ADDENDUM:

SUMMARY

Back in the seventies, nobel winning chemist Linus Pauling predicted the significance of the shape of molecules in regulating their physiological behavior. Nowadays, fifty years later, biopolymer topology is recognized as a key factor influencing biophysical properties such as protein folding rate and pathways as well as binding dynamics. Although the reciprocal influence between structure and function is widely acknowleged, the mechanisms of such interaction remain mostly an open question, far from the depth of understanding the scientific community has gained on the impact of chemical and biological variables. This knowledge gap is partly due to the lack of a unified framework for the topological characterization of biopolymers. Chapter 1 explores a variety of popular applications of topology in the field of folded molecular chains, encompassing protein, RNA, DNA and synthetic polymers. In particular, we explore advantages and limitations of backbone crossing topology (knot theory) and contact-based descriptors – specifically, the relatively new framework of Circuit Topology (CT). Circuit Topology relies on the definition of three fundamental topological relations between pairs of contacts, series (S), parallel (P) and cross (X), defined based on the reciprocal arrangements of their contact sites. Chapter 2 aims at uncovering the role of circuit topology as a folding rate predictor, benchmarking its predictive power against traditional geometric descriptors such as protein size and contact order. We find that CT relations extracted from the native 3D structure of a protein can complement other predictors in providing accurate folding rate estimates and a theoretical insight into the topological principles of protein folding.

However, not all proteins have stable native structures from which to extract topological parameters. In particular, Intrinsically Disordered Proteins (IDPs) present several challenges for their characterization, while being involved in a variety of diseases. In Chapter 3, we extend the theoretical and computational pipeline of CT for the extraction of conformational features from dynamic and disordered systems. We develop dynamic Circuit Topology, a method that goes beyond fold description by also allowing comparison of unstructured polymers and the definition of a topological similarity score between different IDPs. Among other things, this technique allows for a novel representation of the folding pathway on the topological space, capable of highlighting transient states and possibly relate them to key conformations for pharmaceutical applications.

Theoretical frameworks have the double scope of enhancing our understanding of phenomena and assist and guide experimental exploration. In recent years, single molecule force spectroscopy techniques have provided an unprecedented

insight into topological transitions in molecules, allowing for direct observation of transient states in molecular processes such as protein folding. However, interpretation of such experiments can sometimes be challenging, mostly due to the complexity and heterogeneity of interactions that can occur during folding under physiological conditions. Protein folding, in the cellular environment, is inherently a collaborative process. Molecular chaperones are a class of proteins with a variety of functions aimed at assisting the folding process. While Optical Tweezers (OT) and other single molecule assays have studied in depth protein-chaperone interaction in isolation with specific chaperone systems, a full picture of what happens in the cellular cytoplasm is still missing. Chapter 4 aims at taking the first step in this direction, by presenting OT protein and DNA pulling experiments in a complex environment, the diluted E coli cytosol. In our cytosol solution, we find chaperones and other molecules affecting folding dynamics in physiological concentrations. Therefore, we explore the effect of the combined chaperone machinery of the bacterial cytosolic interactome on DNA and Maltose Binding Protein (MBP), here used as a topological sensor. We also draw a parallel between our computational and experimental line of research, by providing an example on how to use the CT formalism to analyze and visualize Force Spectroscopy data. Both chaperones and DNA are important drug targets, and carachterizing their dynamic properties and interactions in the cellular environment would provide insightful information for what concerns drug binding kinetics.

Proteins are not the only biopolymers with complex architectures which are fundamental for the steady completion of their function. The genome folds to function too, albeit through different mechanisms. The 3D arrangement of the human genome inside the nucleus is key to correct gene expression and regulation, and an important component of epigenetic changes in health and disease conditions. In Chapter 5 we develop a multi-scale heterogeneity analysis based on Circuit Topology in order to characterize the 3D folds of single-cell genome structures experimentally derived by Hi-C. We show the potential of CT for the identification of statistically distinct topological states in the chromosomes, while also highlighting highly conserved motifs in chromatin looping. This research line might shed a light on the intricate relationship between genomic activity and chromosome organization, as well as provide a practical tool for identifying disease-related structural biomarkers.

The flexibility of the CT framework relies in the freedom it provides for contact definition, which can be adapted to the system under study without loss of generality. As such, Circuit Topology can be applied to any chain-like object where it is possible to define a direction and intra- (or inter-, in case of multi-chain systems) chain contacts. These polymer-like objects can be abstract and not necessarily belong to the biological field. In Chapter 6 we investigate one of these applications, exploiting the CT formalism to characterize the communication style in written texts in a dataset composed by true and fake news articles. Here, we treat text like a chain of sentences, and define contacts between sentences by semantic similarity: those sentences which are close in meaning create a contact in the textual chain. Semantic similarity is quantified by word embeddings, numerical attributes of textual elements obtained from state-of-the-art pre-trained machine learning models. We demonstrate that, in our dataset, there is statistical variation in the topology of semantic similarity between true and fake news, corresponding to two distinct communication styles.

Finally, in Chapter 7, we draw our conclusions, discussing possible future applications and opportunities based on the current theoretical and computational state of the CT framework. A few main directions are delineated, among which, the embedding of backbone and network topology, multi-chain CT, the localization of CT parameters, and the potential applications of CT in machine learning pipelines for harnessing its predictive potential.