

Combinatorics and topological weights of chromatin loop networks

Andrea Bonato,¹ Michael Chiang ¹ Dom Corbett ² Sergey Kitaev,³ Davide Marenduzzo ²
Alexander Morozov,² and Enzo Orlandini ⁴

¹*Department of Physics, University of Strathclyde, Glasgow G4 0NG, Scotland, United Kingdom*

²*SUPA, School of Physics and Astronomy, The University of Edinburgh, Edinburgh EH9 3FD, Scotland, United Kingdom*

³*Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, Scotland, United Kingdom*

⁴*Department of Physics and Astronomy, University of Padova and INFN, Sezione Padova, Via Marzolo 8, I-35131 Padova, Italy*



(Received 29 August 2023; accepted 15 May 2024; published 14 June 2024)

Polymer physics models suggest that chromatin spontaneously folds into loop networks with transcription units (TUs), such as enhancers and promoters, as anchors. Here we use combinatoric arguments to enumerate the emergent chromatin loop networks, both in the case where TUs are labeled and where they are unlabeled. We then combine these mathematical results with those of computer simulations aimed at finding the inter-TU energy required to form a target loop network. We show that different topologies are vastly different in terms of both their combinatorial weight and energy of formation. We explain the latter result qualitatively by computing the topological weight of a given network—i.e., its partition function in statistical mechanics language—in the approximation where excluded volume interactions are neglected. Our results show that networks featuring local loops are statistically more likely with respect to networks including more nonlocal contacts. We suggest our classification of loop networks, together with our estimate of the combinatorial and topological weight of each network, will be relevant to catalog three-dimensional structures of chromatin fibers around eukaryotic genes, and to estimate their relative frequency in both simulations and experiments.

DOI: [10.1103/PhysRevE.109.064405](https://doi.org/10.1103/PhysRevE.109.064405)

I. INTRODUCTION

Chromatin is a protein-DNA composite polymer that provides the building block of chromosomes, and it constitutes the form in which genomic information is stored in the nuclei of eukaryotic cells. Chromatin also provides the genomic substrate for fundamental intracellular processing of DNA, such as transcription and replication [1,2]. Longstanding observations suggest that the three-dimensional (3D) structure of chromatin is functionally important: for instance, it is known that the 3D structure of a gene locus correlates with its transcriptional activity [3].

Polymer models to determine the chromatin structure in 3D are therefore important in this field, and several coarse-grained potentials have been developed to describe them (see, e.g., [4–8], and [9,10] for a review of some of these). Typically, coarse-grained polymer models view chromatin as a copolymer, or heterogeneous polymer, where different beads may have different properties to reflect, among others, the local sequence and post-translational modification in DNA-binding histone proteins, such as acetylation or methylation (see, e.g., [3,6,11]).

A simple copolymer model for active chromatin [12], which is relevant to our current work, views the fiber as a

semiflexible polymer with interspersed “transcription units” (TUs; the red circles in Fig. 1), representing open chromatin regions such as enhancers or promoters which have high affinity for multivalent chromatin-binding proteins associated with transcription—such as RNA polymerases and transcription factors, or protein complexes including both of these [13,14].

Simulations of more sophisticated polymer models, resolving chromatin-binding proteins, show that TUs come together due to the bridging-induced attraction, a positive-feedback loop associated with multivalent chromatin-protein binding [13]. The bridging-induced attraction leads to microphase separation into clusters of TUs (and their associated proteins) because clustering the TUs creates loops whose entropy grows superlinearly with TU number, eventually balancing the energetic gain of clustering [12,13]. This phenomenon provides a mechanistic model for the formation of transcription factories in mammalian nuclei [15]. This discussion suggests that in a simpler effective model, one can consider the TUs themselves as sticky for each other, and this is the model sketched in Fig. 1.

In the copolymer model of Fig. 1, chromatin loop networks emerge in a steady state due to the sticky nature of TUs. Some natural questions then arise, namely, how to classify the emerging network topologies [such as the one in Fig. 1(b)] and what the statistical likeliness of observing each of such topologies is. A possible way to classify the loop topologies is by computing the entropic exponent associated with the network, as in [16]. However, the issue arises that all networks with the same number of nodes and edges (or legs) emanating from each node would have the same entropic exponent, as

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

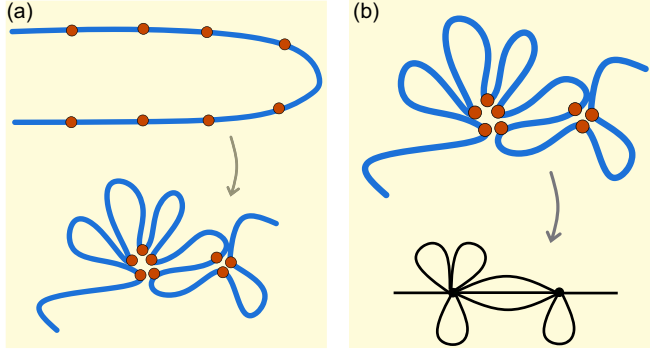


FIG. 1. (a) Top: A chromatin fiber with $n = 8$ TUs. Bottom: A possible structure formed when TUs attract each other, for instance, effectively due to the bridging-induced attraction [13]. The structure is made of two clusters and four local loops. (b) The loop network topology corresponding to this configuration (repeated at the top for clarity) is shown at the bottom of the panel.

they have the same number of nodes and edges [16]. As shown in the companion paper [17], simulations suggest, instead, that the probability of observing different networks is not constant at all, so it would be desirable to go beyond the calculation of the entropic exponent and estimate the statistical weight associated with each loop topology.

We consider two possible classes of chromatin loop networks. First, “labeled” networks are those in which the TUs are numbered. This is often relevant in biological examples where different TUs correspond to different regulatory elements, and it may be important in practice to distinguish networks with the same topology and distribution of clusters, but where different TUs participate in the clusters.

Second, “unlabeled” networks are those where TUs are not numbered, such that different configurations are topologically nonequivalent configurations of our chromatin fiber. For example, the two networks in Fig. 2(a) are different labeled networks, but represent the same topology when counting unlabeled networks. Unlabeled networks are relevant when considering generic topologies, for example, the rosette and watermelon ones in Fig. 2(b), and asking which topology is most often found in gene loci genome-wide. Labeled networks are a lot simpler to count combinatorially with respect to unlabeled ones; this is because it is hard, in general, to count the multiplicity of labeled networks corresponding to a unique, unlabeled network topology.

In the present work, we aim to classify topologies of chromatin loop networks, counting them and finding their statistical weights, which measure the probabilities of observing them in a polymer model. Our article is structured as follows. First, in Sec. II, we provide combinatorial formulas to count labeled networks. As we shall see, the theory of Bell numbers and partitions provides a powerful way to count such networks. We also find a series of recursion relations which constrain the number of labeled networks with specific properties (e.g., without or with singletons). These recursions are associated with an exponential network-generating function for which we find explicit formulas. Second, in Sec. III, we discuss the case of unlabeled, topologically inequivalent networks, and derive a formula to count the number of such

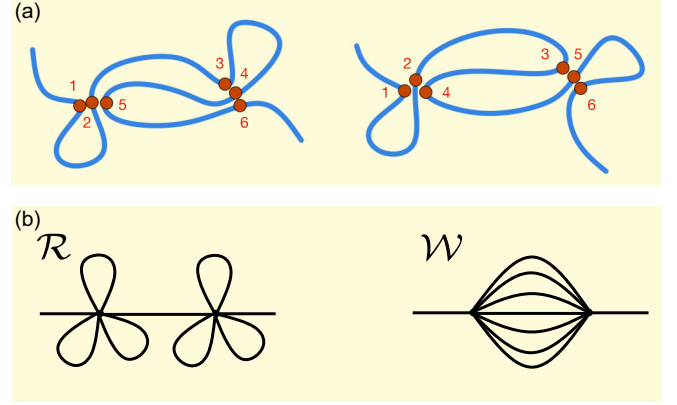


FIG. 2. (a) An example of two different labeled networks with two clusters which yield the same unlabeled topology (neglecting singletons). (b) Rosette $\mathcal{R}$ and watermelon $\mathcal{W}$ topologies.

structures with two clusters, which is of interest in applications to chromatin structures in real gene loci. Section IV contains numerical results obtained by simulating chromatin folding within a specific polymer model, viewing the chromatin fiber as a semiflexible self-avoiding chain with equally spaced sticky sites (the TUs). Here we show that different target topologies require different interaction energies between the TUs to form so that they are, in general, associated with a different entropic cost of formation. These results complement those discussed in the companion paper [17], which show that rosettelike topologies, that are rich in local loops, are much more favored statistically with respect to others with nonlocal loops. In Sec. V, we compute the statistical weight of a generic topology in the simplified case of a phantom freely jointed chain (i.e., without excluded volume interactions). We show that the weights that we compute, although approximate due to the neglect of excluded volume effects, are sufficient to recapitulate the much enhanced statistical likelihood of forming rosettelike networks, in spite of the fact that the combinatoric multiplicities of other topologies are often larger. Finally, Sec. VI contains our conclusion.

II. COMBINATORICS OF LABELED CHROMATIN LOOP NETWORKS

We first consider the case of labeled chromatin loop networks, where more progress can be done analytically. Therefore, in this section, TUs are assumed to be labeled from 1 to n , and we think of TUs as the set $\{1, 2, \dots, n\}$ that is also denoted by $[n]$.

For such labeled networks, we first derive a few enumerative results; we then discuss recursion relations and derive their generating function. Whenever suitable, the asymptotics will be discussed and we will also give references to the Online Encyclopedia of Integer Sequences [18], when the counting sequence in question can be found there.

The combinatoric multiplicities which we will find can be used, for instance, to find all possible configurations of a chromatin segment with a given number of TUs and a list of desired features (such as the number of clusters and of singletons). This provides a useful bound for all possible topologies

that this genomic region can form, in either simulations or experiments.

A. Configurations with an arbitrary number of clusters

To begin with, we note that if we do not care about the number of clusters in the configurations, then the number of different configurations with n TUs is given by the *Bell number* B_n . This is because each configuration can be thought of as a partition of the set $[n]$, where each subset (or block, or part) with at least two TUs will form a cluster, while the singletons will correspond to the TUs not belonging to any clusters. For example, for $n = 6$, the partition $\{\{1, 4\}, \{2\}, \{3, 5, 6\}\}$ encodes a possible configuration with two clusters. It is well known that B_n counts the number of partitions of $[n]$, and this is the sequence A000110 in [18] that begins with

$$1, 2, 5, 15, 52, 203, 877, 4140, \dots \quad (1)$$

The Bell numbers satisfy the recurrence relation $B_{n+1} = \sum_{k=0}^n \binom{n}{k} B_k$. Their *exponential generating function* $\sum_{n \geq 0} B_n \frac{t^n}{n!}$ is $e^{e^t - 1}$, while the ordinary generating function is

$$B(t) = \sum_{k \geq 0} \frac{t^k}{\prod_{j=1}^k (1 - jt)}. \quad (2)$$

Also, the Bell numbers satisfy Dobinski's formula $B_n = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}$ and asymptotically ($n \rightarrow \infty$),

$$B_n \sim \frac{1}{\sqrt{n}} \left(\frac{n}{W(n)} \right)^{n+\frac{1}{2}} \exp \left(\frac{n}{W(n)} - n - 1 \right), \quad (3)$$

where the *Lambert W function* has the same growth as the logarithm [19].

Interestingly, if B_n^* denotes the number of configurations *without singletons* (i.e., each TU is part of a cluster), then the following (well-known) combinatorial argument can be used to show that $B_{n+1}^* = B_n - B_n^*$. Note that $B_n - B_n^*$ is the number of partitions of $[n]$ that have at least one singleton. Now, take all singletons in a partition counted by $B_n - B_n^*$ and add them together along with the element $n+1$ to form a subset in a partition of $[n+1]$ that has no singletons (and hence is counted by B_{n+1}^*). This mapping, between partitions of $[n]$ with singletons and partitions of $[n+1]$ without singletons, is a bijection.

B. Configurations with a fixed number of clusters

We now discuss how to enumerate configurations with a fixed number of clusters. To do so, a useful set of quantities is provided by the *Stirling numbers of the second kind*, $S(n, k)$, which count the number of ways to partition the set $[n]$ into k subsets. Even though $S(n, k)$ does not directly give us the number of configurations with k clusters, below we will make use of these numbers.

We wish to find the number of partitions of $[n]$ into subsets (i.e., the number of configurations) so that precisely k subsets, $1 \leq k \leq n-2$, have two or more elements (i.e., there are exactly k clusters). We call this number $f(n, k)$. We highlight that this quantity counts the partition of n TUs into k clusters, with an arbitrary number of singletons.

1. Number of configurations with one cluster

There are $2^n - n - 1$ configurations corresponding to the case of $k = 1$. Indeed, each binary string $s_1 s_2 \dots s_n$ over the alphabet $\{0, 1\}$ corresponds to a configuration, where $s_i = 0$ indicates that the TU i is a singleton, while $s_i = 1$ indicates that the TU i is included in the only cluster. The number of possibilities is 2^n , but we need to subtract the situations when, at most, one 1 is present in the string because a cluster needs to have at least two TUs. Note that asymptotically, we have $O(2^n)$ such configurations. For $n \geq 1$, the counting sequence begins

$$0, 1, 4, 11, 26, 57, 120, 247, 502, \dots \quad (4)$$

and this is the sequence A000295 in [18].

2. Number of configurations with two clusters

The case of $k = 2$ can be derived similarly to the case of $k = 1$. Instead of binary sequences, we can consider sequences over $\{0, 1, 2\}$ (there are 3^n of them) and then subtract those sequences that do not correspond to configurations with precisely two clusters (for example, sequences with no 2's, or with one 1 and one 2). However, this method is still cumbersome, so we use the following approach instead: Let i correspond to the number of singletons in a configuration (this number cannot be bigger than $n-4$ for us to be able to create two clusters); then, $\binom{n}{i}$ is the number of ways to choose these singletons in $[n]$, and $S(n-i, 2)$ (equal to $2^{n-i-1} - 1$ [20]) counts the number of configurations with two subsets for $(n-i)$ TUs. Subtracting, from this number $(n-i)$, the number of possibilities for subsets receiving a single TU and summing up all possible numbers of singletons, we find

$$\begin{aligned} f(n, 2) &= \sum_{i=0}^{n-4} \binom{n}{i} [S(n-i, 2) - (n-i)] \\ &= \sum_{i=0}^{n-4} \binom{n}{i} (2^{n-i-1} - 1 - n + i) \\ &= \frac{1}{2} (3^n + 1) - (n+2)2^{n-1} + \binom{n}{2} + n. \end{aligned} \quad (5)$$

The last equality can be checked, for instance, by induction. We note that asymptotically, the number of configurations is $O(3^n)$ and the counting sequence begins, for $n \geq 4$, with

$$3, 25, 130, 546, 2037, 7071, \dots; \quad (6)$$

this is the sequence A112495 in [18].

C. Recurrence relations and generating function for loop networks with an arbitrary number of singletons

Using the approaches above is rather cumbersome to produce explicit formulas for arbitrary k . Alternatively, we can produce a recurrence relation for the numbers in question, $f(n, k)$, that can be turned into a partial differential equation for the respective generating function.

Note that for $n \geq 2$, this recursion reads as follows:

$$f(n, k) = (k+1)f(n-1, k) + (n-1)f(n-2, k-1). \quad (7)$$

To prove Eq. (7), we can think of producing, in a unique way, a configuration with n TUs from a smaller configuration by introducing the n th TU. The possible disjoint options are as follows:

(i) n joins an existing cluster (with at least two TUs in it) or n becomes a singleton, and there are $(k+1)f(n-1, k)$ possibilities in this case;

(ii) n ends up in a cluster with exactly two TUs, and there are $(n-1)f(n-2, k-1)$ ways as there are $n-1$ ways to select a TU to share the cluster with n .

The initial conditions of Eq. (7) are $f(0, 0) = 1$ and $f(0, k) = 0$ for $k \neq 0$, and $f(1, 0) = 1$ and $f(1, k) = 0$ for $k \neq 0$, along with $f(n, k) = 0$ for $k < 0$.

We now consider the exponential generating function, defined as

$$F(t, x) = \sum_{n, k \geq 0} f(n, k) \frac{t^n}{n!} x^k. \quad (8)$$

We can write that

$$\begin{aligned} \frac{\partial}{\partial t} F(t, x) &= \sum_{n \geq 1, k \geq 0} f(n, k) \frac{t^{n-1}}{(n-1)!} x^k \\ &= \sum_{n \geq 1, k \geq 1} k f(n-1, k) \frac{t^{n-1}}{(n-1)!} x^k \\ &\quad + \sum_{n \geq 1, k \geq 0} f(n-1, k) \frac{t^{n-1}}{(n-1)!} x^k \\ &\quad + \sum_{n \geq 2, k \geq 1} f(n-2, k-1) \frac{t^{n-1}}{(n-2)!} x^k \\ &= x \sum_{n, k \geq 1} k f(n, k) \frac{t^n}{n!} x^{k-1} + \sum_{n, k \geq 0} f(n, k) \frac{t^n}{n!} x^k \\ &\quad + tx \sum_{n, k \geq 0} f(n, k) \frac{t^n}{n!} x^k \\ &= x \frac{\partial}{\partial x} F(t, x) + F(t, x)(1 + tx). \end{aligned} \quad (9)$$

Therefore, $F(t, x)$ satisfies the following partial differential equation:

$$\frac{\partial}{\partial t} F(t, x) - x \frac{\partial}{\partial x} F(t, x) = (1 + tx)F(t, x). \quad (10)$$

The solution of this equation with the boundary conditions $F(0, x) = 1$ and $F(t, 0) = e^t$ can be explicitly found to be

$$F(t, x) = e^{xe^t - x + (1-x)t}. \quad (12)$$

Equation (12) can be used to find $f(n, k)$ for arbitrary values of n and k , as well as the associated asymptotic behavior. Note that $f(n, k)$ is the sequence known as A124324 in [18], where Eq. (12) is also given.

D. Loop networks with a fixed number of singletons

We can refine Eq. (7) to enumerate configurations with a fixed number of singletons. Let $f(n, k, \ell)$ be the number of configurations with n TUs, k clusters, and ℓ singletons. This

quantity satisfies the following recursion relation:

$$\begin{aligned} f(n, k, \ell) &= k f(n-1, k, \ell) + f(n-1, k, \ell-1) \\ &\quad + (\ell+1) f(n-1, k-1, \ell+1). \end{aligned} \quad (13)$$

Indeed, we can think of producing, in a unique way, a configuration with n TUs from a configuration with $n-1$ TUs by introducing the TU n . The possible disjoint options are as follows:

(i) n joins an existing cluster (with at least two TUs in it) and there are $k f(n-1, k, \ell)$ possibilities in this case;

(ii) n is a singleton and there are $f(n-1, k, \ell-1)$ possibilities in this case;

(iii) n forms a cluster with precisely one other TU, in which case there are $(\ell+1) f(n-1, k-1, \ell+1)$ possibilities.

By iterating the recursion relation (13) with the easily checkable base $f(1, 0, 1) = 1$, and $f(1, k, \ell) = 0$ otherwise, we obtain

(i) $f(2, 0, 1) = 0$, $f(2, 0, 2) = 1$, $f(2, 1, 0) = 1$, recovering the total number of configurations with two TUs, $B_2 = 2$;

(ii) $f(3, 0, 2) = 0$, $f(3, 0, 3) = 1$, $f(3, 1, 0) = 1$, $f(3, 1, 1) = 3$, recovering the total number of configurations with three TUs, $B_3 = 5$;

(iii) $f(4, 0, 4) = 1$, $f(4, 1, 0) = 1$, $f(4, 1, 1) = 4$, $f(4, 1, 2) = 6$, $f(4, 2, 0) = 3$, recovering the total number of configurations with four TUs, $B_4 = 15$;

(iv) $f(5, 0, 5) = 1$, $f(5, 1, 0) = 1$, $f(5, 1, 1) = 5$, $f(5, 1, 2) = 10$, $f(5, 1, 3) = 10$, $f(5, 2, 0) = 10$, $f(5, 2, 1) = 15$, recovering the total number of configurations with five TUs, $B_5 = 52$, and so on.

By using a similar approach as in Sec. II C, we can define the following exponential generating function:

$$F(t, x, y) = \sum_{n, k, \ell \geq 0} f(n, k, \ell) \frac{t^n}{n!} x^k y^\ell, \quad (14)$$

which obeys the following partial differential equation,

$$\frac{\partial}{\partial t} F(t, x, y) - x \frac{\partial}{\partial x} F(t, x, y) - x \frac{\partial}{\partial y} F(t, x, y) = y F(t, x, y). \quad (15)$$

Quite remarkably, the physically relevant solution of this more complex equation can also be explicitly found and is given by

$$F(t, x, y) = e^{yt} e^{x(e^t - 1 - t)}. \quad (16)$$

Note that this solution satisfies the following boundary conditions: (i) $F(0, x, y) = 1$, (ii) $F(t, 0, y) = e^{yt}$, and (iii) $F(t, x, 0) = e^{x(e^t - 1 - t)}$. Once more, Eq. (16) can be expanded to yield coefficients $f(n, k, \ell)$, therefore solving the problem of enumerating all configurations with a fixed number of TUs, clusters, and singletons.

E. Results for networks without singletons

It is sometimes useful, or of interest, to consider the case where there are no singletons in the configuration. This is, for instance, the case that is considered in the companion paper [17]. If we denote by $N(n, k)$ the number of configurations with n TUs, k clusters, and no singletons, such that

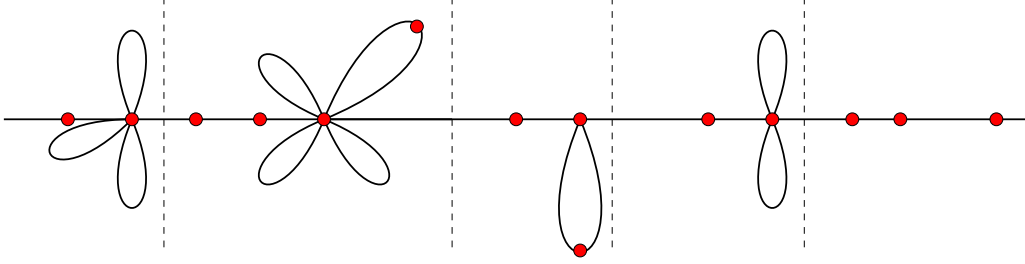


FIG. 3. Example of a reducible network and of its decomposition into irreducible blocks (here separated by dashed vertical lines). The shown configuration is a string of rosettes.

$N(n, k) = f(n, k, 0)$, we find that for $k = 2$,

$$\begin{aligned} N(n, 2) &= 2^{n-1} - n - 1 = f(n-1, 1) - 1 \\ &= f(n-1, 1) - N(n-1, 1). \end{aligned} \quad (17)$$

This equation can be derived by noting that the configurations of the chain can be constructed by assigning the first bead to cluster 0, and computing the number of configurations of the rest of the TUs with a single cluster and an arbitrary number of singletons. The singletons are then put in the same cluster as the first bead. In this way, we obtain all configurations with two clusters and no singletons once we subtract the single configuration which has no singletons in the rest of the chain [as this would lead to a configuration where the first TU is a singleton, which does not contribute to $N(n, 2)$].

A similar argument leads to the general identity

$$N(n, k) = f(n-1, k-1) - N(n-1, k-1), \quad (18)$$

linking the number of configurations with a given number of clusters with and without singletons.

The quantities $N(n, k)$ obey the following recursion relation [21–23]:

$$N(n, k) = kN(n-1, k) + (n-1)N(n-2, k-1). \quad (19)$$

Similarly to what was previously done, starting from Eq. (19), we can find the following exponential generating function for $N(n, k)$:

$$G(t, x) = \sum_{n \geq 0} N(n, k) \frac{t^n}{n!} x^k, \quad (20)$$

to be given by

$$G(t, x) = e^{x(e^t - 1 - t)}. \quad (21)$$

The related quantities

$$g_k(t) = \sum_{n \geq 0} N(n, k) \frac{t^n}{n!} \quad (22)$$

can now be found exactly for each k and are given by [21]

$$g_k(t) = \frac{(e^t - 1 - t)^k}{k!}. \quad (23)$$

F. String of rosettes and reducible networks

A natural question is whether a particular configuration can be broken up, or reduced, into a series of simpler configurations. To characterize such states, we call a configuration with

n TUs *irreducible* if it contains no cluster and only singletons, or it has k clusters, one of which contains TU n , and it is not possible to separate the k clusters into two groups by cutting a single polymer segment.

An example of a reducible network is a string of rosettes, shown in Fig. 3, where each irreducible component has a single cluster, at most. The decomposition into irreducible blocks is always unique, assuming that if the configuration has at least one cluster, then the leftmost irreducible block has a cluster.

We next derive the ordinary generating function $A(t)$ for the number of configurations in a string of rosettes. Note that there are $2^{n-1} - 1$ irreducible configurations with n TUs with a cluster, as this is precisely the number of ways to choose at least one TU to join n in the cluster. The generating function for these numbers is

$$\begin{aligned} I(t) &= \sum_{n \geq 2} (2^{n-1} - 1)t^n = t \sum_{n \geq 2} (2t)^{n-1} - \sum_{n \geq 2} t^n \\ &= t \left(\frac{1}{1-2t} - 1 \right) - \left(\frac{1}{1-t} - t - 1 \right) \\ &= \frac{t^2}{(1-2t)(1-t)}. \end{aligned} \quad (24)$$

Noting that the generating function for irreducible blocks without clusters is $\frac{1}{1-t}$, we have

$$\begin{aligned} A(t) &= \frac{1}{1-I(t)} \frac{1}{1-t} \\ &= \frac{1-2t}{1-3t+t^2} \\ &= 1 + t + 2t^2 + 5t^3 + 13t^4 + 34t^5 + 89t^6 + O(t^7). \end{aligned} \quad (25)$$

The corresponding sequence is A001519 in [18] and it has many combinatorial interpretations. One can derive from the generating function through the recurrence relation that

$$a_n = (\phi^{2n-1} + \phi^{1-2n})/\sqrt{5}, \quad (26)$$

where $\phi = (1 + \sqrt{5})/2$, and hence, asymptotically, the number of configurations in a string of rosettes is $O[(\frac{3+\sqrt{5}}{2})^n] \approx O(2.618^n)$.

The concept of reducible networks would be useful to enumerate configurations of longer chains that we consider in this work. Additionally, strings of rosettes appear often in

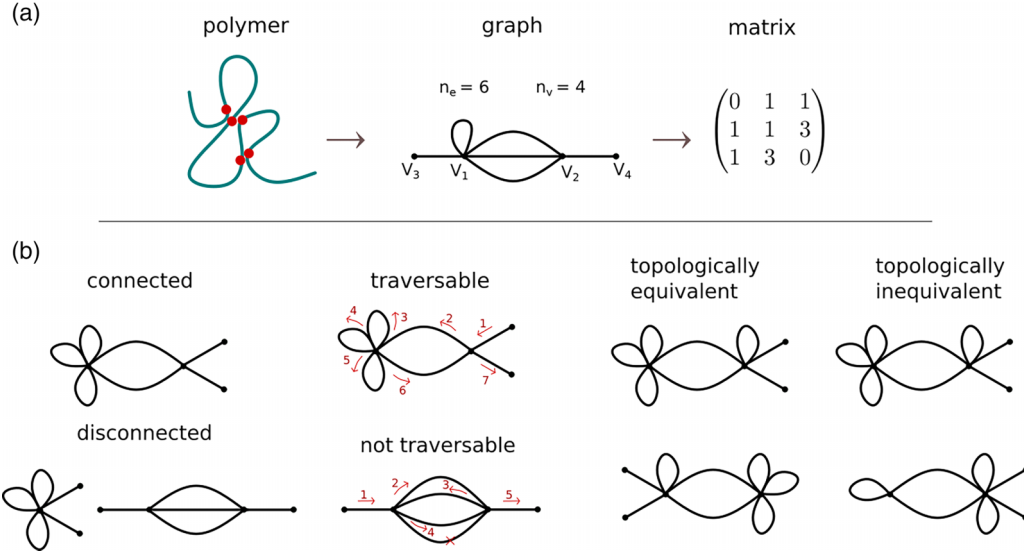


FIG. 4. (a) Schematics showing how polymer networks can be converted into graphs, and graphs to matrices. (b) Examples of connected and disconnected, traversable and not traversable, topologically equivalent and inequivalent graphs (or equivalently polymer networks).

simulations and it is therefore useful to provide a way to separately count the number of possible configurations leading to this specific type of polymer network.

III. COMBINATORICS OF INEQUIVALENT TOPOLOGIES FOR UNLABELED NETWORKS

We now discuss the case of unlabeled networks, which as anticipated is of interest when discussing the relative frequencies of different network topologies, irrespective of the specific labeling that is chosen. This is relevant, for instance, when asking whether, in a simulation or experiment, rosette topologies are more or less common than watermelon ones.

To study this case, we will be mapping polymer networks to graphs and matrices. While this mapping is not necessary to derive the formula we will give below, which holds for $k = 2$ clusters, it provides a useful framework to build, for instance, numerical algorithms which can enumerate all possible inequivalent topologies with a larger number of clusters, k .

Specifically, we begin by noting that the network topologies assembled by joining the TUs of a polymer can be mapped to graphs with n_v vertices and n_e edges [see Fig. 4(a)]. Each vertex of the graph corresponds to either a cluster of TUs or to one of the two polymer ends, while each edge of the graph corresponds to a polymer segment between two TUs or between one TU and one of the polymer ends.

We note that not all graphs can be representations of a polymer with TUs: since they are associated with a folded polymer, graphs representative of a chromatin loop network must be connected and traversable [see Fig. 4(b)]. Additionally, both n_v and n_e are constrained by the number of TUs, n . If none of the TUs coincides with the ends of the polymer, and if singletons are disallowed [as in Fig. 4(a)], then $n_e = n + 1$ and $3 \leq n_v \leq \lfloor \frac{n+1}{2} \rfloor + 2$, where $p = n_e - 2$ and $\lfloor x \rfloor$ denotes the floor of x (the largest integral smaller than or equal to x). Two of these vertices correspond to the polymer ends, and their degree is 1; we will call all the others *internal* vertices,

namely, all the vertices associated with clusters of TUs [V_1 and V_2 in Fig. 4(a)] [24].

A. Enumeration of inequivalent topologies with two clusters

We now proceed to count the number of topologically inequivalent, connected, and traversable graphs with a given number $n + 2$ of edges and 4 vertices, V_1, V_2, V_3, V_4 , two of which, V_3, V_4 , are of degree 1. This is the number of topologically inequivalent networks with n TUs and $k = 2$ clusters, without any singletons, studied in the companion paper [17].

Let $\mathcal{G}$ be a graph of this kind. $\mathcal{G}$ is identified by five numbers: a, b, c, n_1 , and n_2 . Of these, a and b denote the number of edges connecting, respectively, vertex V_1 and vertex V_2 to themselves, c is the number of edges connecting vertex V_1 to vertex V_2 , while n_1 and n_2 are the numbers of vertices of degree 1 connected, respectively, to V_1 and V_2 [for instance, the network in Fig. 4(a) has $n_1 = n_2 = 1$, whereas the top left graph in Fig. 4(b) has $n_1 = 2, n_2 = 0$]. The following symmetric matrix, therefore, identifies $\mathcal{G}$ in a compact way [see Fig. 4(a)]:

$$M(\mathcal{G}) = \begin{pmatrix} 0 & n_1 & n_2 \\ n_1 & a & c \\ n_2 & c & b \end{pmatrix}. \quad (27)$$

To count the number of graphs we are interested in, we remark the following:

(i) $\mathcal{G}$ and $\mathcal{G}'$ are equivalent if and only if

$$P^T M(\mathcal{G}) P = M(\mathcal{G}'), \quad (28)$$

where P is one of the two permutation matrices,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \quad (29)$$

In this case, we say that $M(\mathcal{G})$ and $M(\mathcal{G}')$ are equivalent (i.e., they represent equivalent graphs).

(ii) $\mathcal{G}$ is disconnected if and only if $c = 0$.

(iii) $n - 1 = a + b + c$.

(iv) $\deg(V_1) = n_1 + 2a + c \geq 3$ and $\deg(V_2) = n_2 + 2b + c \geq 3$.

(v) Since a connected graph is traversable if and only if the number of vertices with odd degree is either 0 or 2 [25], $\deg(V_1)$ and $\deg(V_2)$ must be even. Moreover, since $\deg(V_1) = n_1 + 2a + c$ and $\deg(V_2) = n_2 + 2b + c$, we have the following cases: (A) if c is even, either $n_1 = 2$ and $n_2 = 0$, or $n_1 = 0$ and $n_2 = 2$; (B) if c is odd, $n_1 = 1$ and $n_2 = 1$.

Let us call $\{M\}_G$ the set of all inequivalent [according to point (i) above] matrices representing a graph with the desired constraints. To count the inequivalent topologies, let us consider the map $f : (a, b, c) \mid a, b, c \in \mathbb{N}, c \geq 1, a + b + c = n \rightarrow M \in \{M\}_G$ defined as

$$f(a, b, c) = \begin{cases} \begin{pmatrix} 0 & 2 & 0 \\ 2 & a & c \\ 0 & c & b \end{pmatrix} & \text{if } c \text{ is even} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & a & c \\ 1 & c & b \end{pmatrix} & \text{if } c \text{ is odd.} \end{cases} \quad (30)$$

This map covers all the desired inequivalent topologies, but it is not injective [26]. The number of possible combinations of (a, b, c) satisfying the constraints $n = a + b + c$, $a \geq 0$, $b \geq 0$, and $c \geq 1$ is given by

$$\sum_{i=1}^{n-1} (n-i) = \frac{n(n-1)}{2}. \quad (31)$$

From this number, we first need to identify the combinations of (a, b, c) which map to equivalent graphs (or matrices), then to take away those which would lead to multiple counting of the same topology, and finally to remove the combinations which do not satisfy point (iv).

To do so, we note that the graph equivalence condition $P^T M P = M'$ with $M \neq M'$ requires $a = b'$, $b = a'$, $n_1 = n'_2$, and $n_2 = n'_1$. If c is even, this is never met by construction; if c is odd, (a, b, c) and (b, a, c) are mapped to equivalent matrices. To account for this, and avoid double counting of these equivalent topologies, we require $a \geq b$, a condition which removes $\sum_{\text{odd } c \leq n} \frac{c-1}{2}$ possibilities: equivalently, $\sum_{i=1}^{\frac{n-1}{2}} i$ combinations if n is odd, and $\sum_{i=1}^{\frac{n}{2}-1} i$ combinations if n is even.

Finally, to account for point (iv), we also remove the two configurations $(a = n - 2, b = 0, c = 2)$ and $(a = n - 1, b = 0, c = 1)$ from the total count [27].

The total number of inequivalent graphs with n TUs and two clusters is then given by

$$\begin{aligned} N_u(n, 2) &= \frac{n(n-1)}{2} - 2 - \sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} i \\ &= \frac{n(n-1)}{2} - 2 - \frac{\lfloor \frac{n-1}{2} \rfloor \lfloor \frac{n+1}{2} \rfloor}{2}. \end{aligned} \quad (32)$$

B. Network multiplicities

Note that for each of the unlabeled network topologies just found, there are multiple possible labeled configurations that correspond to it. As these combinatorial weights, or multiplicities, are generally different for different topologies, it is

TABLE I. Topology summary table. All topologies with $n = 8$ binding sites and two clusters (graph vertices) are listed, together with their number of ties (n_t), number of loops (n_l), nontrivial vertex orders (L_1 and L_2 for first and second cluster), and multiplicity (Ω). The last two columns give the critical energy between TUs needed to form the topology, ε_c , in units of $k_B T$, together with the 95% confidence interval (CI): these results correspond to the simulations presented in Sec. IV. The set of diagrams can be divided into three classes, each characterized by the same pair of nontrivial vertex orders (L_1, L_2) [or (L_2, L_1)]. Using the order in which these are shown in the table, these classes are given by the first 7 diagrams, the 8 following diagrams, and the final 5 ones.

Diagram	n_t	n_l	L_1	L_2	Ω	$\varepsilon_c/(k_B T)$	CI/($k_B T$)
	1	6	8	8	1	9.1	[9.0, 9.5]
	2	5	8	8	3	9.7	[9.5, 10.1]
	3	4	8	8	9	9.8	[9.5, 10.2]
	4	3	8	8	9	10.4	[10.1, 10.5]
	5	2	8	8	9	11.1	[10.0, 11.1]
	6	1	8	8	3	10.9	[10.6, 11.5]
	7	0	8	8	1	11.2	[11.0, 11.4]
	1	6	6	10	2	8.8	[8.7, 9.3]
	2	5	10	6	4	9.6	[9.4, 10.0]
	2	5	6	10	2	9.0	[8.8, 9.3]
	3	4	6	10	16	9.3	[8.9, 9.4]
	4	3	10	6	12	10.1	[10.0, 11.0]
	4	3	6	10	4	9.2	[9.2, 9.5]
	5	2	6	10	12	9.8	[9.5, 10.2]
	6	1	10	6	4	10.6	[10.5, 11.3]
	1	6	4	12	2	8.7	[8.6, 9.0]
	2	5	12	4	5	9.1	[8.8, 9.3]
	2	5	4	12	1	8.8	[8.5, 9.0]
	3	4	4	12	10	9.0	[8.5, 9.3]
	4	3	12	4	10	9.3	[9.1, 9.3]

desirable to keep track of these. It is, however, difficult to go beyond a case-by-case study. Here, we focus on the case of $n = 8$ TUs and $k = 2$ clusters, studied in [17], for which there

are 20 inequivalent topologies [as predicted by Eq. (32) for $n = 8$].

In this case, for each inequivalent topology, Table I provides the number of ties, n_t , the number of loops, n_l , the degree of the two vertices in the graphs (corresponding to the clusters), L_1 and L_2 , respectively, and the multiplicity of the topology Ω , which is the number of labeled configurations corresponding to that topology. The degrees L_1 and L_2 determine the entropic exponent of the polymer network [16]: it can be seen that there are three classes of such exponents in the 20 topologies considered in Table I.

Regarding multiplicities, we observe that these tend to be larger for hybrid networks which are intermediate between rosettes and watermelons and that multiplicities are, in general, small for networks with low n_l . This would suggest that in the absence of other biases, such configurations would form less often. We shall see in what follows, however, that these topologies are actually easier to form, so there is an interesting competition between the combinatoric multiplicity and the entropic cost of formation of these loop networks.

Note that as required, the total sum of multiplicities for all topologies is $N(8, 2) = 119$, namely, the number of two-cluster configurations with $n = 8$ without singletons found previously [see Eq. (17)].

IV. COARSE-GRAINED MOLECULAR DYNAMICS SIMULATIONS

Having discussed the combinatorial problem of enumerating the possible configurations and inequivalent topologies of a chromatin loop network, we now turn to the associated polymer physics problem and ask what interaction between TUs needs to be included to form a target topology in practice. This calculation requires computer simulations for polymer models representing chromatin fibers, and therefore here we use coarse-grained molecular dynamics simulations to study this problem.

In this section, we will first describe the model that is used and then present our simulation results, whose main outcome will be to show that different topologies require significantly different energy inputs to form. This energy of formation will combine in practical examples with the combinatoric multiplicities discussed above (the relevant ones for the case at hand are those given in Table I) to determine the likeliness of observing a given topology in an unconstrained polymer simulation, such as the one discussed in [17].

A. Model and potentials used

We model a chromatin fiber as a bead-and-spring polymer. The underlying equations of motion are the set of Langevin equations for each bead,

$$m \frac{d^2 \mathbf{x}_i}{dt^2} = -\nabla_i U - \gamma \frac{d\mathbf{x}_i}{dt} + \sqrt{2k_B T \gamma} \boldsymbol{\eta}(t), \quad (33)$$

where m is the bead mass, $\mathbf{x}_i$ is the i th bead position, γ is the drag, and $\boldsymbol{\eta}(t)$ is uncorrelated white noise defined by $\langle \boldsymbol{\eta}(t) \rangle = \mathbf{0}$ and $\langle \eta_\alpha(t) \eta_\beta(t') \rangle = \delta_{\alpha\beta} \delta(t - t')$. This Langevin equation imposes an NVT ensemble on the system within

which fluctuation and dissipation govern the exploration of the configuration space.

In order to reproduce behavior that is appropriate to chromatin, the following potentials U enter into this equation according to the particular beads under consideration. A simple phenomenological Lennard-Jones potential, truncated to include only the repulsive regime (Weeks-Chandler-Andersen potential) acts between all beads in the system enforcing excluded volume, or self-avoidance. This is given by

$$U_{LJ}(r_{ij}) = \begin{cases} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right] + \epsilon & \text{if } r_{ij} < 2^{1/6} \sigma \\ 0 & \text{otherwise,} \end{cases} \quad (34)$$

where σ is the bead diameter.

To capture chain connectivity, finite extensible nonlinear elastic (FENE) bonds are considered, acting only between consecutive beads along the polymer chain,

$$U_{FENE}(r) = -\frac{1}{2} K R_0^2 \ln \left[1 - \left(\frac{r}{R_0} \right)^2 \right] \quad (35)$$

if $r < R_0$, and ∞ otherwise, where $K = 30k_B T / \sigma^2$ is the spring constant and $R_0 = 1.6\sigma$ is the maximum extent of the bond. While FENE springs are used, in line with the literature on chromatin modeling, we expect strong harmonic bonds will lead to equivalent results, and could have been used instead.

Finally, we add a bending or Kratky-Porod potential, which acts on the angle θ between three consecutive beads along the chain and enforces a nonzero persistence length l_p ,

$$U_{\text{bending}} = K_b [1 + \cos(\theta)], \quad (36)$$

where $K_b = k_B T l_p / \sigma = 3k_B T$. Note that the persistence length is artificially raised at the beginning of equilibration to remove overlaps and assist the system in reaching a self-avoiding configuration. The persistence length is then lowered from 10σ to 3σ , which is an appropriate value for flexible chromatin [13,14,28]. The polymer is further equilibrated in the presence of excluded volume interactions only. Subsequently, attractive interactions are switched on for the production runs. We note that as for naked DNA [29], chromatin stiffness may depend on epigenetic modifications and, for instance, be higher in inactive chromatin, or heterochromatin [30].

Specifically, to study the formation of a target topology, we include attractive interactions between beads that should be in the same cluster in the target topology; this procedure is similar to what is done in a Go model approach to study protein folding, where only attractive interactions between residues in contact in the folded state are included [31]. We consider all 20 topologies in Table I; for instance, for a symmetric two-rosette state, we include an interaction between the first four TUs and between the last four. The attraction between the selected TUs is simulated by a Lennard-Jones potential, where part of the attractive tail is retained. The interaction range (cutoff) is set to 1.8σ , while the interaction strength ϵ is varied between $5k_B T$ and $15k_B T$. For large enough ϵ , the target topology is formed, with all interactions realized. (Note that not all topologies may be realizable if the steric interactions between different polymer segments prevent this, but in our case, this is not an issue.)

The transition between an initially unstructured chain and the target topology arises because the energy gained as binding sites come together increases (asymptotically) linearly with the number of binding beads in a cluster, whereas the free energy cost of adding loops to a cluster scales superlinearly [12,16,32]. As such, there is a critical energy ϵ_c at which the energy gain just offsets the entropic loss and this is the transition point. Our goal is to find how the value of ϵ_c depends on topology. Note that all topologies that we compare (i.e., within each of the three classes in Table I) contain the same number of binding site interactions, hence the same total maximum energy. The difference in ϵ_c is then primarily due to the free energy cost of forming that specific target topology.

Our script loads a modular pair coefficient file generated by a simple PYTHON script. This allows the target polymer network topologies to be easily specified and changed, while keeping other elements of the simulation fixed, which is important for reproducibility, scalability, and comparisons. Finally, the system is evolved via Langevin dynamics using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package [33].

B. Target topology simulations: Rosettes, watermelons, and dependence on the number of ties

As discussed above, in the thermodynamic limit, it is expected that the entropic exponent of forming a given topology should depend only on the number of legs ($L_{1,2}$) meeting at its vertices [16,32]. This partitions the set of 20 inequivalent topologies into three classes that have the same values of $L_{1,2}$ (see Table I), so that the results discussed below should be compared only between topologies in the same class.

Among the first class (first seven topologies in Table I), two topologies, namely, the *rosette* (top topology in the class) and the *watermelon* (bottom topology in the class), stand out as particularly illustrative choices to discuss the results of the simulations. For these two topologies, simulations are carried out by varying the interaction affinity ϵ between $5k_B T$ and $15k_B T$. From the estimate of the pairing energy normalized by the number of beads, ϵ_{pair} , we can identify the values of ϵ for which the target topology is formed (Fig. 5).

As expected, for small values of ϵ , the chain remains unfolded. In contrast, for sufficiently large values of ϵ , the targeting topology is formed (examples of folded configurations in this regime for the rosette and watermelon topologies are shown in Figs. 6 and 7, respectively). The point of sharpest variation of the sigmoidal curves in Fig. 5 can be interpreted as the critical interaction affinity ϵ_c required to form the target topology (either rosette or watermelon). Thicker lines indicate the mean of ϵ_{pair} over 100 random initial configurations for each ϵ used. The surrounding shaded regions represent one standard deviation on either side of this mean.

The interaction affinity giving rise to the maximal standard deviation is taken as the recorded transition affinity ϵ_c . Confidence intervals are computed using a bootstrapping procedure. In order to better illustrate the relationship between the standard deviation amongst simulations with the same interaction affinity and the inferred transition affinity, the standard deviations are plotted independently in Fig. 8. From these curves, it is clear that the location of the maximum differs

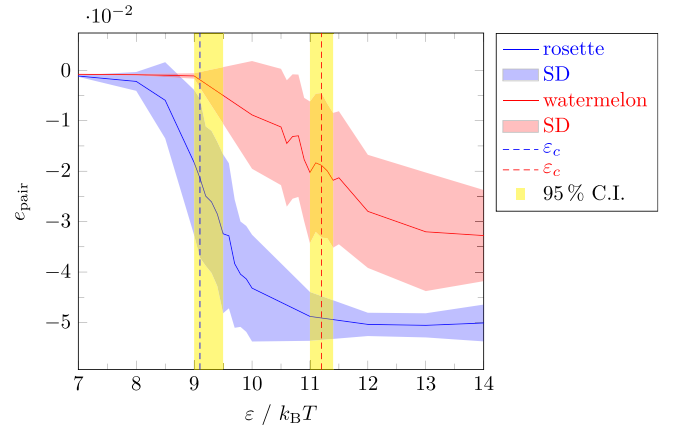


FIG. 5. Plot of the average measured pairing attractive energy per bead, ϵ_{pair} , as a function of the input attractive energy ϵ . The sigmoidal shape of the curve signals a transition to the formation of the target topology.

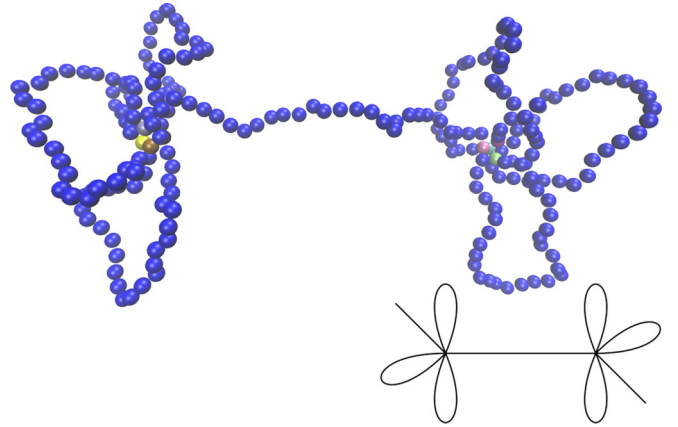


FIG. 6. Simulation snapshot and corresponding topology for the rosette case. Note different TUs are colored differently.

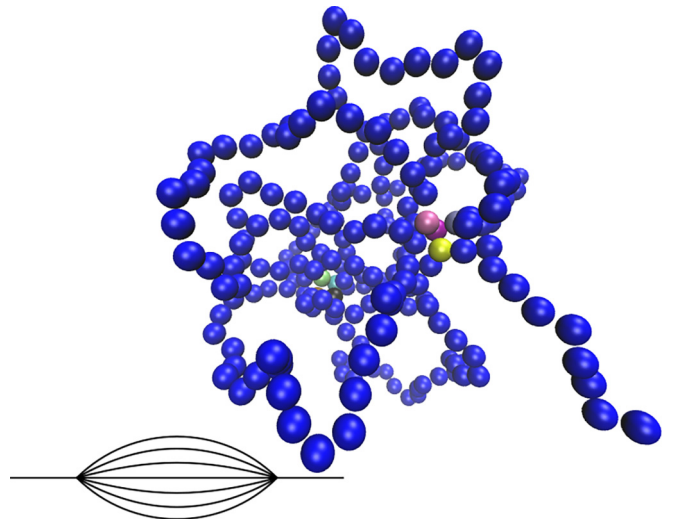


FIG. 7. Simulation snapshot and corresponding topology for the watermelon case. Note different TUs are colored differently.

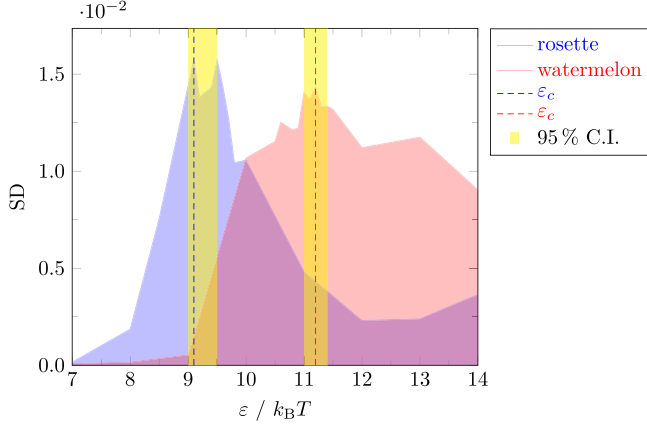


FIG. 8. Plot showing simulation results for the standard deviation of the normalized pairing energy $\varepsilon_{\text{pair}}$ as a function of the attraction between TUs for the rosette (blue) and watermelon (red) topologies. Maxima are used to infer the transition affinities, which are indicated with dashed lines. Bootstrapped 95% confidence intervals are shaded in yellow.

for the rosette and watermelon topologies. In particular, one can observe that the rosette topology forms more easily, as it requires a smaller value of ε or, equivalently, the associated ε_c (corresponding to the peak in Fig. 8) is smaller.

Note that while the rosette and watermelon topologies have the same values of $L_{1,2}$, they differ by the number of ties, n_t ($n_t = 0$ for the rosette topology and $n_t = 7$ for the watermelon one). In general, we observed that the larger n_t in a target topology, the greater the interaction affinity typically required to form it (with exceptions, see Table I). In this respect, a simple class to study is that of the first seven symmetric topologies shown in Table I, of which the rosette and the watermelon constitute the limiting cases. In order to elucidate the relationship between n_t and ε_c for this class, we carry out a bootstrapped linear regression. The corresponding fit is plotted in Fig. 9: a simple linear relationship between n_t

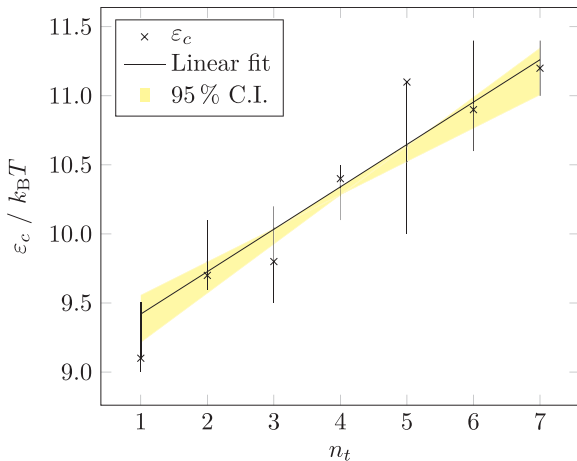


FIG. 9. Plot of the critical energy ε_c against the number of ties, n_t , for the first class of topologies in Table I, with $L_{1,2} = (8, 8)$. Values corresponding to 95% confidence intervals for each value of n_t are found by bootstrapping.

and ε_c holds to a good approximation. As the number of ties increases by one, the number of loops decreases by one too, and so our results indicate that there is a nearly uniform energetic cost each time one loop is exchanged for a tie in a chromatin network. The estimate of the constant cost per tie is $\Delta\varepsilon_c = 0.31k_B T$ (95% confidence interval: $0.24\text{--}0.36k_B T$). The other two classes of topologies reported in Table I still show an increase of ε_c with n_t , but the functional form is less clear (see Table I for a list of values of ε_c found for each inequivalent two-cluster topology).

V. TOPOLOGICAL WEIGHTS OF GAUSSIAN CHROMATIN LOOP NETWORKS

Up to now, we have enumerated the configurations of polymer loop networks, thereby finding their combinatorial weights. We have also seen in the last section that Brownian dynamics simulations show that the energy that is required to offset the free energy cost associated with the formation of these topologies is significantly different. In the companion paper, we have additionally shown that inequivalent (unlabeled) topologies with the same combinatorial weight, such as the rosette and watermelon ones, are observed in polymer models with starkly different frequencies. In this section, we will show that these results can be understood, at least qualitatively, by computing the *topological weight* of a given graph, which is essentially the partition function of a Gaussian polymer network with that topology. (Note that this is equivalent to a freely jointed polymer network with a large number of monomers [34].)

More specifically, to compute the topological weight of a given graph $\mathcal{G}$ associated with an inequivalent topology of a chromatin loop network with n TUs, we need to compute its corresponding partition function,

$$Z_{\mathcal{G}} = \int d\mathbf{x}_0 \dots d\mathbf{x}_{n+1} \delta(\mathcal{G}) \prod_{i=0}^n e^{-\frac{3(\mathbf{x}_{i+1}-\mathbf{x}_i)^2}{2l\sigma}}, \quad (37)$$

where l is the mutual distance between two consecutive TUs, σ is the bead size, and $\delta(\mathcal{G})$ is a product of Dirac delta functions that describes the topology of the network (see below).

Here, $e^{-\frac{3(\mathbf{x}_{i+1}-\mathbf{x}_i)^2}{2l\sigma}}$ can be thought of, in field theoretical terms, as the *propagator* of our Gaussian theory, from the i th to the $(i+1)$ -th TU.

In the remainder of this section, we will first compute in detail the topological weights of unlabeled configurations with two clusters, which are the focus of the numerical simulations in the companion paper [17]. Afterwards, we shall see how to generalize the calculation to compute the topological weight of any given Gaussian polymer loop network. This calculation can be done explicitly because we are approximating the polymer to a Gaussian chain.

Including self-avoidance and mutual avoidance between different polymer segments would require a separate treatment and is outside the scope of the current work. In the special case of two-cluster configurations, self-avoidance, or excluded volume, can be included by using the self-avoiding walk propagator [34] in place of the Gaussian propagator used in the theory just described. The resulting weights are calculated in the companion paper [17]. Excluded volume

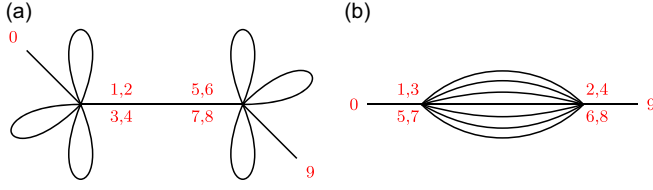


FIG. 10. Loop network configurations and TU labeling used for the calculation of the topological weights of the (a) rosette and (b) watermelon topologies.

effects coming from self-avoidance of polymer segments favor rosettes over watermelon even more than is predicted by the Gaussian theory. It would be of interest to also single out the effect of mutual avoidance between different polymer segments. We note that such excluded volume effects contribute to the difference in entropy between the different topologies because, in the absence of energetic contributions (such as from bending rigidity), all configurations are equal in energy, and hence the difference in weights stems from differences in entropy.

A. Topological weights of two-cluster configurations

We begin by noting that the term $\delta(\mathcal{G})$ in Eq. (37) is a product of Dirac δ functions which specify the topology of the network [16]. For instance, in the case of rosettes [$\mathcal{G} = \mathcal{R}$, Fig. 10(a)] and watermelons [$\mathcal{G} = \mathcal{W}$, Fig. 10(b)], $\delta(\mathcal{G})$ is explicitly given by $\delta(\mathcal{R})$ and $\delta(\mathcal{W})$, with

$$\begin{aligned}\delta(\mathcal{R}) &= \prod_{i=2,3,4} \delta(\mathbf{x}_1 - \mathbf{x}_i) \prod_{j=6,7,8} \delta(\mathbf{x}_5 - \mathbf{x}_j), \\ \delta(\mathcal{W}) &= \prod_{i=3,5,7} \delta(\mathbf{x}_1 - \mathbf{x}_i) \prod_{j=4,6,8} \delta(\mathbf{x}_2 - \mathbf{x}_j).\end{aligned}\quad (38)$$

The topological weight of the rosette topology is therefore given by

$$\begin{aligned}Z_{\mathcal{R}} &= \int d\mathbf{x}_0 \dots d\mathbf{x}_9 \left[\prod_{i=0}^8 e^{-\frac{3(\mathbf{x}_{i+1} - \mathbf{x}_i)^2}{2l\sigma}} \right] \\ &\times \prod_{i=2,3,4} \delta(\mathbf{x}_1 - \mathbf{x}_i) \prod_{j=6,7,8} \delta(\mathbf{x}_5 - \mathbf{x}_j),\end{aligned}\quad (39)$$

where the TU labeling in the integral follows the one in Fig. 10(a).

Noting that

$$\int d\mathbf{x}_0 e^{-\frac{3(\mathbf{x}_1 - \mathbf{x}_0)^2}{2l\sigma}} = \int d\mathbf{x}_0 e^{-\frac{3\mathbf{x}_0^2}{2l\sigma}} = W_0^{-1}, \quad (40)$$

with $W_0 \equiv (\frac{3}{2\pi l\sigma})^{3/2}$, and that an analogous formula holds for the integral over $d\mathbf{x}_9$, we obtain, by making use of the properties of the δ function, that

$$Z_{\mathcal{R}} = W_0^{-2} \int d\mathbf{x}_1 d\mathbf{x}_5 e^{-\frac{3(\mathbf{x}_1 - \mathbf{x}_5)^2}{2l\sigma}} = W_0^{-3} V, \quad (41)$$

where we have called V the volume of the system.

By repeating the same steps for the watermelon topology [see Fig. 10(b) and the associated choice of TU labeling], we

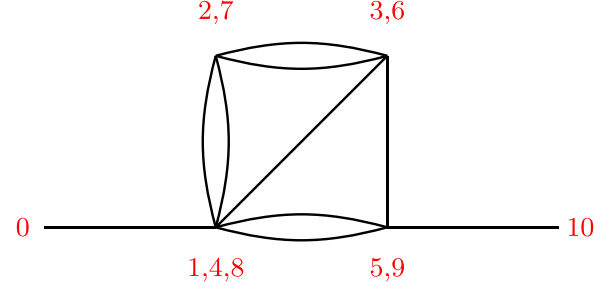


FIG. 11. Loop network configuration and TU labeling used for the calculation of the topological weights of a network with multiple clusters (here, four).

get

$$Z_{\mathcal{W}} = W_0^{-2} \int d\mathbf{x}_1 d\mathbf{x}_2 e^{-7 \frac{3(\mathbf{x}_2 - \mathbf{x}_1)^2}{2l\sigma}} = \frac{W_0^{-3} V}{7^{3/2}} = \frac{Z_{\mathcal{R}}}{7^{3/2}}. \quad (42)$$

Therefore, the topological weight of the watermelon is much smaller than that of the rosette. Additionally, one can generalize the result shown above to hybrid rosette-watermelon configurations with two clusters and n_t ties, obtaining that their topological weight is given by

$$Z_{\mathcal{G}} = \frac{Z_{\mathcal{R}}}{n_t^{3/2}}, \quad (43)$$

which becomes Eq. (42) for $n_t = 7$ (which holds for the watermelon topology). The decrease in topological weight of two-cluster topologies with n_t qualitatively explains why they are seen less frequently in simulations [17] and why the interaction energy between TUs needed to stabilize a topology increases with n_t , as found in the previous section with coarse-grained molecular dynamics simulations.

B. General formulas for the topological weights of Gaussian loop networks

With a bit more work, the topological weight calculation just outlined can actually be generalized to any chromatin loop network.

To see how, let us consider the topology $\mathcal{G}$ shown in Fig. 11. Its associated topological weight is given by

$$\begin{aligned}Z_{\mathcal{G}} &= \int d\mathbf{x}_0 \dots d\mathbf{x}_{10} \left[\prod_{i=0}^9 e^{-\frac{3(\mathbf{x}_{i+1} - \mathbf{x}_i)^2}{2l\sigma}} \right] \delta(\mathbf{x}_1 - \mathbf{x}_4) \delta(\mathbf{x}_1 - \mathbf{x}_8) \\ &\times \delta(\mathbf{x}_2 - \mathbf{x}_7) \delta(\mathbf{x}_3 - \mathbf{x}_6) \delta(\mathbf{x}_5 - \mathbf{x}_9),\end{aligned}\quad (44)$$

which, by using methods similar to those described in the previous section, can also be written as

$$Z_{\mathcal{G}} = W_0^{-2} \int d\mathbf{x}_1 d\mathbf{x}_2 d\mathbf{x}_3 d\mathbf{x}_5 e^{-\frac{3}{2l\sigma} f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5)}, \quad (45)$$

where

$$\begin{aligned}f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5) &= [2(\mathbf{x}_2 - \mathbf{x}_1)^2 + 2(\mathbf{x}_3 - \mathbf{x}_2)^2 \\ &+ (\mathbf{x}_3 - \mathbf{x}_1)^2 + (\mathbf{x}_5 - \mathbf{x}_3)^2 + 2(\mathbf{x}_5 - \mathbf{x}_1)^2].\end{aligned}\quad (46)$$

We now introduce the following matrix:

$$A(\mathcal{G}) = \begin{pmatrix} 0 & 2 & 1 & 2 \\ 2 & 0 & 2 & 0 \\ 1 & 2 & 0 & 1 \\ 2 & 0 & 1 & 0 \end{pmatrix}, \quad (47)$$

which equals to the adjacency matrix of the multigraph corresponding to $\mathcal{G}$ [note that if loops were present in $\mathcal{G}$, they should not be included in the calculation of $A(\mathcal{G})$]. From this, we define the matrix

$$B(\mathcal{G}) = \begin{pmatrix} 5 & -2 & -1 & -2 \\ -2 & 4 & -2 & 0 \\ -1 & -2 & 4 & -1 \\ -2 & 0 & -1 & 3 \end{pmatrix}, \quad (48)$$

where we have changed the sign of the off-diagonal components and the diagonal components have been set equal to the sum of the corresponding row in $A(\mathcal{G})$. With this setup, the argument of the exponential in Eq. (11) can be written in matrix form as

$$f(\mathbf{x}) = \mathbf{x}^T B(\mathcal{G}) \mathbf{x}, \quad (49)$$

where $\mathbf{x}^T = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5)$. Note that $\det[B(\mathcal{G})] = 0$. This is consistent with the fact that the topological weight is proportional to the volume V of the system. By fixing the position of one of the cluster's center of mass, say $\mathbf{x}_1$, and integrating over it, the weight associated with this topology can be given in terms of the determinant of the matrix obtained by removing the first row and column,

$$\det[B'(\mathcal{G})] = \det \begin{pmatrix} 4 & -2 & 0 \\ -2 & 4 & -1 \\ 0 & -1 & 3 \end{pmatrix} = 32, \quad (50)$$

as follows:

$$Z_{\mathcal{G}} = W_0^{-3} V \det[B'(\mathcal{G})]^{-3/2} = \frac{W_0^{-3} V}{32^{3/2}}. \quad (51)$$

It can be verified that as expected, the above result does not depend on which cluster is fixed and integrated upon, as the determinant of any matrix obtained by removing the i th row and column of $B(\mathcal{G})$ is the same. [Indeed, one can also show that all minors of $B(\mathcal{G})$ are the same up to a sign.]

By applying this procedure, for instance, to a string of rosettes with n TUs, one can show that the corresponding topological weight is $W_0^{-3} V$.

Finally, for a general graph $\mathcal{G}$ with n TUs and k clusters, its topological weight can be computed by starting from the corresponding matrix $B(\mathcal{G})$ and computing the determinant of any submatrix $B'(\mathcal{G})$ obtained by removing the i th row and column, for any $i \in [1, k]$. This is given by

$$Z_{\mathcal{G}} = W_0^{-3} V \det[B'(\mathcal{G})]^{-3/2}. \quad (52)$$

As a generic network can be obtained from a string of rosettes by adding a suitable amount of ties between clusters, which leads to a decrease in the integrand in Eq. (45), this means that $\det[B'(\mathcal{G})] \geq 1$. Therefore, the above result confirms that a generic network typically has a (significantly) lower weight with respect to that of a string of rosettes with the same number of TUs, n , in line with the numerical results obtained for $k = 2$.

VI. DISCUSSION AND CONCLUSIONS

In summary, we have presented a combination of analytical and numerical results for the combinatorial and topological weights of chromatin loop networks. These weights are important to determine the relative frequencies with which different topologies arise in polymer models for DNA and chromatin, which are studied in the companion paper [17]. In particular, we are interested here in the relation between these results and the physical properties of the loop networks which arise due to the bridging-induced attraction [13,14,28], in polymer models for the 3D structures formed by chromatin fibers *in vivo*, and which are associated with gene folding. For instance, the statistical, or Boltzmann, weight associated with a given topology shown in Table I, which determines the frequency with which it is observed, for instance, in computer simulations, is proportional to its combinatorial weight (computed in Sec. III) times its topological weight (computed in Sec. V).

We have shown that the enumeration problems associated with counting labeled and unlabeled chromatin loop networks are fundamentally different. When transcription units (TUs) are labeled (Sec. II), the problem can be usefully mapped to that of counting the ways in which n different TUs can be distributed into k clusters with ≥ 2 TUs per cluster. The resulting combinatorial sequences are often related to the Bell or Stirling numbers, and we have shown that it is possible to find explicit formulas for the exponential generating functions associated with a number of different cases, with or without singletons (i.e., TUs not in any clusters). This is useful for providing estimates or upper bounds for the number of topologies which a given chromatin region (with a specified number of TUs) can fold into.

For networks with unlabeled TUs, corresponding to inequivalent topologies (Sec. III), the enumeration problem is related to that of counting multigraphs, which is NP-complete and hence harder. We have, though, provided here a derivation of a formula counting all inequivalent topologies with n TUs and $k = 2$ clusters; for $n = 8$ (a common occurrence in real gene loci [3,17]) and $k = 2$, i.e., the case studied in detail in the continuum paper, this formula gives 20 inequivalent topologies (shown explicitly in Table I).

We also asked what attraction energy is needed to form a target topology. This is a biophysically relevant question regarding chromatin loop networks. For instance, we may want to know whether a rosette topology or a watermelon one forms more easily (i.e., requires less interaction between the TUs), as this may affect the relative frequency with which these two structures may be found in mammalian chromatin. These predictions could then be compared with computer simulations of 3D chromatin folding [3,17]. Previous work based on renormalization group calculations came to the important conclusion that the entropic exponent of a polymer loop network solely depends on the degree of its nodes (the clusters in our terminology) [16,32]. However, this exponent does not completely determine the weight, as there is a prefactor which can, in principle, also be topology dependent. More in detail, rosettes and watermelons, and indeed all first seven network topologies in Table I, have the same entropic exponent, yet they require significantly different energies to form, as we

show in Sec. IV. In particular, by focusing on configurations with $n = 8$ TUs and $k = 2$ clusters, our simulations show that the critical energy to form a target topology tends to increase with the number of ties, or polymer segments, linking the two clusters, which we call n_t .

Finally, in Sec. V, we have computed the topological weight of a chromatin loop network, under the assumption that the polymer is a Gaussian chain. This weight is the partition function of a network with the given topology and, importantly, we have found that it strongly depends on n_t , qualitatively explaining our numerical results in Sec. IV.

In the future, it would be interesting to generalize the topological weight calculations in Sec. V to the case where the polymer network has both self- and mutual avoidance (for a first step in this direction, see the companion paper [17]). From an application perspective, it would be desirable to use

our labeled and inequivalent unlabeled topologies to classify the 3D configurations of chromatin fiber around genes, for instance, the gene loci configuration found by “HiP-HoP” simulations in [3] or the interaction networks and hypergraphs found by chromatin capture experiments accounting for multiway chromatin contacts, such as poreC [35]. We hope that these extensions of our work will be addressed in the future.

ACKNOWLEDGMENTS

We thank Nick Sheridan for useful discussions. This work was supported by the Wellcome Trust (Grant No. 223097/Z/21/Z). The work of S.K. was supported by Leverhulme Research Fellowship (Grant No. RF-2023-065\9). E.O. acknowledges support from Grant No. PRIN 2022R8YXMR funded by the Italian Ministry of University and Research.

-
- [1] C. R. Calladine and H. Drew, *Understanding DNA: The Molecule and How it Works* (Elsevier Academic Press, San Diego, 1997).
- [2] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Taylor & Francis, London, 2014).
- [3] M. Chiang, C. A. Brackley, C. Naughton, R.-S. Nozawa, C. Battaglia, D. Marenduzzo, and N. Gilbert, *bioRxiv* (2022), doi:[10.1101/2022.06.09.495447](https://doi.org/10.1101/2022.06.09.495447).
- [4] A. Rosa and R. Everaers, *PLoS Comp. Biol.* **4**, e1000153 (2008).
- [5] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, *Proc. Natl. Acad. Sci. USA* **109**, 16173 (2012).
- [6] D. Jost, P. Carrivain, G. Cavalli, and C. Vaillant, *Nucleic Acids Res.* **42**, 9553 (2014).
- [7] M. Di Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **113**, 12168 (2016).
- [8] A. M. Chiariello, C. Annunziatella, S. Bianco, A. Esposito, and M. Nicodemi, *Sci. Rep.* **6**, 29775 (2016).
- [9] C. A. Brackley, D. Marenduzzo, and N. Gilbert, *Nat. Methods* **17**, 767 (2020).
- [10] M. Chiang, G. Forte, N. Gilbert, D. Marenduzzo, and C. A. Brackley, *Methods Mol. Biol.* **2301**, 267 (2022).
- [11] M. Chiang, D. Michieletto, C. A. Brackley, N. Rattanaivrotkul, H. Mohammed, D. Marenduzzo, and T. Chandra, *Cell Rep.* **28**, 3212 (2019).
- [12] D. Marenduzzo and E. Orlandini, *J. Stat. Mech.* (2009) L09002.
- [13] C. A. Brackley, J. Johnson, S. Kelly, P. R. Cook, and D. Marenduzzo, *Nucleic Acids Res.* **44**, 3503 (2016).
- [14] C. Brackley, N. Gilbert, D. Michieletto, A. Papantonis, M. Pereira, P. Cook, and D. Marenduzzo, *Nat. Commun.* **12**, 1 (2021).
- [15] P. R. Cook and D. Marenduzzo, *Nucleic Acids Res.* **46**, 9895 (2018).
- [16] B. Duplantier, *J. Stat. Phys.* **54**, 581 (1989).
- [17] A. Bonato, M. Chiang, D. Corbett, S. Kitaev, D. Marenduzzo, A. Morozov, and E. Orlandini, Companion paper, *Phys. Rev. Lett.* **132**, 248403 (2024).
- [18] OEIS Foundation Inc., “The On-Line Encyclopedia of Integer Sequences” (2023), published electronically at <http://oeis.org>.
- [19] L. Lovász, *Combinatorial Problems and Exercises* (Elsevier, Amsterdam, 1993).
- [20] B. C. Rennie and A. J. Dobson, *J. Comb. Theory* **7**, 116 (1969).
- [21] M. Bóna and I. Mező, *Eur. J. Comb.* **51**, 500 (2016).
- [22] F. Beyene and R. Mantaci, *arXiv:2101.07081*.
- [23] O. Nabawanda, F. Rakotondrajao, and A. S. Bamunoba, *arXiv:2007.03821*.
- [24] The upper bound comes from requiring the degree of the internal vertices to be greater than or equal to 3. Since it is tied to traversability, the maximum number of vertices can be lower than this upper bound.
- [25] R. J. Trudeau, *Introduction to Graph Theory* (Dover Publications, New York, 2013).
- [26] Note that, for even c , we *apparently* neglected $n_1 = 0, n_2 = 0$. This is because
- $$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 2 \\ 0 & b & c \\ 2 & c & a \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2 & 0 \\ 2 & a & c \\ 0 & c & b \end{pmatrix}.$$
- [27] ($a = 0, b = n - 1, c = 1$) has already been removed, since $a < b$.
- [28] C. A. Brackley, S. Taylor, A. Papantonis, P. R. Cook, and D. Marenduzzo, *Proc. Natl. Acad. Sci. USA* **110**, E3605 (2013).
- [29] A. Marin-Gonzalez, J. G. Vilhena, F. Moreno-Herrero, and R. Perez, *Phys. Rev. Lett.* **122**, 048102 (2019).
- [30] P. R. Cook and D. Marenduzzo, *J. Cell Biol.* **186**, 825 (2009).
- [31] J. N. Onuchic and P. G. Wolynes, *Curr. Opin. Struct. Biol.* **14**, 70 (2004).
- [32] B. Duplantier, *Phys. Rev. Lett.* **57**, 941 (1986).
- [33] S. Plimpton, *J. Comput. Phys.* **117**, 1 (1995).
- [34] P.-G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).
- [35] G. A. Dotson, C. Chen, S. Lindsly, A. Cicalo, S. Dilworth, C. Ryan, S. Jeyarajan, W. Meixner, C. Stansbury, J. Pickard *et al.*, *Nat. Commun.* **13**, 5498 (2022).