Nano Research
2022, 15(11): 9809–9817

ISSN 1998-0124 CN 11-5974/O4
https://doi.org/10.1007/s12274-022-4355-x

Research Article

A topology framework for macromolecular complexes and condensates

Maziar Heidari[1], Duane Moes[1], Otto Schullian[1,2], Barbara Scalvini[1], and Alireza Mashaghi[1,3] (✉)

[1] *Medical Systems Biophysics and Bioengineering, Leiden Academic Centre for Drug Research, Faculty of Science, Leiden University, Leiden 2333CC, the Netherlands*
[2] *Fachbereich Physik, Freie Universität Berlin, Arnimallee 14, Berlin 14195, Germany*
[3] *Harvard Medical School, Harvard University, 25 Shattuck St, Boston MA 02115, USA*

Received: 2 February 2022 / **Revised:** 21 March 2022 / **Accepted:** 22 March 2022

ABSTRACT

Macromolecular assemblies such as protein complexes and protein/RNA condensates are involved in most fundamental cellular processes. The arrangement of subunits within these nano-assemblies is critical for their biological function and is determined by the topology of physical contacts within and between the subunits forming the complex. Describing the spatial arrangement of these interactions is of central importance to understand their functional and stability consequences. In this concept article, we propose a circuit topology-based formalism to define the topology of a complex consisting of linear polymeric chains with inter- and intrachain interactions. We apply our method to a system of model polymer chains as well as protein assemblies. We show that circuit topology can categorize different forms of chain assemblies. Our multi-chain circuit topology should aid analysis and predictions of mechanistic and evolutionary principles in the design of macromolecular assemblies.

KEYWORDS

topology, macromolecular complex, protein evolution, folding

1 Introduction

Molecular complexes formed between two or more molecules are ubiquitous in nature. Cells contain a large number of macromolecular complexes with specific functionality, ranging from binary ligand/receptor pairs to large multiprotein and protein/nucleic acid complexes (e.g., RNA/DNA polymerases and ribosomes, respectively). Biomolecular nano assemblies are not limited to these molecular complexes. Proteins, peptides, and RNA molecules can dynamically assemble to form liquid-phase nano- and micro-scale condensates such as P-bodies and stress granules in the cytoplasm or nucleoli and Cajal bodies in the nucleus [1–2]. These complexes and condensates form by molecular contacts established within the constituting chains and between different chains. Complexes can then disassemble or rearrange by breaking the existing contacts and forming others with a new inter-contact arrangement.

Whereas a number of widely used frameworks are available for single molecules, or subunits, the literature on topological description of molecular complexes is scarce. Knot theory [3] and circuit topology (CT) [4–5] are the two main topological frameworks with applicability to fold analysis of single linear chains. While knot theory is a mature field with a plethora of demonstrated applications, its applicability to biopolymers is limited, as these molecules are typically unknotted. Moreover, knot theory ignores contacts within molecules. Circuit topology, a notion that has only recently been introduced, informs not only on molecular structure, but also can help predict dynamics during folding of biomolecules [6–8]. The circuit topology framework is based on monitoring the sequential arrangement of intramolecular contacts within linear chains. The arrangement of contacts in a polypeptide or polyribonucleotide is a topological property, because it is conserved through continuous deformations of the chain. Stretching or compressing the chain leaves the circuit topology invariant. Only rupturing existing and/or forming new contacts can alter the spatial arrangement of interactions and thus contact topology. These topological changes typically have profound biological consequences [9].

The definition of frameworks for the topology of molecular complexes has been addressed in a number of studies [10–12]. Teichmann et al. proposed a hierarchical classification of all protein complexes with known three-dimensional (3D) structure, based on a graph-based representation of their main structural features [10]. The authors systematically describe the quaternary structure of protein complexes in terms of topological parameters of the graph describing the assembly of their subunits and found that they cluster within a small number of arrangements compared with all theoretically possible ones. This approach can serve as a classification scheme and analysis tool for all protein complexes in the Protein Data Bank (PDB). Such coarse-grained topological descriptions are however ignorant of the spatial contact pattern in the complexes and treat entire protein subunits as a node in a graph. This is in contrast with other current topological frameworks used for single proteins or RNA molecules, such as circuit topology. In their dependence on information about the spatial arrangement of contacts, these

Address correspondence to a.mashaghi.tabari@lacdr.leidenuniv.nl

approaches are becoming increasingly feasible for the study of molecular complexes because high-resolution 3D structures of such complexes are becoming available.

We propose to extend the notion of circuit topology, originally developed as a single-molecule concept, to macromolecular complexes of two or more interacting linear chains. Such linear chains (e.g., polypeptides or polyribonucleotides) can form contacts, both within the chain and between two or more chains. Our discussion does not assume any particular form of contact and the approach can be applied to an arbitrary definition of contacts such as non-covalent residue-residue contacts, disulfide bonds, β-β contacts in proteins, base-pairing in nucleic acids, and contact defined based on distance threshold. As such, this definition is general in the sense that it encompasses various contact types and their associated contact topology. Interactions in a molecular complex can be divided into either intra-chain or inter-chain contacts and can take a number of different forms in their mutual arrangements. Importantly, a given contact from a biopolymer chain such as a polypeptide or polyribonucleotide, is either binary or can often be effectively reduced to multiple binary contacts as far as topology is concerned. In the following we will thus only consider binary contacts. The assignment of contact topologies to macromolecular complexes can be informed by atomic models of these complexes from experiment (deposited in the Protein Data Bank) [13], from homology modeling [14], or from cross-linking mass spectrometry in combination with atomic models of individual subunits [15–17]. Once these contact topologies have been identified, even intricate changes in intermolecular connectivity can be described in more simple terms by their associated changes in topological parameters. In what follows, we provide a simple algorithm for extension of circuit topology to multi-chain systems and show how the toolbox can be applied to the fields of materials science and biomolecular sciences.

2 Methodology

2.1 Generalisation of circuit topology to multi-chains

In this section we introduce the formalism to describe contacts for multiple chains which is consistent with single-chain circuit topology: If the system consists of only one chain classical circuit topology is recovered. Single-chain CT represents contacts as intervals $[a,b]$, with $a<b$, on the real numbers. Two contacts can either be in parallel, series, or cross relation. The definition is complete, meaning that any two contacts have one, and only one of the relations above [4].

In the case of multiple chains, we define contacts in a similar way. Assume there are n chains that have a direction (for example, N-terminus to C-terminus in proteins). We consider the n-dimensional vector space $\mathbb{R}^n$ where each of the chains is placed on a separate axis such that the direction of the chain is in the positive direction of the axis. A contact is represented by the straight line between the chains (or axes). Figure S1 in the Electronic Supplementary Material (ESM) illustrates this procedure for a contact between position a of chain A and position b of chain B (A and B can be identical).

The straight line is parametrized by $t\in[0,1]$and given by

$$[a_{\mathrm{A}},b_{\mathrm{B}}]=\left\{\begin{pmatrix}a\\0\\\vdots\\0\end{pmatrix}(1-t)+\begin{pmatrix}0\\b\\\vdots\\0\end{pmatrix}t\,|\,t\in[0,1]\right\}\qquad(1)$$

where $\begin{pmatrix}a\\0\\\vdots\\0\end{pmatrix}$ is the vector containing only zeros, except in the A-th position where it contains the value a (similar for the vector $\begin{pmatrix}0\\b\\\vdots\\0\end{pmatrix}$). $[a_{\mathrm{A}},b_{\mathrm{B}}]$ is the set containing a straight segment from location a on chain A to location b on chain B with the requirement of $a,b>0$. In contrast to CT for single chains, contacts are not in 1-dimensional space. However, the classical CT definition of the relations for contacts can be used without any alteration: Let A, B, C, and D positive integers labelling the chains (they need not be different) and $a,b,c,d\in\mathbb{R}$ positions on the chains A, B, C, and D respectively. We define

- parallel relation: $[a_{\mathrm{A}},b_{\mathrm{B}}]\subset[c_{\mathrm{C}},d_{\mathrm{D}}]$ or $[a_{\mathrm{A}},b_{\mathrm{B}}]\supset[c_{\mathrm{C}},d_{\mathrm{D}}]$
- series relation: $[a_{\mathrm{A}},b_{\mathrm{B}}]\cap[c_{\mathrm{C}},d_{\mathrm{D}}]=\varnothing$
- cross relation: $[a_{\mathrm{A}},b_{\mathrm{B}}]\not\subset[c_{\mathrm{C}},d_{\mathrm{D}}]$, $[a_{\mathrm{A}},b_{\mathrm{B}}]\not\supset[c_{\mathrm{C}},d_{\mathrm{D}}]$ and $[a_{\mathrm{A}},b_{\mathrm{B}}]\cap[c_{\mathrm{C}},d_{\mathrm{D}}]\neq\varnothing$

This definition is invariant to the order of the chains on the axis. The proof can be shown in the following manner: The reordering of the chains is represented by the permutation $\sigma:A\mapsto\sigma(A)$, where σ is an element of the symmetric group of degree n and the transformation of the path is given by $[a_{\sigma(\mathrm{A})},b_{\sigma(\mathrm{B})}]=\{O(\sigma)x|x\in[a_{\mathrm{A}},b_{\mathrm{B}}]\}$, where $O(\sigma)$ is the matrix representation of σ. We show that $[a_{\mathrm{A}},b_{\mathrm{B}}]\subset[c_{\mathrm{C}},d_{\mathrm{D}}]\iff[a_{\sigma(\mathrm{A})},b_{\sigma(\mathrm{B})}]\subset[c_{\sigma(\mathrm{C})},d_{\sigma(\mathrm{D})}]$. For the right implication choose any $y\in[a_{\sigma(\mathrm{A})},b_{\sigma(\mathrm{B})}]$. Then there exists an $x\in[a_{\mathrm{A}},b_{\mathrm{B}}]$ such that $y=O(\sigma)x$. Because of the parallel relation, it follows $x\in[c_{\mathrm{C}},d_{\mathrm{D}}]$ and therefore $y=O(\sigma)x\in[c_{\sigma(\mathrm{C})},d_{\sigma(\mathrm{D})}]$. To show the left implication we assume $[a_{\sigma(\mathrm{A})},b_{\sigma(\mathrm{B})}]\subset[c_{\sigma(\mathrm{C})},d_{\sigma(\mathrm{D})}]$ to be true. For any $x\in[a_{\mathrm{A}},b_{\mathrm{B}}]$, it follows that $O(\sigma)x\in[a_{\sigma(\mathrm{A})},b_{\sigma(\mathrm{B})}]$ and therefore $O(\sigma)x\in[c_{\sigma(\mathrm{C})},d_{\sigma(\mathrm{D})}]$. This however means that $x\in[c_{\mathrm{C}},d_{\mathrm{D}}]$, which completes the proof. Two contacts that are in parallel relation for any order of the chains. Similarly, it can be shown that the other relations (series and cross) are invariant under reordering of the chains.

The proofs for completeness and topology rules in Ref. [4] only require set theory and can be applied identically to this new definition. Hence, the definition given here for multiple chains is complete: Any two contacts must have one of the above relations and it must be unique. In addition, the following topology rules, also named chain rules (box 1 in Mashaghi et al. [4]) are true in the case of multiple chains.

Interestingly, for a single chain the vector space reduces to the real numbers and a contact is represented by an interval on the real numbers, i.e., classical circuit topology is recovered without any further restrictions or assumptions. Therefore, the new definition for multi-chain topology presented here is truly a generalization of the classical circuit topology.

Multiple chains increase the possible configurations many times over, so in addition to the parallel, series, and cross relation, an additional classification is proposed: If the contact can be connected to a closed loop by travelling only along the chains or contacts without visiting any section twice or the origin of the vector space $\mathbb{R}^n$, we label the state with L (for "loop"). If the two contacts can be connected by such a path, without it being closed then with a T (for "tandem") and in all other cases with a I (for "independent"). Figure S2 in the ESM (top) shows a possible path for a L-configuration, whereas Fig. S2 in the ESM (bottom) depicts a T-configuration.

The notation describing the whole state is the CT relation (P, X, or S) and the configuration (L, T, or I) as subscript and the number of chains as the superscript. For example, $\mathrm{S}^2_{\mathrm{L}}$ describes two contacts that are in series, which involve 2 chains and form a loop. Figure 1 shows the all possible configurations for two contacts in multi-chain circuit topology. The procedure to determine the contact relation is the following. First, choose a pair of contacts (for the type of contacts and the protocols for

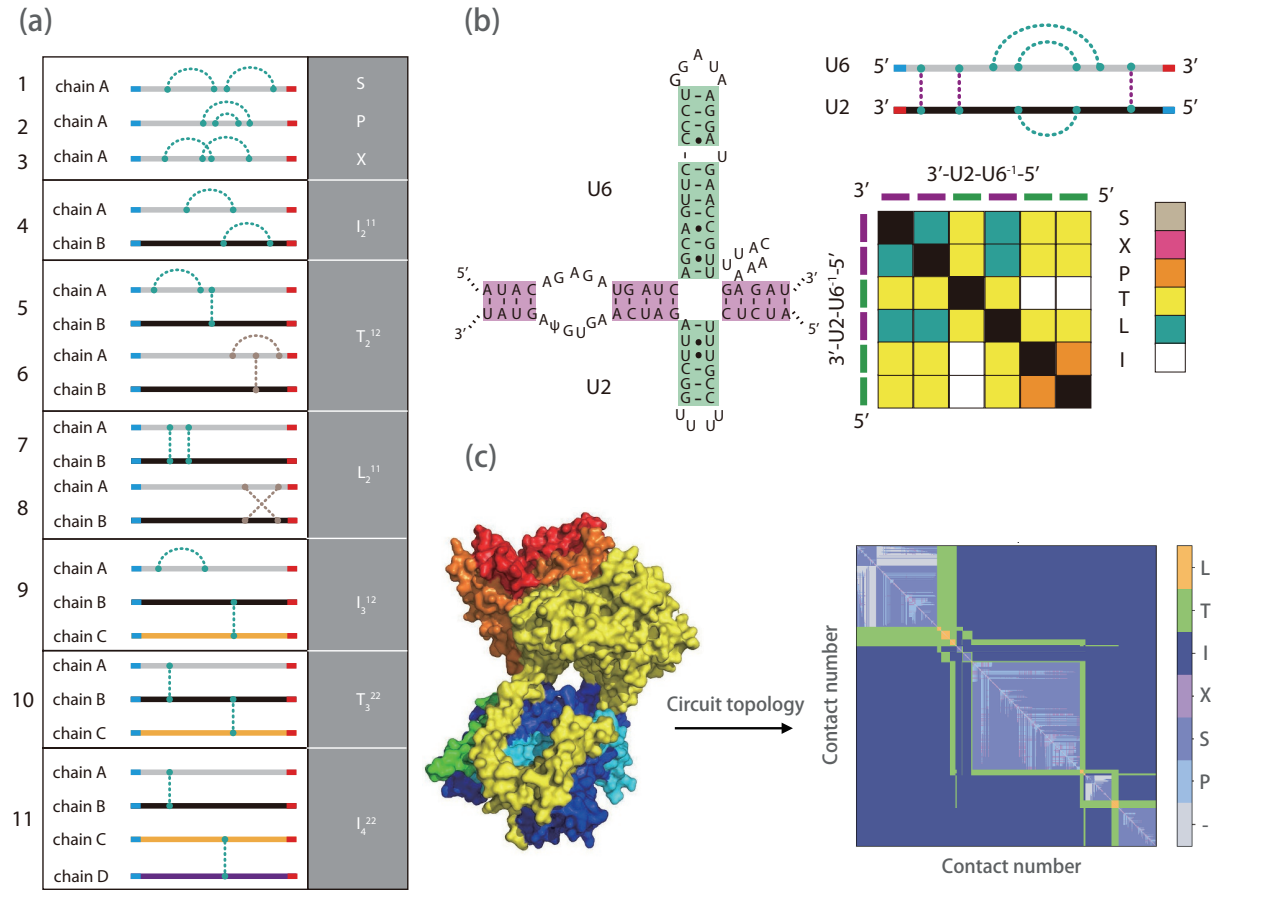

Figure 1 Topology framework for macromolecular complexes (a) circuit topology assignment table. To identify the topology of a complex, we follow a 3-step protocol: (1) choose a pair of binary contacts and their contributing chains; (2) assign a topological relation based on the assignment table; (3) tabulate the assigned relations in the form of a matrix. The topology matrix, with elements taken from (P, S, X, I, T, and L), can serve as a formalized topological description of any given complex. "I" identifies all contact pairs that do not share their chains (I for independent). "L" identifies a contact pair that forms a loop. "T" identifies contact pairs that partly share their chains, but do not form loops. (b) Structure and topology analysis of the spliceosomal U2-U6 complex, with contacts formed by base-pairing of nucleobases. (c) Structure of the yeast Sec complex composed of six different proteins (heterogeneity index $H = 1$) with PDB 6n3q. The corresponding circuit topology matrix is shown on the right side of the protein structure. The complex has 250 inter-protein contacts and 958 intra-protein contacts. The topological fraction of the protein complex is $S = 0.16$, $P = 0.0301$, $X = 0.0134$, $I_2 = 0.5692$, $I_3 = 0.0496$, $I_4 = 0.0793$, $T_2 = 0.0849$, $T_3 = 0.0060$, $L_2 = 0.0021$.

extracting contact information from coordinate files, please see Mashaghi et al. [4]). Second, find the chains that contribute the selected pair of contacts and ignore the remaining chains and contacts in the complex. Up to four chains may be involved in forming a selected pair of contacts (Fig. 1). Finally, determine the contact relation (P, S, or X) and the configuration (L, T, or I) as well as the number of chains.

In the case where the chain does not have a direction, there is no unique way of placing the chain onto the axes. The procedure and notation introduced here can still be used, as flipping only affects one configuration: It changes S_L^2 into X_L^2 and *vice versa*. In this case these two states are indistinguishable. Therefore, we propose to indicate them with S/X_L^2.

For many applications, one may opt to simplify the assignment table to ignore contact relations and include only the configuration types and the number of chains for multi chain systems. For a single chain system, one can reduce the notation to the notation of single-chain circuit topology.

Let us apply the new notation on the spliceosomal U2-U6 complex (Fig. S3(a) in the ESM). First the structure and contacts must be identified, then the direction of the chain needs to be taken into account, flipping each chain such that they point in the same direction (Fig. S3(b) in the ESM). Finally, all relations are determined and represented in a matrix (Fig. S3(c) in the ESM). For the spliceosomal U2-U6 complex, all contacts that connect the two chains are in cross relation which is a direct consequence of the two chains running antiparallelly. There are two contacts that are in parallel configuration and all other relations are in series (either in T- or I-configuration). The topology matrix with simplified notation can also be readily calculated and is given in Fig. 1(b) for the spliceosomal complex. Without losing generality, we will use the simplified notation in the case studies provided in the subsequent section. As another example, the circuit topology matrix of the yeast Sec complex composed of six different proteins is shown in Fig. 1(c). The corresponding circuit topology matrix is shown on the right side of the protein structure. The complex has 250 inter-protein contacts and 958 intra-protein contacts and is enriched in S arrangements.

3 Results and discussions

3.1 Multi-chain circuit topology of polymer models

In the following section, we apply the circuit topology framework on a single-chain and multi-chain system. We use Kremer–Grest model [18] to simulate a coarse-grain model of a polymeric chain of size 90 monomers. We used LAMMPS package to perform the simulations [19]. The interaction potential between the monomers is Lennard–Jones potential whose length and cohesive strength scales are σ and $\tilde{\varepsilon}k_B T$, respectively. Here, $\tilde{\varepsilon}$ is the dimensionless strength, k_B is the Boltzmann constant and T is the temperature. The system is thermalized by Langevin thermostat at temperature corresponding to thermal energy of $k_B T = 1$. The thermostat

damping coefficient is set by 10τ. The mass (m) of all monomers is identical and we define characteristic time scale $\tau = \sqrt{ma^2/k_B T}$. The time step of all simulations is set by 0.01τ and the averaging and standard deviation are calculated out over five different simulation runs, each run has $10^5\tau$ simulation time. The bending energy is introduced by $U_b = \tilde{\kappa} k_B T[1+\cos(\theta)]$, where $\tilde{\kappa}$ is elastic constant and θ is the angle between two successive bonds. Figure 2 shows the gyration radius of a single chain as the function of the cohesive strength. For $\tilde{\kappa} = 0$, the chain transits from coil structure into globular one as the cohesive strength grows. For $\tilde{\varepsilon} \ll 0.2$, the configuration of the chain resembles that of a real chain in good solvent, $R_g = N^{3/5}\sigma/\sqrt{6}$. For the cohesive strength close to $\tilde{\varepsilon} = 0.4$ [20] the chain behaves like θ-polymer and its gyration radius scales similar to an ideal chain $R_g = N^{1/2}\sigma/\sqrt{6}$. For an infinitely rigid chain ($\tilde{\kappa} = \infty$), one can obtain $R_g = N^{1/2}\sigma/\sqrt{12}$. In Fig. 2(b) the topology fractions are shown for fully flexible and elastic chains. For all analysis, two non-bonded monomers are assumed in contact if their distance is less than 1.4σ. As for random coil, the entropic contacts mostly occur locally and independently along the chain, thus the series contact topology is mostly populated compared with the other topologies. There is a decreasing trend in series loops when the chain undergoes coil-globule transition. The reason is that in the globular conformation loops likely collide locally leading to populate the parallel and cross topologies. The circuit topology does not change for $\tilde{\varepsilon} \geq 1.0$ as the chain has already formed globular structure. The topological invariance has been shown for a growing globule as a result of folding chain [21]. However, this is quite different for an elastic chain as the loops are populated mostly in the cross topology. When the elastic energy of the chain dominates the cohesive energy ($\tilde{\varepsilon} < 1.0$), the persistent length of the chain ($l_p \approx \tilde{\kappa}\sigma \approx 10\sigma$) is large enough such that the chain rarely forms any contact. For higher cohesive strength ($1 < \tilde{\varepsilon} < 2$), the single chain forms an oblate collapsed structure (Fig. 2(c)). This can be clarified by the contact probability along the chain as shown in Fig. 2(c). For the coil configuration, the contact probability decays close to an ideal chain contact probability as $P(s) \sim s^{-3/2}$ where s is the contour length distance between the contact pair [22]. However, for a globule chain, the contact probability decays slower with contour length. The globular conformation resembles a fractal globule, whose contact probability decays with $P(s) \sim s^{-1}$ [23]. For the elastic chain, the contact probability increases with the contour length up to 20σ,

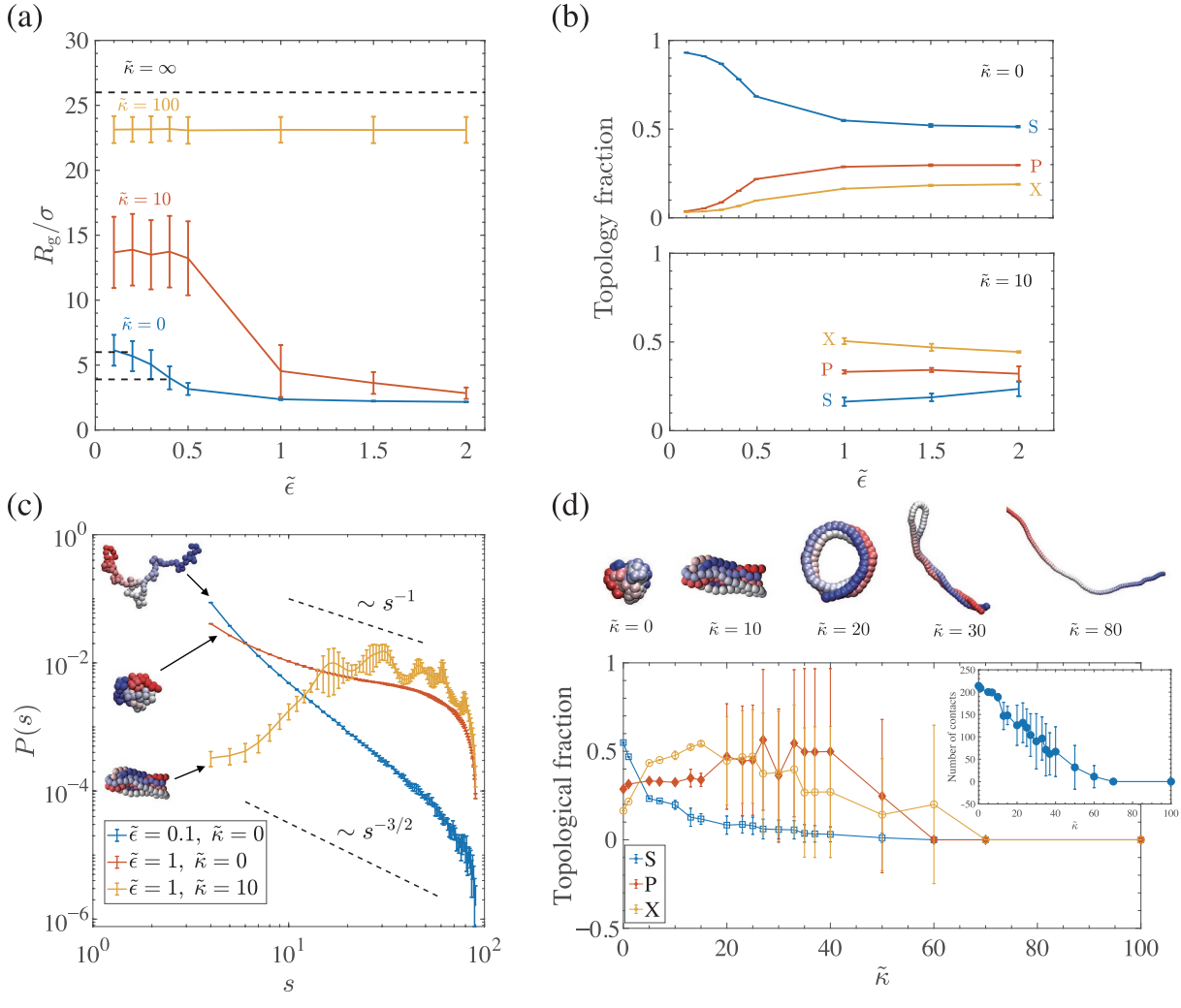

Figure 2 Circuit topology of a single chain. Gyration radius (R_g) of a single chain for different reduced cohesive interaction strength ($\tilde{\varepsilon}$) and bending rigidities ($\tilde{\kappa}$). For an ideal chain and real chain in good solvent, the gyration radius is calculated by $R_g^2 = N\sigma^2/6$ and $N^{6/5}\sigma^2/6$, respectively as shown by dashed lines in (a). The horizontal dashed line represents the gyration radius of a rigid chain ($\tilde{\kappa} = \infty$). The corresponding topology fractions, series (S), parallel (P), and cross (X) of the chains are shown in (b) for two bending rigidities $\tilde{\kappa} = 0, 10$. The topology fraction of the case $\tilde{\kappa} = 10$ is shown for $\tilde{\varepsilon} \geq 1$ since the chain rarely contacts itself below that value. (c) The contact probability for a single chain having different cohesive interaction ($\tilde{\varepsilon}$) and bending rigidities ($\tilde{\kappa}$). The dashed lines represent power-law decaying of contact probability as obtained for a fractal globule ($P(s) \sim s^{-1}$) and for an ideal chain $P(s) \sim s^{-3/2}$. The inset shows the instantaneous conformation of the chains. (d) Circuit topology of a single chain at cohesive strength $\tilde{\varepsilon} = 1$ as a function of different bending stiffness $\tilde{\kappa}$. The snapshots of the configurations at different bending rigidities are shown at the top of (d).

which is the twice of the persistence length ($l_p \approx \tilde{\kappa}\sigma = 10\sigma$) [22] (see the inset of Fig. 2(c)). In Fig. 2(d), we plotted the topology fraction against the bending rigidity. For low bending rigidity $\tilde{\kappa} < 2$, the loops mostly occupy the series topological state as the chain is in the collapsed coiled configuration. For larger bending rigidity, the globule structure turns into oblate globule and the number of loops in cross topology surpasses the series and parallel loops. For larger bending rigidity $2 < \tilde{\kappa} < 20$, the oblate collapsed configuration changes into collapsed ring so as to minimize the bending energy. In the region $20 < \tilde{\kappa} < 40$, in addition to the stable folded ring structure in which cross and parallel loops exceed series loops, there is an unfolded meta-stable state (having no contact) which yields large standard deviations. For $40 < \tilde{\kappa} < 60$, the stable ring structure changes into hairpin structure leading to relative increase in the parallel loops. For $\tilde{\kappa} > 60$, the unfolded structure becomes stable and the chain rarely forms contacts (top row of Fig. 2(d)).

To investigate the circuit topology of a multichain system, we build a model system composed of $M = 100$ chains of length $N = 90$ monomers in a simulation box of size $50\sigma \times 50\sigma \times 50\sigma$. We perform a NVT simulation with similar parameters as those described earlier for the single chain system. The system is initially equilibrated for $10^5\tau$ at low cohesive and zero bending rigidity ($\tilde{\varepsilon} = 0.1, \tilde{\kappa} = 0$). Then the simulation is continued for $10^5\tau$. The configuration of the system is sampled every $1,000\tau$ and the topology fractions are computed over the last 10 sampled configurations. Figures 3(a)–3(d) show the snapshots of the chain system having different cohesive strength and bending rigidities. For fully flexible chains, and low cohesive strength ($\tilde{\varepsilon} = 0.1, \tilde{\kappa} = 0$), the chains have coil configurations and they are uniformly distributed in the system. The monomer concentration of the chains is $\phi = 0.072$ which is comparable to the overlapping concentration $\phi^\star = N/R_F^3 = 0.023$, where $R_F = N^{3/5}\sigma \approx 15\sigma$. In this regime, independent contacts are mostly formed within the chains and the chain-chain collisions are not frequent. Thus, similar to CT of a single chain (Fig. 2(b)), the number of loops in series topology state is larger than parallel and cross and it scales with number of the chains in the system, i.e., $\#S \sim M$. However, the number of loops in the topology state I_2 scales with $\#S \sim M^2$ leading it to be the mostly populated state. By increasing the cohesive strength, the chains collapse and form a large globule (Fig. 3(b)). In this regime, the collision of the chains enhances the topology fractions I_3 and I_4 which are the result of inter-chain contacts. By increasing the bending rigidity ($\tilde{\varepsilon} = 0.1, \tilde{\kappa} = 100$), the chains exhibit elastic spherocylinders configurations and they are randomly oriented in the system. The number density of the chains inside the system is $n = N/V = 8 \times 10^{-4}$. Given the aspect ratio of a single spherocylinder $\frac{D}{L} \sim 10^{-2}$ and the second virial coefficient of a suspension of rods $b = \pi L^2 D/4$, the reduced concentration is obtained by $c = bn \cong 5$. Although the reduced concentration is above the isotropic-nematic phase coexistence concentrations of hard rods ($c_{\text{isotropic}} \cong 3$ and $c_{\text{nematic}} \cong 4$) [24, 25], we have not observed nematic phase. The reason can be explained by the finite elasticity of the chain which potentially changes the

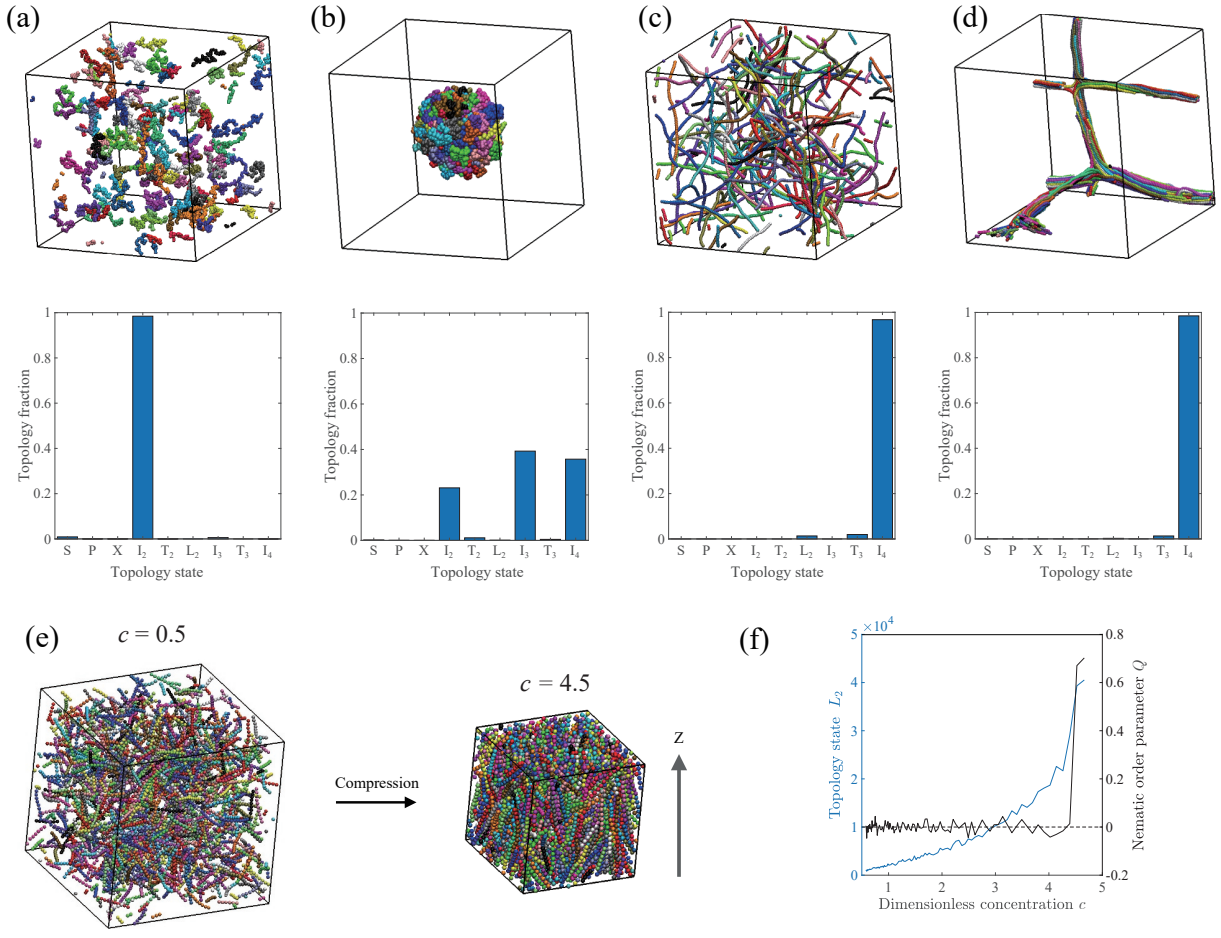

Figure 3 Circuit topology of a multi-chain system. Snapshots of the system and corresponding topological fractions are shown for four systems, sorted with different cohesive strength $\tilde{\varepsilon}$ and bending rigidity $\tilde{\kappa}$. (a) Fully flexible chains with weak cohesive interaction strength, $\tilde{\kappa} = 0$ and $\tilde{\varepsilon} = 0.1$, (b) fully flexible chains with strong cohesive interaction strength, $\tilde{\kappa} = 0$ and $\tilde{\varepsilon} = 0.5$, (c) elastic chains with weak cohesive interaction strength, $\tilde{\kappa} = 100$ and $\tilde{\varepsilon} = 0.1$, and (d) elastic chains with strong cohesive interaction strength, $\tilde{\kappa} = 100$ and $\tilde{\varepsilon} = 0.5$. The size of the system in (a)–(d) is $50\sigma \times 50 \times 50$ and it is composed of 100 chains each having 90 monomers. (e) Compression of a system composed of 900 chains, each having 10 monomers from dimensionless concentration c = 0.5 to 4.5. (f) The topology fraction L_2 and nematic order parameter Q as a function of dimensionless concentration c.

phase coexistence concentrations as it was similarly explained for compressible rods [26]. To examine the effect of isotropic-nematic transition on the circuit topology, we fragmented the length of the chain into 10 monomers which are much shorter than the persistence length ($l \ll l_p = 100\sigma$). Thus, the effect of finite elasticity is negligible and we can assume the chains are rigid rods. We build a system with 900 chains which are initially equilibrated in a cubic box of size 50σ. The size of the cubic box changes from 50σ to 25σ. Figure 3(c) shows two snapshots of the system at isotropic and nematic phases. Since the persistence length of the chain is larger than the chain length, chains rarely form intra-chain contacts. Accordingly, the topology states of S, P, and X are not occupied. Figure 3(d) shows the L_2 loops which are monotonically increasing with the system concentration. To quantify the nematic order of the system, the nematic order parameter is calculated by $Q = 3\cos^2\theta - 1/2$, where θ is the angle between the chains vector (defined by the end-to-end vector) and the system global director. As shown in Fig. 3(d), for the isotropic phase, the nematic director is zero and it abruptly rises for concentrations close to $c = 4.5$. The topological states L_2, T_3, and I_4 monotonically increase with system concentration (Fig. 3(f) and Fig. S4 in ESM). However, only L_2 exhibits an abrupt change with nematic transitions. This is due to the fact that the topology state L_2 can capture the local reordering of the chains at the isotropic-nematic transition.

For elastic chains with high cohesive strength (Fig. 3(d)), the chains percolate along the system. This leads to enhancement of independent loop formation along the chain bundles and accordingly to the increase in topology fraction I_4.

3.2 Circuit topology of protein complexes

Next, we developed an algorithm to determine the circuit topology of protein complexes and condensates. For this aim, we used the most comprehensive database containing stoichiometric and functional information about protein complexes, i.e., the manually curated Complex Portal [27]. In this database, the organism with the most available information on complexes is *Saccharomyces cerevisiae*, with over 616 available protein complexes. Not every available complex contained the full stoichiometric information needed for processing, therefore we limited our dataset to 238 complexes that had all the necessary information. This dataset was then cross-referenced with the structural information provided by the PDB website [28], complexes that did not have a complete PDB entry were removed which distilled the dataset down to 176 protein complexes. Furthermore, we removed 6 of these protein complexes because they either contained faulty structural information or existed solely out of DNA strands with no protein structures. The final 170 protein complexes containing a total of 1,223 proteins were then analyzed.

We developed a python based computational tool that identifies inter- and intra-chain contacts within a given protein complex and identifies their circuit topology arrangement as described in Fig. 1(a). Contacts were retrieved from PDB structures, by defining a spatial cutoff for atom-atom distance (4.5 Å), and a threshold for the minimum number of atoms to be found in contact below the cutoff equals to 5 atoms in order to consider the two residues in contact. Contacts formed by residues which were 3 or less residues apart in the sequence were excluded from the analysis. Contact indexes were assigned as they appear along the chain, left end to right end.

We used the available stoichiometric information from the Complex Portal to calculate the complex heterogeneity index and related that to complex topology. This index was introduced by Ref. [29]. It can be used as a measure of how many different proteins are inside the complex. The heterogeneity index ($H = N_c/M_c$) is calculated by dividing the number of different proteins in the complex (N_c) by the total number of proteins in the complex (M_c), where a score of $H = 1$ means that all proteins inside the complex are structurally different, and a low score means that there is a form of stoichiometry and symmetry present in the complex.

The average topology fractions of the topological states for all protein complexes are shown in Fig. 4(a). The topology I_2 has the highest fractions compared with the other states. This is result of the complexity of the structure and contacts inside each protein of the complex, which is in turn reflected in the rise of the independent contact pairs. There is a good comparison between the results of the Figs. 4(a) and 3(b), although the polymeric assembly is much simpler than the protein complex. The reason behind the analogy can be attributed to the compactness of the individual chains, which enhances the intra-chain contacts and the proximity of the chains inside the assembly, thus increasing in the inter-chain contact probability. In Fig. 4(b), the distribution of the topological states I_2 and I_4 is plotted (for other topological states, see Fig. S5 in the ESM). These states have the largest fractions among all other states based on the respective categorization of the intra-chain and inter-chain group. While the distribution of I_4 and other topological states in inter-chain group (Fig. S5 in the ESM) is localized below 0.3, the distribution of I_2 is much broader. As the former has the information about the connectivity and stability of the proteins inside the complex, the observed narrow distribution of L_2, T_3, and I_4 might provide a universal topological prescription for the stability of the protein complex. Proving this statement would require further analysis, such as molecular dynamics simulations and calculating the binding free energy between the proteins inside the complex, and is beyond the scope of the current study. Sartori and Leibler showed that one of the criteria for a reliable assembly is the high heterogeneity of the complex [29]. For the protein complex of our study, the average heterogeneity is 0.7 and the distribution of the calculated heterogeneity is in agreement with the previously reported one (Fig. S6 in the ESM). Figures 4(c) and 4(d) show distributions of the topological fractions I_2 and I_4 and the complex heterogeneity. As it is shown, the distribution of the both I_2 and I_4 topological fractions is highly localized at high heterogeneity index ($H > 0.7$). The distributions of other topological fractions and their corresponding heterogeneity are shown in Fig. S7 in the ESM.

4 Concluding remarks

Protein complexes and condensates enclose and carry out profound biological processes [1]. For example, nuclear condensates perform regulatory activity such as regulation of chromosome structure [30, 31] and dynamics, DNA repair [32], transcription [33, 34], and replication [35]. Mechanistic insight into multi-chain arrangement in nano-assemblies is going to become crucial in future understanding of the behavior of condensate systems [36]. Here, we presented a CT framework to characterize protein complex structure based on inter and intra-chain contacts. The topology of intramolecular contacts is central to structure and function of biopolymers. Even in strongly denaturing conditions, the arrangement of long-range intramolecular contacts is similar to its native configuration (Shortle and Ackerman, 2001). Both rate and cooperativity of protein folding are defined by contact topology [6, 37, 38]. It is well recognized that folding rate anti-correlates with contact order and correlates with the percentage of entangled relations (P and X) in single chain biopolymers. Similarly, one should be able to correlate the folding and assembly rates of molecular complexes with their topology. Statistical analysis of the distribution and

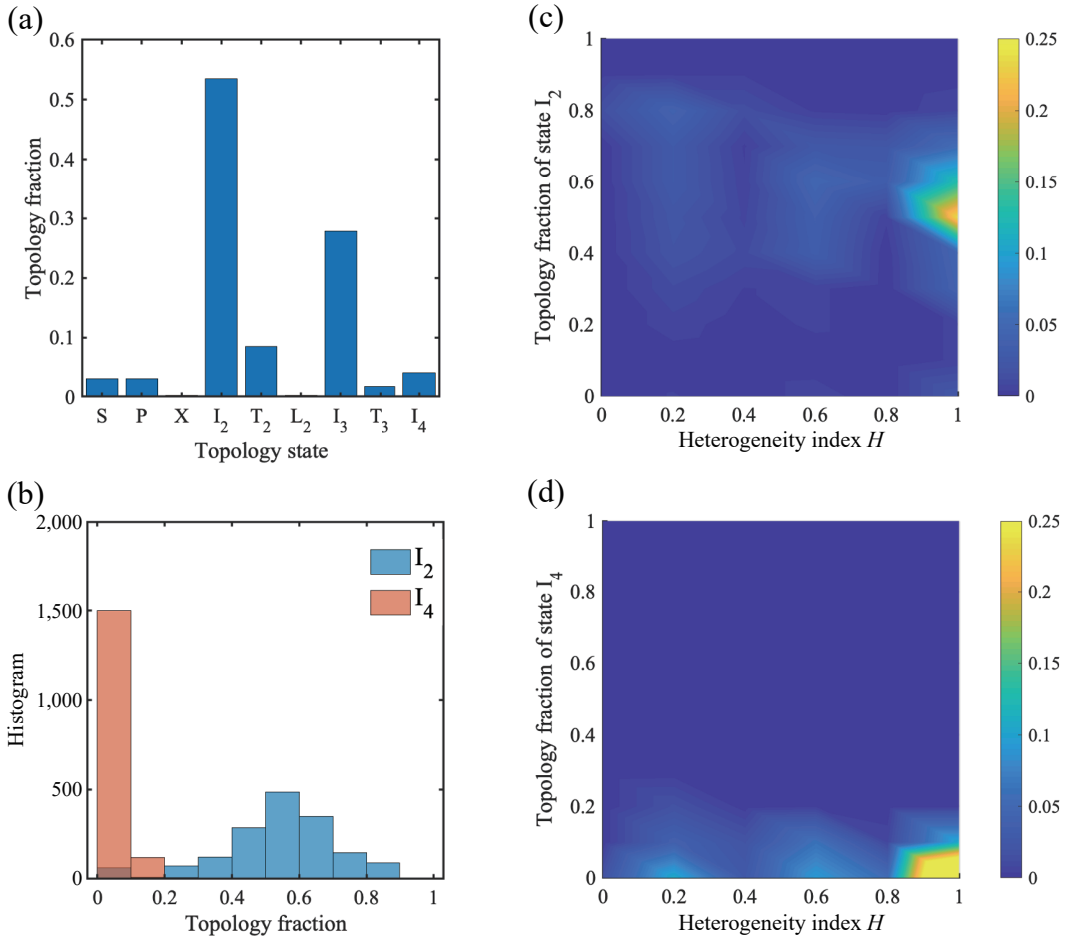

Figure 4 Circuit topology of protein complexes. (a) Averaged topology fractions of the contact pairs for the protein complexes. (b) Histogram of the topology fractions of the topological states I_2 and I_4. The binning width is 0.1. (c) and (d) Bivariate probability distributions of topology fraction and the heterogeneity index of the protein complexes for the topological states I_2 and I_4, respectively. The protein complex set composed of 170 protein complexes (details are given in the text).

frequencies of inter-chain and intra-chain contacts can reveal details on the degree of compactness and entanglement in the complex and might also carry information regarding stability of the structure.

A topological description of macromolecular complexes provides a means for classification of protein interaction interfaces and the identification of interface similarity. As such it can serve as a valuable tool for a range of functional and evolutionary studies. The relation between molecular topology and molecular function is being increasingly appreciated [4, 39, 40]. Additionally, proteins may undergo topological changes to function [41]. Evidences exist for emergence of new protein topologies via gradual evolution through multistep gene rearrangements [42, 43]. These studies were focused on single chains. However, lessons can be learned from these studies with potential application to molecular complexes. Importantly, the framework of contact topology in molecular complexes is defined here such that the relationship of a particular topology and the physical consequence of the associated arrangement of binary contacts in a single chain is preserved also for a related topology of binary inter-chain contact pairs in a macromolecular complex. The framework discussed above offers the possibility to identify and study topological transitions related to evolution and function of macromolecular complexes in more detail. The circuit topology approach is complementary to other topology approaches including network theory, which has been successfully applied to protein complexes previously (see in the ESM).

The repertoire of identified and structurally characterized biomolecular complexes is expanding rapidly [44, 45]. High resolution mapping of many protein–protein, protein–RNA, and RNA–RNA interactions will become a reality in the coming decades. The challenge will then be to relate the arrangement of these interactions to biological function. Standard geometrical descriptions of structures and structural transitions may reach a level of complexity that may obscure mechanistic analysis. It is here where topological descriptions are strongest: Topology studies shed light on the design principles of molecular complexes and provide a simplified framework for relating structure and function of biomolecular complexes. In this paper, we discussed circuit topology of multi chain systems by means of some illustrative examples and large-scale analyses are needed to reveal the utility of the topology approaches to structural biology and protein/RNA evolution. However, the potential of such approach is apparent. Circuit topology can not only provide statistical information about contact arrangement, but also identify the corresponding motifs of protein–protein interaction. For example, the detection of topologies such as L_2 and T_3 might facilitate the identification of functional hubs inside protein complexes. This feature is particularly crucial since protein condensates are not homogenous bodies, and their inner structure allows them to carry out several dynamic processes [46]. The prevalence of certain topological fractions, such as L_2 and T_3, might facilitate the formation of tightly connected sub-clusters of proteins, the characterization of which represents a challenge in the field of protein networks [47, 48].

In addition to structural investigations, circuit topology analysis can be applied to study dynamics. For example, the application of the molecular dynamics simulations in the future studies would

allow one to probe dynamical contact between proteins; this information is obviously lacking in static crystal structure of proteins and thus in static CT analysis. Dynamics is particularly important in the case of condensate of intrinsically disordered proteins. Investigating the dynamic network of contacts between different amino acids can be enabled by large scale molecular dynamics simulations as well as experimental approaches [49]. In this way, one may quantify correlations between topology and the kinetics of folding and assembly. The link between fold topology and kinetics has been successfully demonstrated for single chains and is expected to work for multi-chain systems as well. Topology would then provide a predictive measure for kinetics of complex formation.

The formalism presented here was extended to describe the relationship of two contact involving multiple chains, so, intuitively, the next step will be relations of higher order contacts (i.e., more than two) involving single and multiple chains. Such studies (extension of the formalism and, more importantly, large-scale analysis) further promise new insights and strategies that molecular engineers can exploit for synthesis of molecular origami, new materials, and supramolecular complexes with desired functionalities [50–52].

Acknowledgements

The authors thank Martin Karplus (Harvard University), Arjen Jakobi (Delft University of Technology), and Alexandre Dawid (University Grenoble Alpes) for helpful discussions and critical reading of the manuscript.

Electronic Supplementary Material: Supplementary material (including supplemental text and figures) is available in the online version of this article at https://doi.org/10.1007/s12274-022-4355-x.

The protein complex circuit topology software is available through the following link: https://github.com/circuittopology/circuit_topology.

References

[1] Alberti, S.; Hyman, A. A. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 196–213.

[2] Marx, V. Cell biology befriends soft matter physics. *Nat. Methods* **2020**, *17*, 567–570.

[3] Virnau, P. ; Mirny, L. A. ; Kardar, M. Intricate knots in proteins: Function and evolution. *PLoS Comput. Biol.* **2006**, *2* e122.

[4] Mashaghi, A.; van Wijk, R. J.; Tans, S. J. Circuit topology of proteins and nucleic acids. *Structure* **2014**, *22*, 1227–1237.

[5] Golovnev, A.; Mashaghi, A. Generalized circuit topology of folded linear chains. *iScience* **2020**, *23*, 101492.

[6] Mugler, A.; Tans, S. J.; Mashaghi, A. Circuit topology of self-interacting chains: Implications for folding and unfolding dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 22537–22544.

[7] Scalvini, B.; Sheikhhassani, V.; Mashaghi, A. Topological principles of protein folding. *Phys. Chem. Chem. Phys.* **2021**, *23*, 21316–21328.

[8] Heidari, M.; Schiessel, H.; Mashaghi, A. Circuit topology analysis of polymer folding reactions. *ACS Cent. Sci.* **2020**, *6*, 839–847.

[9] Schullian, O.; Woodard, J.; Tirandaz, A.; Mashaghi, A. A circuit topology approach to categorizing changes in biomolecular structure. *Front. Phys.* **2020**, *8*, 5.

[10] Levy, E. D.; Pereira-Leal, J. B.; Chothia, C.; Teichmann, S. A. 3D complex: A structural classification of protein complexes. *PLoS Comput. Biol.* **2006**, *2*, e155.

[11] Ozawa, Y.; Saito, R.; Fujimori, S.; Kashima, H.; Ishizaka, M.; Yanagawa, H.; Miyamoto-Sato, E.; Tomita, M. Protein complex prediction via verifying and reconstructing the topology of domain–domain interactions. *BMC Bioinformatics* **2010**, *11*, 350.

[12] Mashaghi, A. R.; Ramezanpour, A.; Karimipour, V. Investigation of a protein complex network. *Eur. Phys. J. B* **2004**, *41*, 113–121.

[13] Gutmanas, A.; Alhroub, Y.; Battle, G. M.; Berrisford, J. M.; Bochet, E.; Conroy, M. J.; Dana, J. M.; Montecelo, M. A. F.; van Ginkel, G. et al. PDBe: Protein data bank in Europe. *Nucleic Acids Res.* **2014**, *42*, D285–D291.

[14] Kopp, J.; Schwede, T. The SWISS-MODEL Repository: New features and functionalities. *Nucleic Acids Res.* **2006**, *34*, D315–D318.

[15] Bruce, J. E. *In vivo* protein complex topologies: Sights through a cross-linking lens. *Proteomics* **2012**, *12*, 1565–1575.

[16] Politis, A.; Schmidt, C.; Tjioe, E.; Sandercock, A. M.; Lasker, K.; Gordiyenko, Y.; Russel, D.; Sali, A.; Robinson, C. V. Topological models of heteromeric protein assemblies from mass spectrometry: Application to the yeast eIF3:eIF5 complex. *Chem. Biol.* **2015**, *22*, 117–128.

[17] Zheng, C. X.; Yang, L.; Hoopmann, M. R.; Eng, J. K.; Tang, X. T.; Weisbrod, C. R.; Bruce, J. E. Cross-linking measurements of *in vivo* protein complex topologies. *Mol. Cell. Proteomics* **2011**, *10*, M110.006841.

[18] Kremer, K.; Grest, G. S. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chem. Phys.* **1990**, *92*, 5057–5086.

[19] Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

[20] Alexander-Katz, A.; Schneider, M. F.; Schneider, S. W.; Wixforth, A.; Netz, R. R. Shear-flow-induced unfolding of polymeric globules. *Phys. Rev. Lett.* **2006**, *97*, 138101.

[21] Heidari, M.; Satarifard, V.; Mashaghi, A. Mapping a single-molecule folding process onto a topological space. *Phys. Chem. Chem. Phys.* **2019**, *21*, 20338–20345.

[22] de Gennes, P. G. *Scaling Concepts in Polymer Physics*; Cornell University Press: Ithaca, 1979.

[23] Mirny, L. A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **2011**, *19*, 37–51.

[24] Onsager, L. The effects of shape on the interaction of colloidal particles. *Ann. N. Y. Acad. Sci.* **1949**, *51*, 627–659.

[25] Bolhuis, P.; Frenkel, D. Tracing the phase boundaries of hard spherocylinders. *J. Chem. Phys.* **1997**, *106*, 666–687.

[26] Shundyak, K.; van Roij, R. van der Schoot, P. Theory of the isotropic-nematic transition in dispersions of compressible rods. *Phys. Rev. E* **2006**, *74*, 021710.

[27] Meldal, B. H. M.; Bye-A-jee, H.; Gajdoš, L.; Hammerová, Z.; Horáčková, A.; Melicher, F.; Perfetto, L.; Pokorný, D.; Lopez, M. R.; Türková, A. et al. Complex Portal 2018: Extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* **2019**, *47*, D550–D558.

[28] RCSB PDB. *RCSB PDB* [Online]. https://www.rcsb.org/search (accessed Feb 1, 2022).

[29] Sartori, P.; Leibler, S. Lessons from equilibrium statistical physics

regarding the assembly of protein complexes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 114–120.
[30] Larson, A. G.; Elnatan, D.; Keenen, M. M.; Trnka, M. J.; Johnston, J. B.; Burlingame, A. L.; Agard, D. A.; Redding, S.; Narlikar, G. J. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **2017**, *547*, 236–240.
[31] Gibson, B. A.; Doolittle, L. K.; Schneider, M. W. G.; Jensen, L. E.; Gamarra, N.; Henry, L.; Gerlich, D. W.; Redding, S.; Rosen, M. K. Organization of chromatin by intrinsic and regulated phase separation. *Cell* **2019**, *179*, 470–484.
[32] Kilic, S.; Lezaja, A.; Gatti, M.; Bianco, E.; Michelena, J.; Imhof, R.; Altmeyer, M. Phase separation of 53BP1 determines liquid-like behavior of DNA repair compartments. *EMBO J.* **2019**, *38*, e101379.
[33] Chong, S. S.; Dugast-Darzacq, C.; Liu, Z.; Dong, P.; Dailey, G. M.; Cattoglio, C.; Heckert, A.; Banala, S.; Lavis, L.; Darzacq, X. et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **2018**, *361*, eaar2555.
[34] Cho, W. K.; Spille, J. H.; Hecht, M.; Lee, C.; Li, C.; Grube, V.; Cisse, I. I. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **2018**, *361*, 412–415.
[35] Parker, W. M.; Bell, M.; Mir, M.; Kao, J. A.; Darzacq, X.; Botchan, M. R.; Berger, J. M. A new class of disordered elements controls DNA replication through initiator self-assembly. *eLife* **2019**, *8*, e48562.
[36] Scholl, D.; Deniz, A. A. Conformational freedom and topological confinement of proteins in biomolecular condensates. *J. Mol. Biol.* **2022**, *434*, 167348.
[37] Shank, E. A. ; Cecconi, C. ; Dill, J. W. ; Marqusee, S. ; Bustamante, C. The folding cooperativity of a protein is controlled by its chain topology. *Nature* **2010**, *465*, 637–640.
[38] Baker, D. A surprising simplicity to protein folding. *Nature* **2000**, *405*, 39–42.
[39] Koga, N.; Tatsumi-Koga, R.; Liu, G. H.; Xiao, R.; Acton, T. B.; Montelione, G. T.; Baker, D. Principles for designing ideal protein structures. *Nature* **2012**, *491*, 222–227.
[40] Ong, C. T.; Corces, V. G. CTCF: An architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **2014**, *15*, 234–246.
[41] Rapp, M.; Granseth, E.; Seppälä, S.; von Heijne, G. Identification and evolution of dual-topology membrane proteins. *Nat. Struct. Mol. Biol.* **2006**, *13*, 112–116.
[42] Peisajovich, S. G.; Rockah, L.; Tawfik, D. S. Evolution of new protein topologies through multistep gene rearrangements. *Nat. Genet.* **2006**, *38*, 168–174.
[43] Sato, P. M.; Yoganathan, K.; Jung, J. H.; Peisajovich, S. G. The robustness of a signaling complex to domain rearrangements facilitates network evolution. *PLoS Biol.* **2014**, *12*, e1002012.
[44] de Souza, N. An expanded human interactome. *Nat. Methods* **2015**, *12*, 107.
[45] Rolland, T.; Taşan, M.; Charloteaux, B.; Pevzner, S. J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R. et al. A proteome-scale map of the human interactome network. *Cell* **2014**, *159*, 1212–1226.
[46] Boisvert, F. M.; van Koningsbruggen, S.; Navascués, J.; Lamond, A. I. The multifunctional nucleolus. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 574–585.
[47] Nepusz, T.; Yu, H. Y.; Paccanaro, A. Detecting overlapping protein complexes in protein–protein interaction networks. *Nat. Methods* **2012**, *9*, 471–472.
[48] Spirin, V.; Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12123–12128.
[49] Ganser, L. R.; Myong, S. Methods to study phase-separated condensates and the underlying molecular interactions. *Trends Biochem. Sci.* **2020**, *45*, 1004–1005.
[50] Cao, J.; Gong, H.; Xie, L.; Li, Y.; Zhang, N.; Tian, W.; Zhang, R.; Zhou, J.; Wang, T.; Zhai, Y. et al. Super-assembled carbon nanofibers decorated with dual catalytically active sites as bifunctional oxygen catalysts for rechargeable Zn-air batteries. *Mater. Today Energy* **2021**, *20*, 100682.
[51] Lapenta, F.; Aupič, J.; Strmšek, Ž.; Jerala, R. Coiled coil protein origami: From modular design principles towards biotechnological applications. *Chem. Soc. Rev.* **2018**, *47*, 3530–3542.
[52] Kočar, V.; Schreck, J. S.; Čeru, S.; Gradišar, H.; Bašić, N.; Pisanski, T.; Doye, J. P. K.; Jerala, R. Design principles for rapid folding of knotted DNA nanostructures. *Nat. Commun.* **2016**, *7*, 10803.