P(z,S_{1,b})}{Q(z,S_{1,b})}$ where $z=0$ is a removable singularity since by setting $F(z,S_{1,b}) = \frac{P(z,S_{1,b})}{Q(z,S_{1,b})}$ for $z \neq 0$ and $F(0,1) = 1$, we have $F(z,S_{1,b})$ being analytic in some neighborhood of $z=0$. In what follows, we shall study both, the fixed singularities of $S_{1,b}(z,w)$, which are purely given by eq. (3.5), and the movable singularities of $S_{1,b}(z,w)$ being dependent on the initial condition $S_{1,b}(0,w) = 1$. Since $F(z,S_{1,b})$ is a rational function of $S_{1,b}$ with coefficients which are polynomials of z, we are assured that $S_{1,b}$ only has finitely many fixed singularities, and all movable singularities are either poles or branch points. We start by determining the fixed singularities of $S_{1,b}$. First there is no singularity arising from the coefficients indexed by z from $P(z,S_{1,b})$ and $Q(z,S_{1,b})$, i.e., $z(1-z)$, $1-z$, $1-z+bz^2(wz-1)$ and $bz^2(1-z)$ are analytic on the complex plane $\mathbb{C}$. Secondly there is one singularity $z=1$ arising from $Q(z,S_{1,b}) \equiv 0$ unless $w=1$ since we have already removed the singularity $z=0$ by setting $F(0,1) = \lim_{z\to 0}F(z,S_{1,b}) = 1$. Thirdly there is a singularity $z=1$ arising from the case when both $P(z,S_{1,b})$ and $Q(z,S_{1,b})$ vanish unless $w=1$, i.e., there is a common root of $P(z, S_{1,b}) = 0$ and $Q(z, S_{1,b}) = 0$ except the removable singularity $z = 0$ and $w = 1$. Furthermore, by substituting $S_{1,b} = \frac{1}{T_{1,b}}$ we transform eq. (3.5) into

(3.6)
$$T'_{1,b} = -\frac{T_{1,b}^2 P(z, \frac{1}{T_{1,b}})}{Q(z, \frac{1}{T_{1,b}})} = \frac{(1-z)T_{1,b}^2 - [1 - z + bz^2(wz-1)]T_{1,b} - bz^2(1-z)}{z(1-z)T_{1,b}} = \frac{\bar{P}(z, T_{1,b})}{\bar{Q}(z, T_{1,b})}.$$

By inspection, the fixed singularities of eq. (3.6) are also the fixed singularities of eq. (3.5), and only the common roots of $\bar{P}(z, T_{1,b}) = 0$ and $\bar{Q}(z, T_{1,b}) = 0$ can potentially contribute to singularities. In this case there is no more singularity from $\bar{P}(z, T_{1,b}) = 0$ and $\bar{Q}(z, T_{1,b}) = 0$ except $z = 0$ and $z = 1$, based on which we conclude that the only fixed singularity of $S_{1,b}(z, w)$ is the point at infinity and $z = 1$ unless $w = 1$. Due to Painlevé's determinateness theorem [4], the movable singularities of the solution of eq. (3.5) can only be poles or algebraic branch points. First we claim that the dominant singularity of $S_{1,b}(z, w)$ is unique (single dominant singularity) since $[z^n]S_{1,b}(z, w) = \sum_{k\geq 1} \mathbb{E}_{\#}^{1,b}(\mathcal{S}_{n,k})w^k > 0$ holds for any n and $w \in (1-\epsilon, 1+\epsilon)$, thus the support of $S_{1,b}(z, w)$, which is the set of all n such that $[z^n]S_{1,b}(z, w) \neq 0$, is equal to $\{1, 2\cdots\} = \mathbb{N}$. Thus $S_{1,b}(z, w)$ is aperiodic and therefore its dominant singularity is unique [6]. Now, let $z = \alpha_{1,b}(w) \in \mathbb{R}^+$ denote the unique dominant movable singularity of $S_{1,b}(z, w)$. To ensure a single value for $S_{1,b}(z, w)$ at $z = \alpha_{1,b}(w)$, we consider the $\Delta_{\alpha_{1,b}(w)}$-domain given by

$$\Delta_{\alpha_{1,b}(w)}(M, \phi) = \{z \mid |z| < M, z \neq \alpha_{1,b}(w), |\arg(z - \alpha_{1,b}(w))| > \phi\}$$

where $M > \alpha_{1,b}(w)$ and $0 < \phi < \frac{\pi}{2}$. Let $U_{\alpha_{1,b}(w)}$ be the intersection of $\Delta_{\alpha_{1,b}(w)}(M, \phi)$ and the neighborhood of $\alpha_{1,b}(w)$, i.e.,

$$U_{\alpha_{1,b}(w)} = \{z \mid 0 < |z - \alpha_{1,b}(w)| < r, |\arg(z - \alpha_{1,b}(w))| > \phi\}.$$

Then we have $\lim_{z\to\alpha_{1,b}(w), z\in U_{\alpha_{1,b}(w)}} T_1(z) = 0$ and we transform eq. (3.6) into

$$\frac{\partial z}{\partial T_{1,b}} = T_{1,b} G(T_{1,b}, z) \text{ where} \tag{3.7}$$

$$G(T_{1,b}, z) = \frac{z(1-z)}{(1-z)T_{1,b}^2 - [1 - z + bz^2(wz-1)]T_{1,b} - bz^2(1-z)}, \tag{3.8}$$

with the initial condition $z(0, w) = \alpha_{1,b}(w)$, which corresponds to the fact $S_{1,b}(\alpha_{1,b}(w), w) = \infty$ resp. $T_{1,b}(\alpha_{1,b}(w), w) = 0$. Here $G(T_{1,b}, z)$ is analytic at $(0, \alpha_{1,b}(w))$ and $G(0, \alpha_{1,b}(w)) \neq 0$. By expanding $z(T_{1,b})$ at $T_{1,b} = 0$, we obtain $G(T_{1,b}, z)$ with $z \in U_{\alpha_{1,b}(w)}$ as an infinite polynomial of $T_{1,b}$. Based on this representation, eq. (3.7) has a unique solution of the form

$$1 - \frac{z}{\alpha_{1,b}(w)} = T_{1,b}^2 \cdot (c_0(w) + c_1(w)T_{1,b} + \cdots) \text{ and } c_0(w) \neq 0,$$

which leads to the solution of eq. (3.6), i.e.,

$$T_{1,b}(z, w) = \sum_{j=1}^{\infty} d_j(w) \left(1 - \frac{z}{\alpha_{1,b}(w)}\right)^{\frac{j}{2}} \text{ and } d_1(w) \neq 0,$$

by which we obtain the solution of eq. (3.5)

$$S_{1,b}(z, w) = \left(1 - \frac{z}{\alpha_{1,b}(w)}\right)^{-\frac{1}{2}} \sum_{j=0}^{\infty} e_j(w) \left(1 - \frac{z}{\alpha_{1,b}(w)}\right)^{\frac{j}{2}} \text{ and } e_0(w) = \frac{1}{\alpha_{1,b}(w)}\sqrt{\frac{1}{2b}} \neq 0. \tag{3.9}$$

This solution holds for $z \in U_{\alpha_{1,b}(w)}$ and $z = \alpha_{1,b}(w)$ is the unique dominant algebraic branching point for any $w \in (1-\epsilon, 1+\epsilon)$. Now we shall show eq. (3.9) uniformly holds for any $|w - 1| < \epsilon$. According to Theorem 2, we have for $w \in (1-\epsilon, 1+\epsilon)$

$$\sum_{k\geq 1} \mathbb{E}_{\#}^{1,b}(\mathcal{S}_{n,k})w^k = e_o(w)\Gamma(\frac{3}{2})n^{-\frac{1}{2}}\alpha_{1,b}(w)^{-n}(1 + O(\frac{1}{n})). \tag{3.10}$$

Both sides of eq. (3.10) are analytic at $|w - 1| < \epsilon$ and they coincide for $w \in (1-\epsilon, 1+\epsilon)$, therefore they are identical for $|w - 1| < \epsilon$, namely eq. (3.9) holds uniformly for $|w - 1| < \epsilon$. Finally we need Theorem 3 (Theorem IX.12 in [6]) to prove the Gaussian distribution of $X_n^{c,b}$, which can be phrased as

Theorem 3. *Let $F(z,u)$ be a function that is bivariate analytic at $(z,u)=(0,0)$ and has non-negative coefficients. Assume the following conditions:*

(1) *Analytic perturbation: there exist three functions A, B, C, analytic in a domain $\mathbb{D}=\{|z|\leq r\}\times\{|u-1|<\epsilon\}$, such that for some small r_0 with $0<r_0\leq r$ and $\epsilon>0$, the following representation holds with $\alpha\notin\mathbb{Z}_{\leq 0}$*

$$F(z,u)=A(z,u)+B(z,u)C(z,u)^{-\alpha};$$

furthermore, assume that in $|z|\leq r$, there exists a unique root ρ of the equation $C(z,1)=0$, that this root is simple, and that $B(\rho,1)\neq 0$.

(2) *Non-degeneracy: one has $\partial_z C(\rho,1)\cdot\partial_u C(\rho,1)\neq 0$, ensuring the existence of a non-constant $\rho(u)$ analytic at $u=1$, such that $C(\rho(u),u)=0$ and $\rho(1)=\rho$.*

(3) *Variability $\sigma^*(\frac{\rho(1)}{\rho(u)})\neq 0$ where $\sigma^*(f)=\frac{f''(1)}{f(1)}+\frac{f'(1)}{f(1)}-(\frac{f'(1)}{f(1)})^2$.*

Then the random variable with probability generating function $p_n(u)=\frac{[z^n]F(z,u)}{[z^n]F(z,1)}$ converges in distribution to a Gaussian variable with a speed of convergence that is $O(n^{-\frac{1}{2}})$. The mean μ_n and variance σ_n^2 are asymptotically linear in n.

Here in our case $C(z,u)=1-\frac{z}{\alpha_{1,b}(w)}$ and $\alpha=\frac{1}{2}$ and we can compute the unique dominant singularity $\alpha_{1,b}(w)$ for $w\in(1-\epsilon,1+\epsilon)$ by rk45 methods, which indicates $\alpha_{1,b}(w)$ is not a constant. In particular, we obtain for $b=1,2$, $\alpha_{1,1}(1)\approx 0.597993$ and $\alpha_{1,2}(1)\approx 0.583274$. Furthermore, the probability generating function of $X_n^{1,b}$ is

$$\sum_{k\geq 1}\frac{\mathbb{E}_{\#}^{1,b}(\mathcal{S}_{n,k})}{\mathbb{E}_{\#}^{1,b}(\mathcal{S}_n)}w^k=\frac{e_0(w)}{e_0(1)}\left[\frac{\alpha_{1,b}(1)}{\alpha_{1,b}(w)}\right]^n(1+O(\frac{1}{n})),$$

from which we can calculate the mean $\mathbb{E}(X_n^{1,b})=n\cdot(-\frac{\alpha'_{1,b}(1)}{\alpha_{1,b}(1)})$ and variance $\sigma^2(X_n^{1,b})=n\cdot\sigma^*(\frac{\alpha_{1,b}(1)}{\alpha_{1,b}(w)})$ asymptotically. On the other hand, mean $\mathbb{E}(X_n^{1,b})$ and variance $\sigma^2(X_n^{1,b})$ can also be computed by differentiating at w on both sides of eq. (3.9) and setting $w=1$ afterwards, which, in combination of Theorem 2, gives

$$\begin{aligned}
\mathbb{E}(X_n^{1,1}) &= \frac{[z^n]\frac{\partial S_1(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_1(z,1)}=\frac{[z^n]S_{1,w}(z,1)}{[z^n]S_1(z,1)}=\frac{\Gamma(\frac{1}{2})k_0n^{\frac{1}{2}}\alpha_{1,b}^{-n}}{\Gamma(\frac{3}{2})e_0n^{-\frac{1}{2}}\alpha_{1,b}^{-n}}\left(1+\mathcal{O}(n^{-1})\right)\\
&= 0.1313n+0.1313+\mathcal{O}(n^{-1}),\\
\mathbb{E}(X_n^{1,2}) &= 0.1462n+0.1460+\mathcal{O}(n^{-1}).\\
\sigma(X_n^{1,1}) &= 0.1871\sqrt{n}(1+\mathcal{O}(n^{-1})),\\
\sigma(X_n^{1,2}) &= 0.1857\sqrt{n}(1+\mathcal{O}(n^{-1})).
\end{aligned}$$

Here – and in all the analysis that will follow – the constants have been determined numerically by inspection of precise series coefficients. From the computation we can see variability $\sigma^*(X_n^{1,b})\neq 0$ and Theorem 3 tells us that $X_n^{1,b}$ is Gaussian distributed with mean and variance linear in n. We next consider the case $c=2$ thus setting $c=2$ in eq. (3.2). We find

$$\frac{T_2(z,w)}{z} \quad = \quad b\sum_{n\geq 3}\sum_{k\geq 1}\frac{1}{(n+1)^2}\mathbb{E}_{\#}^{2,b}(\mathcal{S}_{n,k})w^kz^{n+1}+\sum_{n\geq 1}\frac{b}{(n+1)^2}wz^{n+1}.$$

In combination with eq. (3.3), we obtain for $b\geq 1$ the functional identity for $S_2=S_2(z,w)$. Let S_2'', S_2' denote $\frac{\partial^2 S_2(z,w)}{\partial^2 z}$, and $\frac{\partial S_2(z,w)}{\partial z}$, then

$$S_2''=\frac{2}{S_2}(S_2')^2+\frac{1}{z}S_2'+\frac{S_2}{z^2}+(\frac{b}{1-z}(wz-1)-\frac{1}{z^2})S_2^2+bS_2^3. \tag{3.11}$$

By substituting $S_2 = 1/u$, we transform eq. (3.11) into

$$u'' = \frac{1}{z}u' + \left[-\frac{b}{u} + \frac{1}{z^2} - \frac{b(wz-1)}{1-z} - \frac{u}{z^2}\right] \tag{3.12}$$

with initial condition $u(0,w) = 1$. First of all, $z = 0$ is a removable singularity of $u(z,w)$. Furthermore, the major contribution of u'' can only be counteracted by $-\frac{b}{u}$, from which we are assured that the dominant singularity is of log-algebraic type and the singular solution of eq. (3.12) is
(3.13)

$$u(z,w) = (1 - \frac{z}{\alpha_{2,b}(w)}) \left[\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}})\right]^{\frac{1}{2}} \left[a_0 + a_1(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{-1} + \cdots\right]$$

where $w \in (1-\epsilon, 1+\epsilon)$ and $1/u(z,w)$ has nonnegative coefficients. As a result, we can compute

$$\begin{aligned} \frac{1}{z}u' + \left[-\frac{b}{u} + \frac{1}{z^2} - \frac{b(wz-1)}{1-z} - \frac{u}{z^2}\right] &= \log^{-\frac{1}{2}}(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}})\frac{1}{z-\alpha_{2,b}(w)}(\frac{b\alpha_{2,b}(w)}{a_0} + o(1)) \\ u'' &= \log^{-\frac{1}{2}}(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}})\frac{1}{z-\alpha_{2,b}(w)}(\frac{a_0}{2\alpha_{2,b}(w)} + o(1)). \end{aligned}$$

In view of eq. (3.12), comparing the coefficients of the leading term yields,

$$a_0(w) = \sqrt{2b}\,\alpha_{2,b}(w) > 0, \tag{3.14}$$

and we shall analyze the singular expansion of $S_2(z,w)$ based on eq. (3.13), which gives

(3.15)

$$S_2(z,w) = (1 - \frac{z}{\alpha_{2,b}(w)})^{-1} \left[\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}})\right]^{-\frac{1}{2}} (\frac{1}{a_0(w)} + \mathcal{O}(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{-1}),$$

which holds for $z \in U_{\alpha_{2,b}(w)}$ and $z = \alpha_{2,b}(w)$ is the unique dominant singularity for any $w \in (1-\epsilon, 1+\epsilon)$. According to Theorem 2, we derive the n-th coefficients of $S_2(z,w)$ as

$$\sum_{k\geq 1} \mathbb{E}^{2,b}_{\#}(\mathcal{S}_{n,k})w^k = \frac{1}{a_0(w)}(\log n)^{-\frac{1}{2}}(\alpha_{2,b}(w))^{-n},\ n \to \infty. \tag{3.16}$$

Both sides of eq. (3.16) are analytic at $|w-1| < \epsilon$ and they coincide for $w \in (1-\epsilon, 1+\epsilon)$, therefore they are identical for $|w-1| < \epsilon$, and thus eq. (3.15) holds uniformly for $|w-1| < \epsilon$. Finally we will use Theorem 3 with a logarithmic multiplier to prove the Gaussian distribution of $X_n^{2,b}$. Again we can compute the unique dominant singularity $\alpha_{2,b}(w)$ for $w \in (1-\epsilon, 1+\epsilon)$ by rk45 methods, which indicates $\rho(w)$ is not a constant. In particular, we obtain for $b = 1, 2$, $\alpha_{2,1}(1) \approx 0.765120$ and $\alpha_{2,2}(1) \approx 0.680739$. Furthermore, the probability generating function of $X_n^{2,b}$ is

$$\sum_{k\geq 1} \frac{\mathbb{E}^{2,b}_{\#}(\mathcal{S}_{n,k})}{\mathbb{E}^{2,b}_{\#}(\mathcal{S}_n)}w^k = \frac{a_0(1)}{a_0(w)}\left[\frac{\alpha_{2,b}(1)}{\alpha_{2,b}(w)}\right]^n (1 + O(\frac{1}{\log n})),$$

from which we can calculate the mean $\mathbb{E}(X_n^{2,b}) = n \cdot (-\frac{\alpha'_{2,b}(1)}{\alpha_{2,b}(1)})$ and variance $\sigma^2(X_n^{2,b}) = n \cdot \sigma^*(\frac{\alpha_{2,b}(1)}{\alpha_{2,b}(w)})$ asymptotically. On the other hand, mean $\mathbb{E}(X_n^{2,b})$ and variance $\sigma^2(X_n^{2,b})$ can also be computed by differentiating at w on both sides of eq. (3.9) and setting $w = 1$ afterwards, which,

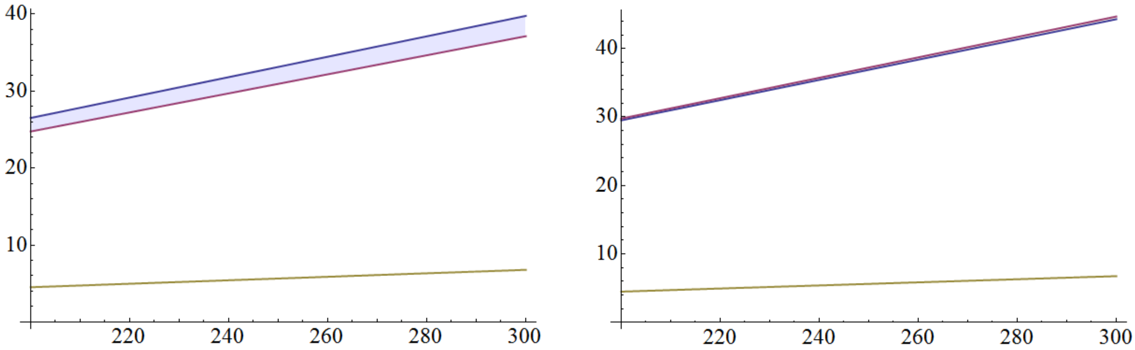

FIGURE 5. Plots of the average number of hairpins as a function of the structure's size n within our polymer-zeta model ($b = 1$ left, $b = 2$ right). The blue (resp. red) line corresponds to case $c = 1$ (resp. $c = 2$), the greenish line shows the behavior of native RNA secondary structures (as derived from the stochastic model from [11]).

in combination of Theorem 2, gives

$$
\begin{aligned}
\mathbb{E}(X_n^{2,1}) &= \frac{[z^n]\frac{\partial S_2(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_2(z,1)} = \frac{[z^n]S_{2,w}(z,1)}{[z^n]S_2(z,1)} = \frac{\lambda_{2,b}\, n\, (\log n)^{-\frac{1}{2}}\alpha_{2,b}^{-n}}{(\log n)^{-\frac{1}{2}}\alpha_{2,b}^{-n}}\left(1+\mathcal{O}((\log n)^{-1})\right) \\
&= 0.1202\, n + 0.0376 \frac{n}{\log n} + \mathcal{O}(\frac{n}{\log^2 n}), \\
\mathbb{E}(X_n^{2,2}) &= 0.1458\, n + 0.0259 \frac{n}{\log n} + \mathcal{O}(\frac{n}{\log^2 n}). \\
\sigma(X_n^{2,1}) &= 0.2047\sqrt{n}(1+\mathcal{O}((\log n)^{-1})), \\
\sigma(X_n^{2,2}) &= 0.2022\sqrt{n}(1+\mathcal{O}((\log n)^{-1})).
\end{aligned}
$$

From the computation we can see variability $\sigma^*(X_n^{2,b}) \neq 0$ and Theorem 3 tells us that $X_n^{2,b}$ is Gaussian distributed with mean and variance linear in n. □

Figure 5 shows two plots of our results for the average number of hairpins for $b = 1$ (left) and $b = 2$ (right). Assuming a continuous transition of the average number of hairpins when changing parameter c from 1 to 2, the behavior for the case $c = 1.5$ (or $c = 1.47$) – which corresponds to the *real* polymer-zeta property – should be located in the shaded area spanned by the two lines depicted. As we can see, the number of hairpins in any of our polymer-zeta models is ways to large compared to the native behavior.

4. Average length of hairpin-loops

Let $Y_n^{c,b}$ denote the random variable counting the total length of all hairpin-loops in a secondary structure of size n in our polymer-zeta model with parameters c, b, and $\mathbb{E}(Y_n^{c,b})$ (resp. $\sigma(Y_n^{c,b})$) be the expectation (resp. standard deviation) of $Y_n^{c,b}$. We have:

Theorem 4. *Under the assumption of the (c,b)-polymer-zeta-model, $c \in \{1, 2\}$, the length of a hairpin-loop in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\frac{\mathbb{E}(Y_n^{1,b})}{\mathbb{E}(X_n^{1,b})} = y_{1,b}(1+\mathcal{O}(n^{-1})), \quad \frac{\mathbb{E}(Y_n^{2,b})}{\mathbb{E}(X_n^{2,b})} = y_{2,b}(1+\mathcal{O}((\log n)^{-1}))$$

and standard deviation

$$\sigma(Y_n^{1,b}) = \frac{y_{1,b}}{\sqrt{n}}(1+\mathcal{O}(n^{-1})), \quad \sigma(Y_n^{2,b}) = \frac{y_{2,b}}{\sqrt{n}}(1+\mathcal{O}((\log n)^{-1}))$$

where $y_{c,b}$ and $y_{c,b}$ are positive constants. For $b = 1, 2$, we have

$$(y_{1,1}, y'_{1,1}) \approx (1.7262, 3.6003) \qquad (y_{1,2}, y'_{1,2}) \approx (1.5291, 3.5570)$$
$$(y_{2,1}, y'_{2,1}) \approx (1.7367, 3.3058) \qquad (y_{2,2}, y'_{2,2}) \approx (1.5467, 3.0230)$$

Since the proof of this and all the subsequent theorems pretty much proceed along the same lines as the previous one, we have moved them to the appendix. We assume this paper to be a nicer read when leaving out the ever repeating same mathematical arguments, focusing on the interpretation of our findings.

Compared to the hairpin length observed for stochastic model from [11], we observe no choice for the parameters c, b which comes close to a fit. This might be related to the assumption of a minimal hairpin-loop length of 1 in our analysis. However, the largest loops are observed for $c = 2$, $b = 1$ where the averaged length is given by 1.7367. Adding a constant of 2 – which could be assumed to overestimate the result for a polymer-zeta model with a minimal loop length of 3 – would still result in a too small length of loop. Thus we can conclude, the our model will not behave realistically with respect to the expected length of hairpin loops.
Considering the total number of unpaired nucleotides residing in hairpin-loops by multiplying expected length and number brings model and real world molecules closer together. We find about $0.1667n$ unpaired nucleotides in hairpins for the native and about $0.2257n$ for the $(1, 2)$-polymer-zeta structures. Thus, in total the number of unpaired nucleotides inside hairpin-loops is overestimated by about 6% within our model. It will be interesting to see, how the overall number of unpaired position behaves in comparison of the (c, b)-polymer-zeta-model and native molecules.

5. Number of unpaired bases

Let $U_n^{c,b}$ denote the random variable counting the number of unpaired bases found in a secondary structure of size n according to the (c, b)-polymer-zeta-model, and $\mathbb{E}(U_n^{c,b})$ (resp. $\sigma(U_n^{c,b})$) be the expectation (resp. standard deviation) of $U_n^{c,b}$. We have:

Theorem 5. *Under the assumption of the (c, b)-polymer-zeta-model, $c \in \{1, 2\}$, the number of unpaired bases in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\mathbb{E}(U_n^{1,b}) = u_{1,b}n(1 + \mathcal{O}(n^{-\frac{1}{2}})), \quad \mathbb{E}(U_n^{2,b}) = u_{2,b}n(1 + \mathcal{O}((\log n)^{-\frac{1}{2}}))$$

and standard deviation

$$\sigma(U_n^{1,b}) = u'_{1,b}\sqrt{n}(1 + \mathcal{O}(n^{-\frac{1}{2}})), \quad \sigma(U_n^{2,b}) = u'_{2,b}\sqrt{n}(1 + \mathcal{O}((\log n)^{-\frac{1}{2}}))$$

where $u_{c,b}$ and $u'_{c,b}$ are positive constants. For $b = 1, 2$, we have

$$(u_{1,1}, u'_{1,1}) \approx (0.5918, 0.4410) \quad (u_{1,2}, u'_{1,2}) \approx (0.5266, 0.4256)$$
$$(u_{2,1}, u'_{2,1}) \approx (0.7046, 0.4605) \quad (u_{2,2}, u'_{2,2}) \approx (0.6323, 0.4507).$$

It is obvious that for all the choices (c, b) considered here, the number of unpaired bases is too large compared to native RNA molecules. There we observe about 45% of unpaired nucleotides for tRNA and roughly 52% for large subunit rRNAs (based on which the results of [11] have been derived). In connection with only about 6% of *supernumerous* unpaired positions within the hairpin-loops, the total of up to 25% superfluous unpaired nucleotides could be explained by either

- other loops like bulges or interior-loops being longer than in native molecules, or
- loops in general being observed more often (as already proven for hairpin-loops) within our model than in nature.

The second possibility would imply that secondary structures look scattered in our polymer-zeta model in the sense that stems are way too short and often interrupted by runs of unpaired nucleotides. We will continue to analyze structural motifs like bulges and interior loops in order to identify which explanation applies.

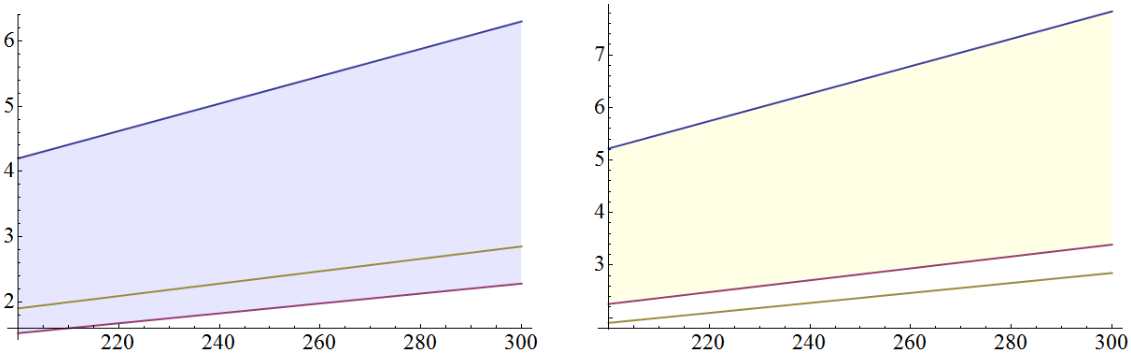

FIGURE 6. Plots of the average number of bulges as a function of the structure's size n within our polymer-zeta model ($b = 1$ left, $b = 2$ right). The blue (resp. red) line corresponds to case $c = 1$ (resp. $c = 2$), the greenish line shows the behavior of native RNA secondary structures (as derived from the stochastic model from [11]).

6. The average number of bulges

A bulge of length ℓ is defined as a sequence of $\ell \geq 1$ unpaired positions $s_{b+1}, s_{b+2}, \cdots, s_{b+\ell}$ that are in between two nested arcs $(s_a, s_{b+\ell+1})$ and (s_{a+1}, s_b), or arcs $(s_{b+\ell+1}, s_a)$ and (s_b, s_{a+1}). For $\mathcal{S}_{n,k}$ the class of secondary structures of size n with k bulges $n \geq 3k + 3$ must hold. Let $B_n^{c,b}$ denote the random variable counting the number of bulges found in a secondary structure of size n according to the (c, b)-polymer-zeta-model, and $\mathbb{E}(B_n^{c,b})$ (resp. $\sigma(B_n^{c,b})$) be the expectation (resp. standard deviation) of $B_n^{c,b}$. We find:

Theorem 6. *Under the assumption of the (c, b)-polymer-zeta-model, $c \in \{1, 2\}$, the number of bulges in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\mathbb{E}(B_n^{1,b}) = \beta_{1,b}\, n(1 + \mathcal{O}(n^{-1})), \quad \mathbb{E}(B_n^{2,b}) = \beta_{2,b}\, n(1 + \mathcal{O}((\log n)^{-1}))$$

and standard deviation

$$\sigma(B_n^{1,b}) = \beta'_{1,b}\sqrt{n}(1 + \mathcal{O}(n^{-1})), \quad \sigma(B_n^{2,b}) = \beta'_{2,b}\sqrt{n}(1 + \mathcal{O}((\log n)^{-1}))$$

where $\beta_{c,b}$ and $\beta'_{c,b}$ are positive constants. For $b = 1, 2$, we have

$$(\beta_{1,1}, \beta'_{1,1}) \approx (0.0210, 0.1379) \quad (\beta_{1,2}, \beta'_{1,2}) \approx (0.0261, 0.1517)$$
$$(\beta_{2,1}, \beta'_{2,1}) \approx (0.0076, 0.0848) \quad (\beta_{2,2}, \beta'_{2,2}) \approx (0.0113, 0.1024).$$

Figure 6 shows a plot of the averages just derived in comparison to the native behavior of RNA molecules (as indicated by the results from [11]). We observe that for $b = 1$, the behavior of our polymer-zeta model nicely fits with natural RNAs; for $b = 2$ the model slightly overestimates the number of bulges. Thus the number of bulges cannot explain the large spread for the number of unpaired positions. Next we will consider the length of bulges.

7. The average length of bulges

Let $L_n^{c,b}$ denote the random variable counting the total length of all the bulges (-loop) found in a secondary structure of size n according to the (c, b)-polymer-zeta-model, and $\mathbb{E}(L_n^{c,b})$ (resp. $\sigma(L_n^{c,b})$) be the expectation (resp. standard deviation) of $L_n^{c,b}$. We have:

Theorem 7. *Under the assumption of the (c, b)-polymer-zeta-model, $c \in \{1, 2\}$, the length of a bulge in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\frac{\mathbb{E}(L_n^{1,b})}{\mathbb{E}(B_n^{1,b})} = l_{1,b}(1 + \mathcal{O}(n^{-1})), \quad \frac{\mathbb{E}(L_n^{2,b})}{\mathbb{E}(B_n^{2,b})} = l_{2,b}(1 + \mathcal{O}((\log n)^{-1}))$$

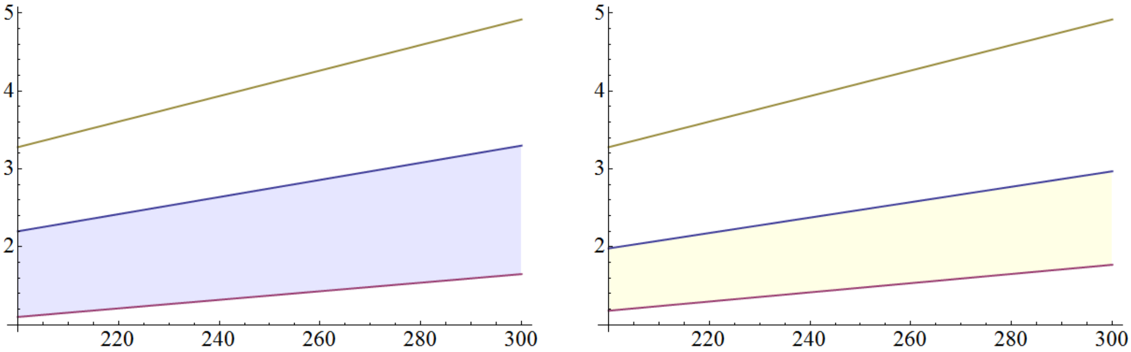

FIGURE 7. Plot of the average number of interior loops as a function of the structure's size n within our polymer-zeta model ($b = 1$ left, $b = 2$ right). The blue (resp. red) line corresponds to case $c = 1$ (resp. $c = 2$), while those lines connected by the blue (resp. yellow) shading correspond to case $b = 1$ (resp. $b = 2$). The greenish line shows the behavior of native RNA secondary structures (as derived from the stochastic model from [11]).

and standard deviation

$$\frac{\sigma(L_n^{1,b})}{\mathbb{E}(B_n^{1,b})} = \frac{l'_{1,b}}{\sqrt{n}}(1 + \mathcal{O}(n^{-1})), \quad \frac{\sigma(L_n^{2,b})}{\mathbb{E}(B_n^{2,b})} = \frac{l'_{2,b}}{\sqrt{n}}(1 + \mathcal{O}((\log n)^{-1}))$$

where $l_{c,b}$ and $l'_{c,b}$ are positive constants. For $b = 1, 2$,

$$(l_{1,1}, l'_{1,1}) \approx (2.0476, 16.7438) \quad (l_{1,2}, l'_{1,2}) \approx (1.7625, 12.3834)$$
$$(l_{2,1}, l'_{2,1}) \approx (2.4079, 35.4115) \quad (l_{2,2}, l'_{2,2}) \approx (2.0354, 23.3449).$$

Comparing our findings to the average bulge length observed for native molecules we have to conclude an almost realistic behavior of our polymer-zeta model; it slightly overestimates the length of bulges for all choices of (c, b). However, the slightly larger length and number of bulges in our model are not sufficient to explain the overall too large number of unpaired positions. Even worth, their effect is counteracted by the interior loops which will turn out to be shorter.

8. AVERAGE NUMBER OF INTERIOR LOOPS

A single interior loop consists of two non-empty runs of unpaired bases $s_{a+1}, \cdots, s_{a+\ell_1}$, and $s_{b+1}, \cdots, s_{b+\ell_2}$, $a < b$, together with the base pairs $(s_a, s_{b+\ell_2+1})$ and $(s_{a+\ell_1+1}, s_b)$. Let $I_n^{c,b}$ denote the random variable counting the number of interior loops found in a secondary structure of size n according to the (c, b)-polymer-zeta-model, and $\mathbb{E}(I_n^{c,b})$ (resp. $\sigma(I_n^{c,b})$) be the expectation (resp. standard deviation) of $I_n^{c,b}$. The following theorem holds:

Theorem 8. *Under the assumption of the (c, b)-polymer-zeta-model, $c \in \{1, 2\}$, the number of interior loops in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\mathbb{E}(I_n^{1,b}) = i_{1,b}\, n(1 + \mathcal{O}(n^{-1})), \quad \mathbb{E}(I_n^{2,b}) = i_{2,b}\, n(1 + \mathcal{O}((\log n)^{-1}))$$

and standard deviation

$$\sigma(I_n^{1,b}) = i'_{1,b}\sqrt{n}(1 + \mathcal{O}(n^{-1})), \quad \sigma(I_n^{2,b}) = i'_{2,b}\sqrt{n}(1 + \mathcal{O}((\log n)^{-1}))$$

where $i_{c,b}$ and $i'_{c,b}$ are positive constant. For $b = 1, 2$, we have

$$(i_{1,1}, i'_{1,1}) \approx (0.0110, 0.0996) \quad (i_{1,2}, i'_{1,2}) \approx (0.0099, 0.0954)$$
$$(i_{2,1}, i'_{2,1}) \approx (0.0055, 0.0706) \quad (i_{2,2}, i'_{2,2}) \approx (0.0059, 0.0739).$$

Figure 7 shows a plot of the averages just presented together with the native behavior of the number of interior loops. We observe that in the (c, b)-polymer-zeta-model (for all choices of (c, b)), the number of interior loops is observably smaller than in native molecules. Accordingly, if the average length of interior loops within the model is not notably larger than for real world molecules the large number of unpaired bases sill remains unexplained.

9. The average length of interior loops

We define the length of an interior loop to be the sum of the sizes of both its runs of unpaired positions (given by $\ell_1 + \ell_2$ according to our definition of an interior loop). Let $P_n^{c,b}$ denote the random variable counting the total length of interior loops found in a secondary structure of size n according to the (c,b)-polymer-zeta-model, and $\mathbb{E}(P_n^{c,b})$ (resp. $\sigma(P_n^{c,b})$) be the expectation (resp. standard deviation) of $P_n^{c,b}$. We have:

Theorem 9. *Under the assumption of the (c,b)-polymer-zeta-model, $c \in \{1, 2\}$, the average length of a single interior loop found in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\frac{\mathbb{E}(P_n^{1,b})}{\mathbb{E}(I_n^{1,b})} = p_{1,b}(1 + \mathcal{O}(n^{-1})), \quad \frac{\mathbb{E}(P_n^{2,b})}{\mathbb{E}(I_n^{2,b})} = p_{2,b}(1 + \mathcal{O}((\log n)^{-1}))$$

and standard deviation

$$\frac{\sigma(P_n^{1,b})}{\mathbb{E}(I_n^{1,b})} = p'_{1,b}\frac{1}{\sqrt{n}}(1 + \mathcal{O}(n^{-1})), \quad \frac{\sigma(P_n^{2,b})}{\mathbb{E}(I_n^{2,b})} = p'_{2,b}\frac{1}{\sqrt{n}}(1 + \mathcal{O}((\log n)^{-1}))$$

where $p_{c,b}$ and $p'_{c,b}$ are positive constants. For $b = 1, 2$, we have

$$(p_{1,1}, p'_{1,1}) \approx (4.2364, 43.5036) \quad (p_{1,2}, p'_{1,2}) \approx (3.6162, 38.7243)$$
$$(p_{2,1}, p'_{2,1}) \approx (5.3455, 81.0401) \quad (p_{2,2}, p'_{2,2}) \approx (4.4068, 63.1238).$$

In summary for all choices of (c,b) considered here, both the number and the length of interior loops within our polymer-zeta model is smaller than observed for real RNA molecule. Accordingly, interior loops cannot explain the large number of unpaired nucleotides observed in our model.

10. The average number of multiloops

A multiloop results, if at least two irreducible secondary (sub-) structures are enclosed by an arc. The number of irreducible structures below that arc plus one is called the degree of the multiloop. Let $M_n^{c,b}$ be the random variable counting the number of multiloops found in a secondary structure of size n according to the (c,b)-polymer-zeta-model, and $\mathbb{E}(M_n^{c,b})$ (resp. $\sigma(M_n^{c,b})$) be the expectation (resp. standard deviation) of $M_n^{c,b}$. We have:

Theorem 10. *Under the assumption of the (c,b)-polymer-zeta-model, $c \in \{1, 2\}$, the number of multiloops found in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\mathbb{E}(M_n^{1,b}) = m_{1,b}^{(1)} n + m_{1,b}^{(2)} + \mathcal{O}(n^{-1}), \quad \mathbb{E}(M_n^{2,b}) = m_{2,b}^{(1)} n + m_{2,b}^{(2)} \frac{n}{\log n} + \mathcal{O}(\frac{n}{\log^2 n})$$

and standard deviation

$$\sigma(M_n^{1,b}) = m'_{1,b}\sqrt{n}(1 + \mathcal{O}(n^{-1})), \quad \sigma(M_n^{2,b}) = m'_{2,b}\sqrt{n}(1 + \mathcal{O}((\log n)^{-1}))$$

where $m_{c,b}$ and $m'_{c,b}$ are positive constants. For $b = 1, 2$, we have

$$(m_{1,1}^{(1)}, m_{1,1}^{(2)}, m'_{1,1}) \approx (0.0330, -0.4683, 0.1188) \quad (m_{1,2}^{(1)}, m_{1,2}^{(2)}, m'_{1,2}) \approx (0.0390, -0.4618, 0.1277)$$
$$(m_{2,1}^{(1)}, m_{2,1}^{(2)}, m'_{2,1}) \approx (0.0112, 0.0024, 0.0730) \quad (m_{2,2}^{(1)}, m_{2,2}^{(2)}, m'_{2,2}) \approx (0.0220, -0.0521, 0.0872).$$

Figure 8 compares the average number of multiloops in our polymer-zeta model to that of native RNA molecules. Only the behavior of the $(2,1)$-polymere-zeta-model comes close to the native characteristics. For all the other parameter choices, model and reality are in no agreement at all.

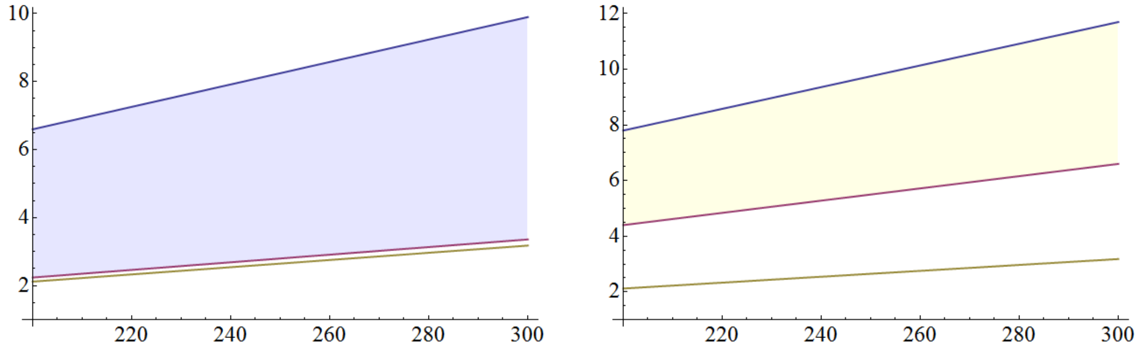

Figure 8. Plot of the average number of multiloops as a function of the structure's size n within our polymer-zeta model ($b = 1$ left, $b = 2$ right). The blue (resp. red) line corresponds to case $c = 1$ (resp. $c = 2$), while those lines connected by the blue (resp. yellow) shading correspond to case $b = 1$ (resp. $b = 2$). The greenish line shows the behavior of native RNA secondary structures (as derived from the stochastic model from [11]).

11. Average degree of multiloops

As already explained before, the degree of a multiloop is given by one plus the number of irreducible substructures embedded below its closing arc. Accordingly, let $D_n^{c,b}$ denote the random variable counting the total degree of all the multiloops found in a secondary structure of size n according to the (c, b)-polymer-zeta-model, and $\mathbb{E}(D_n^{c,b})$ (resp. $\sigma(D_n^{c,b})$) be the expectation (resp. standard deviation) of $D_n^{c,b}$. We find the following theorem:

Theorem 11. *Under the assumption of the (c, b)-polymer-zeta-model, $c \in \{1, 2\}$, the degree of a multiloop found in a secondary structure of size n is asymptotically Gaussian distributed with mean*

$$\frac{\mathbb{E}(D_n^{1,b})}{\mathbb{E}(M_n^{1,b})} = d_{1,b}(1 + \mathcal{O}(n^{-\frac{1}{2}})), \qquad \frac{\mathbb{E}(D_n^{2,b})}{\mathbb{E}(M_n^{2,b})} = d_{2,b}(1 + \mathcal{O}((\log n)^{-\frac{1}{2}}))$$

and standard deviation

$$\frac{\sigma(D_n^{1,b})}{\mathbb{E}(M_n^{1,b})} = d'_{1,b}\frac{1}{\sqrt{n}}(1 + \mathcal{O}(n^{-\frac{1}{2}})), \qquad \frac{\sigma(D_n^{2,b})}{\mathbb{E}(M_n^{2,b})} = d'_{2,b}\frac{1}{\sqrt{n}}(1 + \mathcal{O}((\log n)^{-\frac{1}{2}}))$$

where $d_{c,b}$ and $d'_{c,b}$ are positive constants. For $b = 1, 2$, we have

$$(d_{1,1}, d'_{1,1}) \approx (5.9848, 14.0273) \qquad (d_{1,2}, d'_{1,2}) \approx (5.5615, 12.6077)$$
$$(d_{2,1}, d'_{2,1}) \approx (12.5536, 25.9643) \qquad (d_{2,2}, d'_{2,2}) \approx (8.3636, 16.3182).$$

We observe a rather realistic behavior of our model which for $c = 1$ and $b \in \{1, 2\}$ underestimates the average degree of a multiloop by only about $\frac{3}{10}$.

However, we so far have no good explanation for the large number of unpaired nucleotides as well as the large number of hairpins in our polymer-zeta model. Therefore, we continue our analysis by studying the structure of the exterior loop. Here, the number and length of single stranded regions are of interest.

12. The total length of single stranded regions of the exterior loop

In a secondary structure the exterior loop consists of all positions that are not enclosed by any arc (does not lie in between any base pair). Accordingly, the exterior loop is not empty if any only if we are not dealing with an irreducible structure. On the other hand, a non-empty exterior loop consists of an alternation of single stranded (unpaired) nucleotides and irreducible structures, where for the latter the outermost arc is assumed to be part of the exterior loop. Note, that this way of decomposing an arbitrary secondary structure along its exterior loop gave rise to eq. (3.3).

Now let $E_n^{c,b}$ denote the random variable counting the total number of unpaired positions residing in the exterior loop of a secondary structure of size n according to the (c, b)-polymer-zeta-model, and $\mathbb{E}(E_n^{c,b})$ (resp. $\sigma(E_n^{c,b})$) be the expectation (resp. standard deviation) of $E_n^{c,b}$. Here the limiting distribution of $E_n^{c,b}$ is different to the parameters we discussed before since the dominant singularity is a constant for $|w - 1| < \epsilon$. However, we can explicitly estimate any r-factorial moment of $E_n^{c,b}$. We have:

Theorem 12. *Under the assumption of the (c, b)-polymer-zeta-model, $c \in \{1, 2\}$, the average number of unpaired positions residing in the exterior loop of a secondary structure of size n is asymptotically given by*

$$\mathbb{E}(E_n^{1,b}) = r_{1,b} n^{\frac{1}{2}}(1 + \mathcal{O}(n^{-\frac{1}{2}})), \quad \mathbb{E}(E_n^{2,b}) = r_{2,b} n (\log n)^{-\frac{1}{2}}(1 + \mathcal{O}((\log n)^{-1}))$$

with standard deviation

$$\sigma(E_n^{1,b}) = r'_{1,b}\, n^{\frac{1}{2}}(1 + \mathcal{O}(n^{-\frac{1}{2}})), \quad \sigma(E_n^{2,b}) = r'_{2,b}\, n (\log n)^{-\frac{1}{2}}(1 + \mathcal{O}((\log n)^{-1}))$$

where $r_{c,b}$ and $r'_{c,b}$ are positive constants. Furthermore, for $c = \{1, 2\}$, the r-th factorial moment of $E_n^{c,b}$ is

$$\mathbb{E}(E_n^{1,b}(E_n^{1,b} - 1) \cdots (E_n^{1,b} - r + 1)) = r!(2b)^{-\frac{r}{2}} \frac{(n - r)^{\frac{r+1}{2}-1}}{n^{-\frac{1}{2}}} \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{r+1}{2})}(1 + O(\frac{1}{n}))$$

$$\mathbb{E}(E_n^{2,b}(E_n^{2,b} - 1) \cdots (E_n^{2,b} - r + 1)) = r!(2b)^{-\frac{r}{2}} \frac{(n - r)^r}{\Gamma(r + 1)} \frac{(\log(n - r))^{-\frac{r+1}{2}}}{(\log n)^{-\frac{1}{2}}}(1 + O(\frac{1}{\log n})).$$

where for $b = 1, 2$, we have

$$(r_{1,1}, r'_{1,1}) \approx (1.2066, 0.6842) \quad (r_{1,2}, r'_{1,2}) \approx (0.8668, 0.5254)$$
$$(r_{2,1}, r'_{2,1}) \approx (0.7978, 0.2411) \quad (r_{2,2}, r'_{2,2}) \approx (0.5974, 0.2000).$$

Note that for the first time during our analysis we observe a change of the rate of grows when switching from $c = 1$ to $c = 2$. Nevertheless, as claimed for all the other parameters, the behavior for $c = 1.5$, i.e., in case of the *true* parameter choice as determined by theory, smoothly fits into the interval spanned by our formulæ, see Figure 9. As a consequence, we have to assume the average number of unpaired bases within the exterior loop to lie in the interval $[\sqrt{n}, n/\sqrt{\log n}]$ (modulo constant factors). This is an important observation since for native secondary structures the total number of unpaired bases in the exterior loop behaves different – here quite often long spanning interactions are of importance which give rise a short 5'–3' distance for the molecule and to fewer single strands in the exterior loop; however long spanning interactions are rather unlikely in the polymer-zeta model. For instance, LSU rRNAs with an average length of 2311 (as derived from the database of Wuyts [15]) show in the mean about 112 unpaired nucleotides within the exterior loop, see Figure 9. For tRNA – even if being a rather special family of molecules – we observe an average of 2.2 [13]. In connection with the rate of growth observed for $c = 1.5$ by simulations and in comparison to the statistics derived from databases, we are willing to believe that the number of unpaired positions is notably overestimated within our polymer-zeta model. Accordingly, the large number of unpaired positions in our model can most likely be explained by the growing lengths and/or number of the single stands in the exterior loop. This conclusion also explains the large number of hairpin loops observed .

Finally, let $P_n^{c,b}$ denote the random variable counting the total number of irreducible structures residing in the exterior loop of a secondary structure of size n according to the (c, b)-polymer-zeta-model, and $\mathbb{E}(P_n^{c,b})$ (resp. $\sigma(P_n^{c,b})$) be the expectation (resp. standard deviation) of $P_n^{c,b}$.

Theorem 13. *Under the assumption of the (c, b)-polymer-zeta-model, $c \in \{1, 2\}$, the average degree of the exterior found in a secondary structure of size n is asymptotically given by*

$$\mathbb{E}(P_n^{1,b}) = p_{1,b} n^{\frac{1}{2}}(1 + \mathcal{O}(n^{-1})), \quad \mathbb{E}(P_n^{2,b}) = p_{2,b} n (\log n)^{-\frac{1}{2}}(1 + \mathcal{O}((\log n)^{-1}))$$

with standard deviation

$$\sigma(P_n^{1,b}) = p'_{1,b}\, n^{\frac{1}{2}}(1 + \mathcal{O}(n^{-1})), \quad \sigma(P_n^{2,b}) = p'_{2,b}\, n^{\frac{1}{2}} (\log n)^{-\frac{1}{4}}(1 + \mathcal{O}((\log n)^{-1}))$$

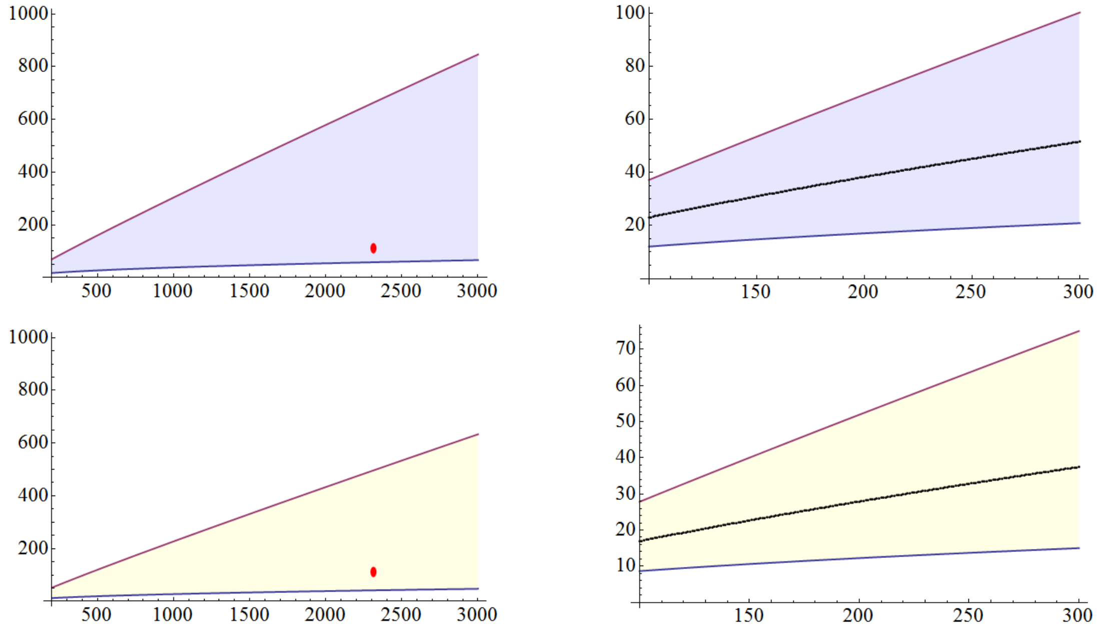

FIGURE 9. Plot of the average number of unpaired nucleotides in the exterior loop as a function of the structure's size n within our polymer-zeta model. The two upper (resp. lower) graphics show the case $b = 1$ (resp. $b = 2$) where the blue (resp. red) line in each case corresponds to case $c = 1$ (resp. $c = 2$). The red dots in the left plots show the average of 112 unpaired nucleotides observed for the exterior loop of LSU rRNAs (as determined from [15]; the corresponding average size of molecules is 2311). The dotted black lines of the right plots correspond to simulation results performed for the case $c = 1.5$ and the choices $b = 1$ (top) and $b = 2$ (bottom).

where $p_{c,b}$ and $p'_{c,b}$ are positive constants. Furthermore, for $c = \{1, 2\}$, the r-th factorial moment of $P_n^{c,b}$ is

$$\begin{aligned}
\mathbb{E}(P_n^{1,b}(P_n^{1,b}-1)\cdots(P_n^{1,b}-r+1)) &= \frac{[z^n]\partial_w^r S_1(z,w)|_{w=1}}{[z^n]S_1(z,1)} \\
&= r!(2b)^{-\frac{r}{2}}(\frac{1}{\alpha_{1,b}}-1)^r n^{\frac{r}{2}}\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{r+1}{2})}(1+O(\frac{1}{n})), \\
\mathbb{E}(P_n^{2,b}(P_n^{2,b}-1)\cdots(P_n^{2,b}-r+1)) &= \frac{[z^n]\partial_w^r S_2(z,w)|_{w=1}}{[z^n]S_2(z,1)} \\
&= r!(2b)^{-\frac{r}{2}}(\frac{1}{\alpha_{2,b}}-1)^r \frac{n^r}{\Gamma(r+1)}(\log n)^{-\frac{r}{2}},
\end{aligned}$$

where for $b = 1, 2$, we have

$$\begin{aligned}
&(p_{1,1}, p'_{1,1}) \approx (0.8426, 0.4404) \quad (p_{1,2}, p'_{1,2}) \approx (0.6332, 0.3310), \\
&(p_{2,1}, r'_{2,1}) \approx (0.2171, 0.4659) \quad (p_{2,2}, p'_{2,2}) \approx (0.2345, 0.4842).
\end{aligned}$$

Remark: With this result, we can show that in our model the so-called conservation law holds. According to that law, the number of hairpins should equal the number of multiloops times the average number of irreducible structures within the multiloop minus one (i.e. according to our definition the average degree of a multiloop minus two) plus the degree of the exterior loop. With the notation used in this paper, we have to consider

$$\mathbb{E}(X_n^{c,b}) = \mathbb{E}(D_n^{c,b}) - 2\mathbb{E}(M_n^{c,b}) + \mathbb{E}(P_n^{c,b})$$

and find for example in case of $(c, b) = (1, 1)$ (using the representations for the respective expected values as given in the proofs found in the Appendix where one can partly find lower order terms, too)

$$\begin{aligned}&\mathbb{E}(X_n^{1,1}) - \mathbb{E}(D_n^{1,1}) + 2\mathbb{E}(M_n^{1,1}) - \mathbb{E}(P_n^{1,1})\\ &= 0.1313n - 0.1975n + 2 \times 0.0330n + 0.8428\sqrt{n} - 0.8425\sqrt{n} \approx 0.\end{aligned}$$

Note that we cannot assume to observe a precise value of 0 since the constants within our asymptotic formulae have been determined by numerical means.

13. Conclusions

In this paper we have shown, that in [14] the authors made faulty use of the polymer-zeta model when analyzing their algorithm CandidateFold and the efficiency of sparsification. Their mistake was to apply unconditioned probabilities where a conditioning on a base pair to constitute the outermost pair of an irreducible structure would have been needed. In order to see, if it is realistic to assume that such conditional probabilities still behave in a polymer-zeta style, i.e., behave like $\frac{b}{d^c}$ for some constants b and c, we introduce a corresponding probability model for RNA secondary structures and examined the resulting expected shape of RNA foldings. To this end, for different choices of (c, b) we determined the average behavior of various structural motives like the number and length of hairpin-, bulge- and interior-loops as well as the number of unpaired bases, the number and average degree of multiloops and the number of unpaired positions residing in the exterior loop. We also proved related standard deviations and for most of the considered parameters to follow a Gaussian distribution.

Compared to the results from [11] – which have proven to nicely reflect the native behavior of RNA – and in light of some statistics computed from RNA databases we observe many parameters in our model to behave almost realistic with respect to both, rate of growth and constants involved. For the number of unpaired bases and the number and length of hairpin loops we experience the opposite: the first two being too large, the third ways too small. Nevertheless, the rate of grows in those three cases is still equal for model and native molecules. To this end, the average number of unpaired nucleotides residing in the exterior loop is special. Here we observe in the model a growth depending on n opposed to a (presumably) constant behavior in real RNAs. This result however also nicely explains the overestimated number of hairpin loops: All our findings make perfect sense if we assume a typical secondary structure in our polymer-zeta model to mostly behave like a sequence of (relatively small) cloverleaf like structures (to be understood as synonym for simple structures without much nesting) emerging one after the other out of the structure's backbone (i.e., emerging from the exterior loop). Accordingly, a large number of bases reside in lots of single strands ($\sqrt{n}$ many as we have proved) in the exterior loop and – because of missing nesting – almost all stems contribute to the number of hairpins. However, parameters like the number and length of bulges and interior loops or the degree of multiloops can still behave realistic (inside the cloverleaf like structures). This also coincides with the discussion from [14], where mRNA data, for which most parts are unpaired and only seldom a simple stem or hairpin can be observed, has been considered. All in all, we conclude that for most families of RNA our polymer-zeta model cannot be assumed realistic. This discovery brings us back to the motivation of our analysis taken from [14]. We conclude that for most families of RNA it is unrealistic to assume the conditional probabilities (as needed to conclude a quadratic running time for CandidateFold) for span d base pairs to behave like $\frac{b}{d^c}$.

Of course, one could think of choices for b and c other than those assumed for our analysis, hoping for a realistic behavior of the corresponding polymer-zeta model. However, the following needs to be considered:

(1) A choice for $c < 1$ would increase the probability for long distance interactions, but at the same time implies a running-time for the CandidateFold algorithm strictly worse than n^2. Therefore, this should not be considered a valid choice for our application.
(2) Choices $c > 2$ will imply an even smaller probability for long distance interactions and thus would lead to a continued degeneration of the exterior loop and a growth of the already too large number of unpaired positions.
(3) Large values for b may be used to introduce high pairing probabilities even for far apart nucleotides. However, since b is constant this can hardly counteract varying sizes of molecules and the corresponding grows of potential distances of two nucleotide positions. Even worse, large values for b imply values strictly larger 1 for some pairing *probabilities* b/d^c. To work around this problem, the minimal hairpin-loop length must be chosen at least $b^{1/c}$, which should not be considered beneficial for the model's quality.

We conclude that the (c, b)-polymer-zeta-model can not be assumed to go along with native families of RNA even for parameter choices not investigated here in detail.

To the best of our knowledge, this is the first paper presenting an analysis of structural parameters for random RNA secondary structures in our polymer-zeta model. Even if we attacked quite a number of parameters, some questions remain open. Most importantly it would be informative to get access to the expected order of secondary structures in our model. This parameter is related to the (balanced) nesting depth of hairpins and has been analyzed e.g. in [9, 10] for the combinatorial and the Bernoulli model of RNA structure. Corresponding results (proving a small expected order) would help to strengthen our interpretation of present results. Last but not least it could be interesting to restrict the model to saturated foldings, i.e., to foldings where no arc can be inserted without violating the definition of secondary structures. This change could counteract the large number of unpaired positions. However, it is not quite clear how saturation could be incorporated into our model.

References

[1] Current Protocols in Bioinformatics (2006), Chapter 12, Unit 12, Copyright by John Wiley & Sons, Inc.

[2] F. Amman, S. H. Bernhart, G. Doose, I: L. Hofacker, J. Qin, P. F. Stadler, S. Will *The Trouble with Long-Range Base Pairs in RNA Folding*, Advances in Bioinformatics and Computational Biology, LNCS **8213**, 2013, pp 1-11

[3] E. L. Ince, Ordinary differential equations, *Pure and Applied Mathematics*, Dover Publications Inc. 1956.

[4] E. Hille, Ordinary differential equations in the complex domain, *Pure and Applied Mathematics*, A Wiley-Interscience series of texts, monographs & Tracts. **60** (2010), 37-48.

[5] I. Hofacker *Vienna RNA secondary structure server* Nucleic Acid Research **13** (2003), 3429-3431.

[6] P. Flajolet and R. Sedgewick, Analytic combinatorics, ISBN-13:9780521898065 Cambridge University Press, 2010.

[7] E. Goodwin, P. Okkema, T.C. Evans et al. *Translational regulation of tra-2 by its 30 untranslated region controls sexual identity in c. elegans.* Cell **75** (1993), 329-339.

[8] Fenix WD Huang and Christian M Reidys *On the combinatorics of sparsification* Algorithms for Molecular Biology **7**:28 (2012).

[9] M. E. Nebel *Combinatorial Properties of RNA Secondary Structures* Journal of Computational Biology **9** (2003), 541-573

[10] M. E. Nebel *Investigation of the Bernoulli model for RNA secondary structures* Bulletin of Mathematical Biology **66** (2004), 925-964.

[11] M. E. Nebel *Identifying Good Predictions of RNA Secondary Structure* Proceedings of the Pacific Symposium on Biocomputing 2004, 423-434

[12] M. Zuker and D. Sankoff *RNA secondary structures and their prediction* Bulletin of Mathematical Biology **46** (1984), 591-621-

[13] M. Sprinzl, K. S. Vassilenko, J. Emmerich and F. Bauer, *Compilation of tRNA sequences and sequences of tRNA genes*, (20 December, 1999) http://www.uni-bayreuth.de/departments/biochemie/trna/.

[14] Y. Wexler, C. Zilberstein and M. Ziv-Ukelson *A Study of Accessible Motifs and RNA Folding Complexity* Journal of Computational Biology **14** (2007), 856-872.

[15] Wuyts J., De Rijk P., Van de Peer Y., Winkelmans T., De Wachter R., *The European Large Subunit Ribosomal RNA database*, Nucleic Acids Res. **29** (2001), 175-177.

[16] M. Zuker *Computer prediction of RNA structure* Methods Enzymol. **180** (1989), 262-288.

[17] M. Zuker *Mfold web server for nucleic acid folding and hybridization prediction* Nucleic Acid Research **13** (2003), 3406-3415.

14. Appendix

In the sequel, we will present all the proofs left open in the main part of that paper. To this end we shall use $\mathcal{T}_{n,k}$ resp. $\mathcal{S}_{n,k}$ to denote the class of irreducible secondary structures resp. secondary structures of size n with the parameter currently under consideration (changes in different subsections) being equal to k. A single structure from that class will be denoted $\mathsf{T}_{n,k}$ resp. $\mathsf{S}_{n,k}$. Accordingly, we will *reuse* generating functions $T(z,w)$ and $S(z,w)$, which in different subsections will have different meanings; however, they will always be associated with classes of structures $\mathcal{T}_{n,k}$ and $\mathcal{S}_{n,k}$ just considered, using variable w to keep track of the parameter discussed in the respective subsection. Furthermore, we will index the generating functions by i, i.e., use T_i resp. S_i, in order to indicate that the corresponding representation of $T(z,w)$ resp. $S(z,w)$ assumes parameter choice $c=i$ and let $\alpha_{c,b}(w)$ be the dominant singularity for $S(z,w)$ and $T(z,w)$ for the different parameter choices. If furthermore indexed by variable z or w, this is used to denote the partial derivative of the generating function with respect to that variable. This way, $S_{1,w}$ for example will be used to represent the partial derivative with respect to w of the generating function associated to $\mathcal{S}_{n,k}$ assuming $c=1$.

Proof of Theorem 4.

Proof. Let $\mathcal{S}_{n,k}$ denote the class of secondary structures of size n and a total number of k unpaired positions residing in a hairpin-loop, then we have

$$\mathbb{E}(Y_n^{c,b}) \quad = \quad \sum_{k\geq 1} k \cdot \frac{\mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k})}{\mathbb{E}_{\#}^{c,b}(\mathcal{S}_n)} / \mathbb{E}(X_n^{c,b}).$$

Here we shall follow the same line of thoughts as for the average number of hairpin loops and we will omit some details if appropriate. The first step is to find a functional identity of the bivariate generating function associated to $\mathcal{S}_{n,k}$, i.e.,

$$S_c(z,w) = \sum_{n\geq 3} \sum_{\substack{k\geq 1\\ k\neq n}} \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k}) w^k z^n + \frac{1}{1-z}.$$

In connection with the corresponding irreducible structures $\mathcal{T}_{n,k}$ we find for the case $n \neq k \geq 1$,

$$\mathbb{E}_{\#}^{c,b}(\mathcal{T}_{n+2,k}) = \frac{b}{(n+1)^c} \cdot \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k}), \tag{14.1}$$

and for the case $k=n$,

$$\mathbb{E}_{\#}^{c,b}(\mathcal{T}_{n+2,n}) = \frac{b}{(n+1)^c}.$$

Let $T_c(z,w)$ be the corresponding double generating function of $\mathcal{T}_{n+2,k}$. Based on eq. (14.1), we have

$$T_c(z,w) \quad = \quad \sum_{n\geq 3} \sum_{\substack{k\geq 1\\ k\neq n}} \frac{b}{(n+1)^c} \cdot \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k}) w^k z^{n+2} + \sum_{n\geq 1} \frac{b}{(n+1)^c} w^n z^{n+2}. \tag{14.2}$$

On the other hand, eq. (3.3) also holds in connection with $S(z,w)$ and $T(z,w)$ as defined for this proof. Again, we start our analysis from the case $c=1$. Dividing by z and differentiating with respect to z on both sides of eq. (14.2), using eq. (3.3), we find

$$\begin{aligned} (\frac{T_1(z,w)}{z})_z \quad &= \quad b \sum_{n\geq 3} \sum_{\substack{k\geq 1\\ k\neq n}} \mathbb{E}_{\#}^{1,b}(\mathcal{S}_{n,k}) w^k z^n + b \sum_{n\geq 1} w^n z^n \\ &= \quad b(S_1(z,w) - \frac{1}{1-z}) + \frac{bwz}{1-wz} \end{aligned}$$

and for $S_{1,b} = S_{1,b}(z, w)$,

$$\frac{\partial S_{1,b}}{\partial z} \quad = \quad -\frac{1}{z}S_{1,b} + \left[\frac{1}{z} - \frac{bz}{1-z} + \frac{bwz^2}{1-wz}\right] S_{1,b}^2 + bzS_{1,b}^3. \tag{14.3}$$

As the proof for Theorem 1, for $z \in U_{\alpha_{1,b}(w)}$ and $|w-1| < \epsilon$, the singular expansion of $S_{1,b}(z, w)$ is of the same type as eq. (3.9) except for the constants $e_j(w)$. By again applying Theorem 2 and 3, we can conclude

$$\begin{aligned}
\frac{\mathbb{E}(Y_n^{1,1})}{\mathbb{E}(X_n^{1,1})} &= \frac{0.2289}{0.1326}\left(1 + \mathcal{O}(n^{-1})\right) = 1.7262\left(1 + \mathcal{O}(n^{-1})\right), \\
\frac{\mathbb{E}(Y_n^{1,2})}{\mathbb{E}(X_n^{1,2})} &= \frac{0.2257}{0.1476}\left(1 + \mathcal{O}(n^{-1})\right) = 1.5291\left(1 + \mathcal{O}(n^{-1})\right). \\
\frac{\sigma(Y_n^{1,1})}{\mathbb{E}(X_n^{1,1})} &= \frac{0.4774}{0.1326\sqrt{n}}\left(1 + \mathcal{O}(n^{-1})\right) = \frac{3.6003}{\sqrt{n}}\left(1 + \mathcal{O}(n^{-1})\right), \\
\frac{\sigma(Y_n^{1,2})}{\mathbb{E}(X_n^{1,2})} &= \frac{0.2257}{0.1476\sqrt{n}}\left(1 + \mathcal{O}(n^{-1})\right) = \frac{3.5570}{\sqrt{n}}\left(1 + \mathcal{O}(n^{-1})\right).
\end{aligned}$$

For the case $c = 2$, let S_2'', S_2' denote $\frac{\partial^2 S_2(z,w)}{\partial^2 z}$, and $\frac{\partial S_2(z,w)}{\partial z}$, then the functional identity for $S_2(z, w) = S_2$ is

$$S_2'' \quad = \quad \frac{2}{S_2}(S_2')^2 + \frac{S_2'}{z} + bS_2^3 - \left[\frac{b}{1-z} - b\frac{wz}{1-wz} + \frac{1}{z^2}\right] S_2^2 + \frac{S_2}{z^2}. \tag{14.4}$$

By substituting $S_2 = 1/u$, eq. (14.4) is transformed into

$$u'' = \frac{1}{z}u' + \left[-\frac{b}{u} + \frac{1}{z^2} + \frac{b}{1-z} - \frac{bwz}{1-wz} - \frac{u}{z^2}\right],$$

from which we have the singular expansion of $S_2(z, w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and 3, we obtain

$$\begin{aligned}
\frac{\mathbb{E}(Y_n^{2,1})}{\mathbb{E}(X_n^{2,1})} &= \frac{0.2150}{0.1238}(1 + \mathcal{O}(\log n)^{-1}) = 1.7367(1 + \mathcal{O}(\log n)^{-1}), \\
\frac{\mathbb{E}(Y_n^{2,2})}{\mathbb{E}(X_n^{2,2})} &= \frac{0.2303}{0.1489}(1 + \mathcal{O}(\log n)^{-1}) = 1.5467(1 + \mathcal{O}(\log n)^{-1}). \\
\frac{\sigma(Y_n^{2,1})}{\mathbb{E}(X_n^{2,1})} &= \frac{0.4093}{0.1238\sqrt{n}}\left(1 + \mathcal{O}(\log n)^{-1}\right) = \frac{3.3058}{\sqrt{n}}\left(1 + \mathcal{O}(\log n)^{-1}\right), \\
\frac{\sigma(Y_n^{2,2})}{\mathbb{E}(X_n^{2,2})} &= \frac{0.4501}{0.1489\sqrt{n}}\left(1 + \mathcal{O}(\log n)^{-1}\right) = \frac{3.0230}{\sqrt{n}}\left(1 + \mathcal{O}(\log n)^{-1}\right).
\end{aligned}$$

Henceforth Theorem 4 follows. □

Proof of Theorem 5. In contrast to the hairpin case, we shall in the sequel use the generating functions for the irreducible secondary structures directly to determine the average number of unpaired bases, number and length of bulges, number and length of interior loops, number and degree of multiloops and the number of unpaired nucleotides in the exterior loop.

Proof. The class $\mathcal{T}_{n+2}$ of irreducible structures of size $n+2$ can be decomposed according to the following cases:

(1) The outermost arc (s_1, s_{n+2}) encloses a run of unpaired positions (each contributing to our parameter), or
(2) the outermost arc (s_1, s_{n+2}) encloses a sequence of alternating unpaired regions and irreducible structures (where the former all contribute to our parameter directly and the later are considered recursively).

Note that the second case already has been used to derive eq. (3.3). Denoting by $T_1(z,w) = T_1$ the bivariate generating function for $\mathbb{E}_{\#}^{1,b}(\mathcal{T}_{n+2,k})$, $n \geq 1$, $k \geq 0$, where each unpaired position is marked by w, we find by using above decomposition

$$\begin{aligned}(\frac{T_1}{z})_z &= b\left[\sum_{i\geq 1}(\frac{1}{1-wz})^{i+1} \times T_1^i\right] + \frac{bwz}{1-wz}\\ &= \frac{b}{1-zw-T_1} - b,\end{aligned}$$

which can be simplified as

$$\frac{\partial T_1(z,w)}{z\partial z} - \frac{T_1(z,w)}{z^2} = \frac{b}{1-zw-T_1(z,w)} - b.$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_1(z,w)$ is of type $(1-\frac{z}{\alpha_{1,b}(w)})^{\frac{1}{2}}$. According to the functional composition for subcritical case, the singular expansion of $S_{1,b}(z,w)$ is of the same type as eq. (3.9) except for the constants $e_j(w)$. By applying Theorem 2 and 3, we finally obtain

$$\begin{aligned}\mathbb{E}(U_n^{1,1}) &= \frac{[z^n]\frac{\partial S_1(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_1(z,1)} = \frac{[z^n]S_{1,w}(z,1)}{[z^n]S_1(z,1)}\\ &= 0.5918n\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right),\\ \mathbb{E}(U_n^{1,2}) &= 0.5266n\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right).\\ \sigma(U_n^{1,1}) &= 0.4410\sqrt{n}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right)\\ \sigma(U_n^{1,2}) &= 0.4256\sqrt{n}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right)\end{aligned}$$

Next we consider the case $c = 2$. Using the same line of reasoning as for $c = 1$ now taking twice the partial derivative with respect to z in order to cancel the denominator $(n+1)^2$ we find

$$\begin{aligned}(z(\frac{T_2}{z})_z)_z &= b\left[\sum_{i\geq 1} T_2^i(\frac{1}{1-wz})^{i+1}\right] + \frac{bwz}{1-wz}\\ &= \frac{b}{1-zw-T_2} - b,\end{aligned}$$

which can be simplified as

$$\frac{\partial^2 T_2(z,w)}{\partial z^2} - \frac{\partial T_2(z,w)}{z\partial z} + \frac{T_2(z,w)}{z^2} = \frac{b}{1-zw-T_2(z,w)} - b.$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_2(z,w)$ is of type $(1-\frac{z}{\alpha_{2,b}(w)})(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{\frac{1}{2}}$, from which we have the singular expansion of $S_2(z,w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and

3, we can derive

$$\begin{aligned}
\mathbb{E}(U_n^{2,1}) &= \frac{[z^n]\frac{\partial S_2(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_2(z,1)} = \frac{[z^n]S_{2,w}(z,1)}{[z^n]S_2(z,1)} \\
&= 0.7046\, n\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right), \\
\mathbb{E}(U_n^{2,2}) &= 0.6323\, n\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right). \\
\sigma(U_n^{2,1}) &= 0.4605\sqrt{n}\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right), \\
\sigma(U_n^{2,2}) &= 0.4507\sqrt{n}\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right).
\end{aligned}$$

Therefore Theorem 5 follows. □

Proof of Theorem 6.

Proof. We decompose an irreducible structure from $\mathcal{T}_{n+2}$ according to the following cases:

1. The outermost arc (s_1, s_{n+2}) encloses a run of n unpaired positions (no bulge but hairpin-loop of length n),
2. the outermost arc (s_1, s_{n+2}) encloses at least two irreducible structures with (potentially empty) unpaired regions before, after and in between (no bulge but a multiloop),
3. the outermost arc (s_1, s_{n+2}) encloses a single irreducible structure IS.

For the third case no bulge (but an interior loop) results if IS has a preceding and succeeding nonempty run of unpaired positions. Otherwise, i.e. if there is either no preceding or no succeeding run of unpaired positions, a bulge results (which is marked by variable w). Finally, if both runs do not exist again no bulge is to be reported. This way, denoting by $T_1(z,w) = T_1$ the bivariate generating function for $\mathcal{T}_{n+2,k}$, $n \geq 1$, $k \geq 0$ and $c = 1$, we arrive at

$$\begin{aligned}
(\frac{T_1}{z})_z &= b\left[\sum_{i\geq 2}(\frac{1}{1-z})^{i+1}\times T_1^i\right] + \frac{bz}{1-z} + b((\frac{z}{1-z})^2 + 2w(\frac{z}{1-z}) + 1)\times T_1 \\
&= \frac{b}{1-z-T_1} - b + 2b(w-1)T_1\frac{z}{1-z},
\end{aligned}$$

which can be simplified as

$$\frac{\partial T_1(z,w)}{z\partial z} - \frac{T_1(z,w)}{z^2} = \frac{b}{1-z-T_1(z,w)} - b + 2b(w-1)\frac{zT_1(z,w)}{1-z}.$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_1(z,w)$ is of type $(1-\frac{z}{\alpha_{1,b}(w)})^{\frac{1}{2}}$. According to the functional composition for subcritical case, the singular expansion of $S_{1,b}(z,w)$ is of the same type as eq. (3.9) except for the constants $e_j(w)$. By applying Theorem 2 and 3, we finally obtain the average number of bulges with standard deviation for $c = 1$ to be given by

$$\begin{aligned}
\mathbb{E}(B_n^{1,1}) &= \frac{[z^n]\frac{\partial S_1(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_1(z,1)} = \frac{[z^n]S_{1,w}(z,1)}{[z^n]S_1(z,1)} \\
&= 0.0210n\left(1+\mathcal{O}(n^{-1})\right), \\
\mathbb{E}(B_n^{1,2}) &= 0.0261n\left(1+\mathcal{O}(n^{-1})\right). \\
\sigma(B_n^{1,1}) &= 0.1379\sqrt{n}\left(1+\mathcal{O}(n^{-1})\right), \\
\sigma(B_n^{1,2}) &= 0.1517\sqrt{n}\left(1+\mathcal{O}(n^{-1})\right).
\end{aligned}$$

We continue by considering the case $c = 2$ where we shall use $T_2 = T_2(z,w)$ to denote the generating function for irreducible secondary structures with bulges marked by variable w. Using the same decomposition discussed above we find

$$
\begin{aligned}
(z(\frac{T_2}{z})_z)_z &= b\left[\sum_{i\geq 2}(\frac{1}{1-z})^{i+1}\times T_2^i\right]+\frac{bz}{1-z}+b((\frac{z}{1-z})^2+2w(\frac{z}{1-z})+1)\times T_2\\
&= \frac{b}{1-z-T_2}-b+2b(w-1)T_2\frac{z}{1-z},
\end{aligned}
$$

which can be simplified as

$$
\frac{\partial^2 T_2(z,w)}{\partial z^2}-\frac{\partial T_2(z,w)}{z\partial z}+\frac{T_2(z,w)}{z^2}=\frac{b}{1-z-T_2(z,w)}-b+2b(w-1)\frac{zT_2(z,w)}{1-z}.
$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_2(z,w)$ is of type $(1-\frac{z}{\alpha_{2,b}(w)})(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{\frac{1}{2}}$, from which we have the singular expansion of $S_2(z,w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and 3, we can derive

$$
\begin{aligned}
\mathbb{E}(B_n^{2,1}) &= \frac{[z^n]\frac{\partial S_2(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_2(z,1)}=\frac{[z^n]S_{2,w}(z,1)}{[z^n]S_2(z,1)}\\
&= 0.0076\,n\left(1+\mathcal{O}((\log n)^{-1})\right),\\
\mathbb{E}(B_n^{2,2}) &= 0.0113\,n\left(1+\mathcal{O}((\log n)^{-1})\right).\\
\sigma(B_n^{2,1}) &= 0.0848\sqrt{n}\left(1+\mathcal{O}((\log n)^{-1})\right),\\
\sigma(B_n^{2,2}) &= 0.1024\sqrt{n}\left(1+\mathcal{O}((\log n)^{-1})\right).
\end{aligned}
$$

Consequently Theorem 6 follows. □

Proof of Theorem 7.

Proof. By using the same decomposition as for the previous proof only changing the way w is used for marking (now every unpaired position belonging to a bulge instead of just a single one) we find for $T_1(z,w)=T_1$, $n\geq 1$, $k\geq 0$ and $c=1$:

$$
\begin{aligned}
(\frac{T_1}{z})_z &= b\left[\sum_{i\geq 2}(\frac{1}{1-z})^{i+1}\times T_1^i\right]+\frac{bz}{1-z}+b((\frac{z}{1-z})^2+2\frac{wz}{1-wz}+1)\times T_1\\
&= \frac{b}{1-z-T_1}-b+2b(\frac{wz}{1-wz}-\frac{z}{1-z})\times T_1,
\end{aligned}
$$

which can be simplified as

$$
\frac{\partial T_1(z,w)}{z\partial z}-\frac{T_1(z,w)}{z^2}=\frac{b}{1-z-T_1(z,w)}-b+2b(\frac{wz}{1-wz}-\frac{z}{1-z})\times T_1(z,w).
$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_1(z,w)$ is of type $(1-\frac{z}{\alpha_{1,b}(w)})^{\frac{1}{2}}$. According to the functional composition for subcritical case, the singular expansion of $S_{1,b}(z,w)$ is of the same type as eq. (3.9) except for the constants $e_j(w)$. By applying Theorem 2 and 3, we finally obtain the average length of a single bulge-loop in the

$(1,b)$-polymer-zeta-model with standard deviation is given by

$$\begin{aligned}
\frac{\mathbb{E}(L_n^{1,1})}{\mathbb{E}(B_n^{1,1})} &= \frac{0.0430}{0.0210}\left(1+\mathcal{O}(n^{-1})\right) = 2.0476\left(1+\mathcal{O}(n^{-1})\right),\\
\frac{\mathbb{E}(L_n^{1,2})}{\mathbb{E}(B_n^{1,2})} &= \frac{0.0460}{0.0261}\left(1+\mathcal{O}(n^{-1})\right) = 1.7625\left(1+\mathcal{O}(n^{-1})\right).\\
\frac{\sigma(L_n^{1,1})}{\mathbb{E}(B_n^{1,1})} &= \frac{0.3516}{0.0210\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right) = \frac{16.7348}{\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right),\\
\frac{\sigma(L_n^{1,2})}{\mathbb{E}(B_n^{1,2})} &= \frac{0.3232}{0.0261\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right) = \frac{12.3834}{\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right).
\end{aligned}$$

Similarly for the case $c=2$, we have

$$(z(\frac{T_2}{z})_z)_z = \frac{b}{1-z-T_2} - b + 2b(\frac{wz}{1-wz} - \frac{z}{1-z}) \times T_2,$$

which can be simplified as

$$\frac{\partial^2 T_2(z,w)}{\partial z^2} - \frac{\partial T_2(z,w)}{z\partial z} + \frac{T_2(z,w)}{z^2} = \frac{b}{1-z-T_2(z,w)} - b + 2b(\frac{wz}{1-wz} - \frac{z}{1-z}) \times T_2(z,w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_2(z,w)$ is of type $(1-\frac{z}{\alpha_{2,b}(w)})(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{\frac{1}{2}}$, from which we have the singular expansion of $S_2(z,w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and 3, we can derive

$$\begin{aligned}
\frac{\mathbb{E}(L_n^{2,1})}{\mathbb{E}(B_n^{2,1})} &= \frac{0.0183}{0.0076}\left(1+\mathcal{O}((\log n)^{-1})\right) = 2.4079\left(1+\mathcal{O}((\log n)^{-1})\right),\\
\frac{\mathbb{E}(L_n^{2,2})}{\mathbb{E}(B_n^{2,2})} &= \frac{0.0230}{0.0113}\left(1+\mathcal{O}((\log n)^{-1})\right) = 2.0354\left(1+\mathcal{O}((\log n)^{-1})\right).\\
\frac{\sigma(L_n^{2,1})}{\mathbb{E}(B_n^{2,1})} &= \frac{0.2691}{0.0076\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right) = \frac{35.4115}{\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right),\\
\frac{\sigma(L_n^{2,2})}{\mathbb{E}(B_n^{2,2})} &= \frac{0.2638}{0.0113\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right) = \frac{23.3449}{\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right).
\end{aligned}$$

Consequently Theorem 7 follows. □

Proof of Theorem 8.

Proof. Again we stick to the decomposition of $\mathcal{T}_{n+2}$ used to prove Theorem 6. There we already indicated which case will contribute interior loops to a structure. Denoting by $T_1(z,w) = T_1$ the bivariate generating function for $\mathbb{E}_{\#}^{c,b}(T_{n+2,k})$, $n \geq 1$, $k \geq 0$ the number of interior loops and $c=1$, we find

$$\begin{aligned}
(\frac{T_1}{z})_z &= b\left[\sum_{i\geq 2}(\frac{1}{1-z})^{i+1} \times T_1^i\right] + \frac{bz}{1-z} + b(w(\frac{z}{1-z})^2 + 2\frac{z}{1-z} + 1) \times T_1\\
&= \frac{b}{1-z-T_1} - b + b(w-1)(\frac{z}{1-z})^2 T_1,
\end{aligned}$$

which can be simplified as

$$\frac{\partial T_1(z,w)}{z\partial z} - \frac{T_1(z,w)}{z^2} = \frac{b}{1-z-T_1(z,w)} - b + b(w-1)(\frac{z}{1-z})^2 T_1(z,w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_1(z,w)$ is of type $(1-\frac{z}{\alpha_{1,b}(w)})^{\frac{1}{2}}$. According to the functional composition for subcritical case, the singular expansion of $S_{1,b}(z,w)$ is of the same type as eq. (3.9) except for the constants $e_j(w)$.

By applying Theorem 2 and 3, we finally obtain the average number of interior loops in the $(1,b)$-polymer-zeta-model with standard deviation:

$$\begin{aligned}\mathbb{E}(I_n^{1,1}) &= 0.0110n\left(1+\mathcal{O}(n^{-1})\right),\\ \mathbb{E}(I_n^{1,2}) &= 0.0099n\left(1+\mathcal{O}(n^{-1})\right).\\ \sigma(I_n^{1,1}) &= 0.0996\sqrt{n}\left(1+\mathcal{O}(n^{-1})\right),\\ \sigma(I_n^{1,2}) &= 0.0954\sqrt{n}\left(1+\mathcal{O}(n^{-1})\right).\end{aligned}$$

Similarly for the case $c=2$, we have

$$(z(\frac{T_2}{z})_z)_z = \frac{b}{1-z-T_2} - b + b(w-1)(\frac{z}{1-z})^2 T_2,$$

which can be simplified as

$$\frac{\partial^2 T_2(z,w)}{\partial z^2} - \frac{\partial T_2(z,w)}{z\partial z} + \frac{T_2(z,w)}{z^2} = \frac{b}{1-z-T_2(z,w)} - b + b(w-1)(\frac{z}{1-z})^2 T_2(z,w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_2(z,w)$ is of type $(1-\frac{z}{\alpha_{2,b}(w)})(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{\frac{1}{2}}$, from which we have the singular expansion of $S_2(z,w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and 3, we find

$$\begin{aligned}\mathbb{E}(I_n^{2,1}) &= 0.0055n\left(1+\mathcal{O}((\log n)^{-1})\right),\\ \mathbb{E}(I_n^{2,2}) &= 0.0059n\left(1+\mathcal{O}((\log n)^{-1})\right).\\ \sigma(I_n^{2,1}) &= 0.0706\sqrt{n}\left(1+\mathcal{O}((\log n)^{-1})\right),\\ \sigma(I_n^{2,2}) &= 0.0739\sqrt{n}\left(1+\mathcal{O}((\log n)^{-1})\right).\end{aligned}$$

Consequently Theorem 8 follows. □

Proof of Theorem 9.

Proof. It is straightforward to adapt the functional equation for the number of interior loops to keep track of their total size (by means of variable w). Denoting by $T_1(z,w)=T_1$ the corresponding bivariate generating function, assuming $n\geq 1$ and $c=1$, we find

$$\begin{aligned}(\frac{T_1}{z})_z &= b\left[\sum_{i\geq 2}(\frac{1}{1-z})^{i+1}\times T_1^i\right] + \frac{bz}{1-z} + b((\frac{wz}{1-wz})^2 + 2\frac{z}{1-z} + 1)\times T_1\\ &= \frac{b}{1-z-T_1} - b + bT_1\left[(\frac{wz}{1-wz})^2 - (\frac{z}{1-z})^2\right],\end{aligned}$$

which can be simplified as

$$\frac{\partial T_1(z,w)}{z\partial z} - \frac{T_1(z,w)}{z^2} = \frac{b}{1-z-T_1(z,w)} - b + bT_1(z,w)\left[(\frac{wz}{1-wz})^2 - (\frac{z}{1-z})^2\right].$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_1(z,w)$ is of type $(1-\frac{z}{\alpha_{1,b}(w)})^{\frac{1}{2}}$. According to the functional composition for subcritical case, the singular expansion of $S_{1,b}(z,w)$ is of the same type as eq. (3.9) except for the constants $e_j(w)$. By applying Theorem 2 and 3, we finally obtain the average size of an interior loop in the

$(1, b)$-polymer-zeta-model with standard deviation:

$$\begin{aligned}
\frac{\mathbb{E}(P_n^{1,1})}{\mathbb{E}(I_n^{1,1})} &= \frac{0.0466}{0.0110}\left(1+\mathcal{O}(n^{-1})\right) = 4.2364\left(1+\mathcal{O}(n^{-1})\right),\\
\frac{\mathbb{E}(P_n^{1,2})}{\mathbb{E}(I_n^{1,2})} &= \frac{0.0358}{0.0099}\left(1+\mathcal{O}(n^{-1})\right) = 3.6162\left(1+\mathcal{O}(n^{-1})\right).\\
\frac{\sigma(P_n^{1,1})}{\mathbb{E}(I_n^{1,1})} &= \frac{0.4785}{0.0110\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right) = \frac{43.5036}{\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right),\\
\frac{\sigma(P_n^{1,2})}{\mathbb{E}(I_n^{1,2})} &= \frac{0.3834}{0.0099\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right) = \frac{38.7243}{\sqrt{n}}\left(1+\mathcal{O}(n^{-1})\right).
\end{aligned}$$

Similarly for the case $c = 2$, we have

$$(z(\frac{T_2}{z})_z)_z = \frac{b}{1-z-T_2} - b + bT_2\left[(\frac{wz}{1-wz})^2 - (\frac{z}{1-z})^2\right],$$

which can be simplified as

$$\frac{\partial^2 T_2(z,w)}{\partial z^2} - \frac{\partial T_2(z,w)}{z\partial z} + \frac{T_2(z,w)}{z^2} = \frac{b}{1-z-T_2(z,w)} - b + b(w-1)(\frac{z}{1-z})^2 T_2(z,w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_2(z,w)$ is of type $(1-\frac{z}{\alpha_{2,b}(w)})(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{\frac{1}{2}}$, from which we have the singular expansion of $S_2(z,w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and 3, we find

$$\begin{aligned}
\frac{\mathbb{E}(P_n^{2,1})}{\mathbb{E}(I_n^{2,1})} &= \frac{0.0294}{0.0055}\left(1+\mathcal{O}((\log n)^{-1})\right) = 5.3455\left(1+\mathcal{O}((\log n)^{-1})\right),\\
\frac{\mathbb{E}(P_n^{2,2})}{\mathbb{E}(I_n^{2,2})} &= \frac{0.0260}{0.0059}\left(1+\mathcal{O}((\log n)^{-1})\right) = 4.4068\left(1+\mathcal{O}((\log n)^{-1})\right).\\
\frac{\sigma(P_n^{2,1})}{\mathbb{E}(I_n^{2,1})} &= \frac{0.4457}{0.0055\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right) = \frac{81.0401}{\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right),\\
\frac{\sigma(P_n^{2,2})}{\mathbb{E}(I_n^{2,2})} &= \frac{0.3724}{0.0059\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right) = \frac{63.1238}{\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-1})\right).
\end{aligned}$$

Consequently Theorem 9 follows. □

Proof of Theorem 10.

Proof. The decomposition of subsection 6 already addressed multiloops; case (2) is the only way a multiloop may result. It is an easy exercise to derive a functional equation for the bivariate generating function $T_1(z,w) = T_1$ according to that decomposition of $\mathcal{T}_{n+2}$, marking every occurrence of a multiloop by variable w. Assuming $n \geq 1$ and $c = 1$, we find

$$\begin{aligned}
(\frac{T_1}{z})_z &= bw\left[\sum_{i\geq 2}(\frac{1}{1-z})^{i+1}\times T_1^i\right] + \frac{bz}{1-z} + b(\frac{1}{1-z})^2\times T_1\\
&= \frac{bw}{1-z-T_1} + b\frac{z-w}{1-z} + b(1-w)(\frac{1}{1-z})^2 T_1,
\end{aligned}$$

which can be simplified as

$$\frac{\partial T_1(z,w)}{z\partial z} - \frac{T_1(z,w)}{z^2} = \frac{bw}{1-z-T_1(z,w)} - b + b\frac{z-w}{1-z} + b(1-w)(\frac{1}{1-z})^2 T_1(z,w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_1(z,w)$ is of type $(1-\frac{z}{\alpha_{1,b}(w)})^{\frac{1}{2}}$. According to the functional composition for subcritical case, the singular expansion of $S_{1,b}(z,w)$ is of the same type as eq. (3.9) except for the constants

$e_j(w)$. By applying Theorem 2 and 3, we finally obtain the average number of multiloops in the $(1, b)$-polymer-zeta-model with standard deviation:

$$\begin{aligned}
\mathbb{E}(M_n^{1,1}) &= 0.0330n - 0.4683 + \mathcal{O}(n^{-1}), \\
\mathbb{E}(M_n^{1,2}) &= 0.0390n - 0.4618\left(1 + \mathcal{O}(n^{-1})\right). \\
\sigma(M_n^{1,1}) &= 0.1188\sqrt{n}\left(1 + \mathcal{O}(n^{-1})\right), \\
\sigma(M_n^{1,2}) &= 0.1277\sqrt{n}\left(1 + \mathcal{O}(n^{-1})\right).
\end{aligned}$$

Similarly, we have for the case $c = 2$

$$(z(\frac{T_2}{z})_z)_z = \frac{bw}{1 - z - T_2} + b\frac{z - w}{1 - z} + b(1 - w)(\frac{1}{1 - z})^2 T_2,$$

which can be simplified as

$$\frac{\partial^2 T_2(z, w)}{\partial z^2} - \frac{\partial T_2(z, w)}{z\partial z} + \frac{T_2(z, w)}{z^2} = \frac{bw}{1 - z - T_2(z, w)} + b\frac{z - w}{1 - z} + b(1 - w)(\frac{1}{1 - z})^2 T_2(z, w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_2(z, w)$ is of type $(1 - \frac{z}{\alpha_{2,b}(w)})(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{\frac{1}{2}}$, from which we have the singular expansion of $S_2(z, w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and 3, we find

$$\begin{aligned}
\mathbb{E}(M_n^{2,1}) &= 0.0112n + 0.0024\frac{n}{\log n} + \mathcal{O}(\frac{n}{\log^2 n}), \\
\mathbb{E}(M_n^{2,2}) &= 0.0220n - 0.0521\frac{n}{\log n} + \mathcal{O}(\frac{n}{\log^2 n}). \\
\sigma(M_n^{2,1}) &= 0.0730\sqrt{n}\left(1 + \mathcal{O}((\log n)^{-1})\right), \\
\sigma(M_n^{2,2}) &= 0.0872\sqrt{n}\left(1 + \mathcal{O}((\log n)^{-1})\right).
\end{aligned}$$

Consequently Theorem 10 follows. □

Proof of Theorem 11.

Proof. To analyze the average degree of a multiloop we only need to slightly change the functional equation of the last proof; if a multiloop is made out of i irreducible structures enclosed by an arc (which for $c = 1$ is considered within the functional equation by a term $(\frac{1}{1-z})^{i+1} \times T_1^i$) we have to mark it by w^{i+1} (instead of w). Accordingly, we find

$$\begin{aligned}
(\frac{T_1}{z})_z &= b\left[\sum_{i\geq 2}(\frac{w}{1 - z})^{i+1} \times T_1^i\right] + \frac{bz}{1 - z} + b(\frac{1}{1 - z})^2 \times T_1 \\
&= \frac{bw}{1 - z - wT_1} + b\frac{z - w}{1 - z} + b(1 - w^2)(\frac{1}{1 - z})^2 T_1,
\end{aligned}$$

which can be simplified as

$$\frac{\partial T_1(z, w)}{z\partial z} - \frac{T_1(z, w)}{z^2} = \frac{bw}{1 - z - wT_1(z, w)} + b\frac{z - w}{1 - z} + b(1 - w^2)(\frac{1}{1 - z})^2 T_1(z, w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_1(z, w)$ is of type $(1 - \frac{z}{\alpha_{1,b}(w)})^{\frac{1}{2}}$. According to the functional composition for subcritical case, the singular expansion of $S_{1,b}(z, w)$ is of the same type as eq. (3.9) except for the constants $e_j(w)$. By applying Theorem 2 and 3, we finally obtain the average degree of a multiloop in the

$(1,b)$-polymer-zeta-model with standard deviation

$$\begin{aligned}
\frac{\mathbb{E}(D_n^{1,1})}{\mathbb{E}(M_n^{1,1})} &= \frac{0.1975n-0.8429\sqrt{n}+O(1)}{0.0330n-0.4683+O(n^{-1})} = 5.9848\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right),\\
\frac{\mathbb{E}(D_n^{1,2})}{\mathbb{E}(M_n^{1,2})} &= \frac{0.2169n-0.7265\sqrt{n}+O(1)}{0.0390n-0.4618+O(n^{-1})} = 5.5615\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right).\\
\frac{\sigma(D_n^{1,1})}{\mathbb{E}(M_n^{1,1})} &= \frac{0.4629}{0.0330\sqrt{n}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right) = \frac{14.0273}{\sqrt{n}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right),\\
\frac{\sigma(D_n^{1,2})}{\mathbb{E}(M_n^{1,2})} &= \frac{0.4917}{0.0390\sqrt{n}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right) = \frac{12.6077}{\sqrt{n}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right).
\end{aligned}$$

Similarly for the case $c=2$, we have

$$(z(\frac{T_2}{z})_z)_z = \frac{bw}{1-z-wT_2} + b\frac{z-w}{1-z} + b(1-w^2)(\frac{1}{1-z})^2 T_2,$$

which can be simplified as

$$\frac{\partial^2 T_2(z,w)}{\partial z^2} - \frac{\partial T_2(z,w)}{z\partial z} + \frac{T_2(z,w)}{z^2} = \frac{bw}{1-z-wT_2(z,w)} + b\frac{z-w}{1-z} + b(1-w^2)(\frac{1}{1-z})^2 T_2(z,w).$$

By the same analysis as for Theorem 1, we have the dominant term in the singular expansion of $T_2(z,w)$ is of type $(1-\frac{z}{\alpha_{2,b}(w)})(\frac{\alpha_{2,b}(w)}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}(w)}}))^{\frac{1}{2}}$, from which we have the singular expansion of $S_2(z,w)$ that is the same as eq. (3.15) except the constant $a_0(w)$. By applying Theorem 2 and 3, we find

$$\begin{aligned}
\frac{\mathbb{E}(D_n^{2,1})}{\mathbb{E}(M_n^{2,1})} &= \frac{0.1406n-0.2396\frac{n}{\log n}+O(\frac{n}{\log^2 n})}{0.0112n+0.0024\frac{n}{\log n}+\mathcal{O}(\frac{n}{\log^2 n})} = 12.5536\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right),\\
\frac{\mathbb{E}(D_n^{2,2})}{\mathbb{E}(M_n^{2,2})} &= \frac{0.1840n-0.3006\frac{n}{\log n}+O(\frac{n}{\log^2 n})}{0.0220n-0.0521\frac{n}{\log n}+\mathcal{O}(\frac{n}{\log^2 n})} = 8.3636\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right).\\
\frac{\sigma(D_n^{2,1})}{\mathbb{E}(M_n^{2,1})} &= \frac{0.2908}{0.0112\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right) = \frac{25.9643}{\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right),\\
\frac{\sigma(D_n^{2,2})}{\mathbb{E}(M_n^{2,2})} &= \frac{0.3590}{0.0220\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right) = \frac{16.3182}{\sqrt{n}}\left(1+\mathcal{O}((\log n)^{-\frac{1}{2}})\right).
\end{aligned}$$

Consequently Theorem 11 follows. □

Proof of Theorem 12.

Proof. For $c=1,2$ let $S_c(z,w)$ be the generating function of secondary structures with every unpaired bases in the exterior loop labeled by variable w. Then $S_c(z,w) = \frac{1}{1-T_c(z,1)-wz}$ where $T_c(z,1)$ represents the generating function of irreducible structures. By differentiating $S_c(z,w)$ with respect to w and setting $w=1$ afterwards, we find $S_{c,w}(z,1) = zS_c(z,1)^2$ for $c=1,2$. Recalling the singular expansion of $S_1(z)=S_1(z,1)$ and $S_2(z)=S_2(z,1)$ we arrive at

$$\begin{aligned}
S_1(z) &= \left(1-\frac{z}{\alpha_{1,b}}\right)^{-\frac{1}{2}}\sum_{j=0}^{\infty} e_j\left(1-\frac{z}{\alpha_{1,b}}\right)^{\frac{j}{2}} \quad \text{and } e_0 = \frac{1}{\alpha_{1,b}}\sqrt{\frac{1}{2b}} \neq 0.\\
S_2(z) &= \left(1-\frac{z}{\alpha_{2,b}}\right)^{-1}\left[\frac{\alpha_{2,b}}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}}})\right]^{-\frac{1}{2}}\left[\frac{1}{a_0}+\mathcal{O}\left[\frac{\alpha_{2,b}}{z}\log(\frac{1}{1-\frac{z}{\alpha_{2,b}}})\right]^{-1}\right].
\end{aligned}$$

Now we can derive r-th moment of $E_n^{c,b}$, i.e.

$$\begin{aligned}
\mathbb{E}(E_n^{1,b}(E_n^{1,b}-1)\cdots(E_n^{1,b}-r+1)) &= \frac{[z^n]\partial_w^r S_1(z,w)|_{w=1}}{[z^n]S_1(z,1)}\\
&= r!(2b)^{-\frac{r}{2}}\frac{(n-r)^{\frac{r+1}{2}-1}}{n^{-\frac{1}{2}}}\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{r+1}{2})}(1+O(\frac{1}{n}))\\
\mathbb{E}(E_n^{2,b}(E_n^{2,b}-1)\cdots(E_n^{2,b}-r+1)) &= \frac{[z^n]\partial_w^r S_2(z,w)|_{w=1}}{[z^n]S_2(z,1)}\\
&= r!(2b)^{-\frac{r}{2}}\frac{(n-r)^r}{\Gamma(r+1)}\frac{(\log(n-r))^{-\frac{r+1}{2}}}{(\log n)^{-\frac{1}{2}}}(1+O(\frac{1}{\log n})).
\end{aligned}$$

Accordingly, we finally obtain for $c=1$ the average number of unpaired bases found in the exterior loop to be asymptotically given by

$$\begin{aligned}
\mathbb{E}(E_n^{1,1}) &= \frac{[z^n]\frac{\partial S_1(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_1(z,1)} = \frac{[z^n]S_{1,w}(z,1)}{[z^n]S_1(z,1)}\\
&= 1.2066\, n^{\frac{1}{2}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right),\\
\mathbb{E}(E_n^{1,2}) &= 0.8668\, n^{\frac{1}{2}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right).
\end{aligned}$$

We can further calculate the standard deviation $\sigma(E_n^{1,b}) = r'_{1,b}\, n^{\frac{1}{2}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right)$ where $(r'_{1,1}, r'_{1,2}) \approx (0.6842, 0.5254)$. For $c=2$ we get

$$\begin{aligned}
\mathbb{E}(E_n^{2,1}) &= \frac{[z^n]\frac{\partial S_2(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_2(z,1)} = \frac{[z^n]S_{2,w}(z,1)}{[z^n]S_2(z,1)}\\
&= 0.7978\, n(\log n)^{-\frac{1}{2}}\left(1+\mathcal{O}((\log n)^{-1})\right),\\
\mathbb{E}(E_n^{2,2}) &= 0.5974\, n(\log n)^{-\frac{1}{2}}\left(1+\mathcal{O}((\log n)^{-1})\right).
\end{aligned}$$

The standard deviation is $\sigma(E_n^{2,b}) = r'_{2,b}n(\log n)^{-\frac{1}{2}}\left(1+\mathcal{O}((\log n)^{-1})\right)$ where $(r'_{1,1}, r'_{1,2}) \approx (0.2411, 0.2000)$. □

Proof of Theorem 13.

Proof. For $c=1,2$ let $S_c(z,w)$ be the generating function of secondary structures with every irreducible structure in the exterior loop labeled by variable w. Then $S_c(z,w) = \frac{1}{1-wT_c(z,1)-z}$ where $T_c(z,1)$ represents the generating function of irreducible structures. We can derive the r-th moment of $P_n^{c,b}$, i.e.,

$$\begin{aligned}
\mathbb{E}(P_n^{1,b}(P_n^{1,b}-1)\cdots(P_n^{1,b}-r+1)) &= \frac{[z^n]\partial_w^r S_1(z,w)|_{w=1}}{[z^n]S_1(z,1)}\\
&= r!(2b)^{-\frac{r}{2}}(\frac{1}{\alpha_{1,b}}-1)^r n^{\frac{r}{2}}\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{r+1}{2})}(1+O(\frac{1}{n}))\\
\mathbb{E}(P_n^{2,b}(P_n^{2,b}-1)\cdots(P_n^{2,b}-r+1)) &= \frac{[z^n]\partial_w^r S_2(z,w)|_{w=1}}{[z^n]S_2(z,1)}\\
&= r!(2b)^{-\frac{r}{2}}(\frac{1}{\alpha_{1,b}}-1)^r\frac{n^r}{\Gamma(r+1)}(\log n)^{-\frac{r}{2}}.
\end{aligned}$$

Accordingly, we finally obtain for $c = 1$ the average degree of exterior loop to be asymptotically given by

$$\begin{aligned}\mathbb{E}(P_n^{1,1}) &= \frac{[z^n]\frac{\partial S_1(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_1(z,1)} = \frac{[z^n]S_{1,w}(z,1)}{[z^n]S_1(z,1)}\\ &= 0.8426\, n^{\frac{1}{2}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right),\\ \mathbb{E}(P_n^{1,2}) &= 0.6332\, n^{\frac{1}{2}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right).\end{aligned}$$

We can further calculate the standard deviation $\sigma(P_n^{1,b}) = p'_{1,b}\, n^{\frac{1}{2}}\left(1+\mathcal{O}(n^{-\frac{1}{2}})\right)$ where $(p'_{1,1}, p'_{1,2}) \approx (0.4404, 0.3310)$. For $c = 2$ we get

$$\begin{aligned}\mathbb{E}(P_n^{2,1}) &= \frac{[z^n]\frac{\partial S_2(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_2(z,1)} = \frac{[z^n]S_{2,w}(z,1)}{[z^n]S_2(z,1)}\\ &= 0.2171\, n(\log n)^{-\frac{1}{2}}\left(1+\mathcal{O}((\log n)^{-1})\right),\\ \mathbb{E}(P_n^{2,2}) &= 0.2345\, n(\log n)^{-\frac{1}{2}}\left(1+\mathcal{O}((\log n)^{-1})\right).\end{aligned}$$

The standard deviation is $\sigma(P_n^{2,b}) = p'_{2,b} n^{\frac{1}{2}}(\log n)^{-\frac{1}{4}}\left(1+\mathcal{O}((\log n)^{-1})\right)$ where $(p'_{2,1}, p'_{2,2}) \approx (0.4659, 0.4842)$.

□