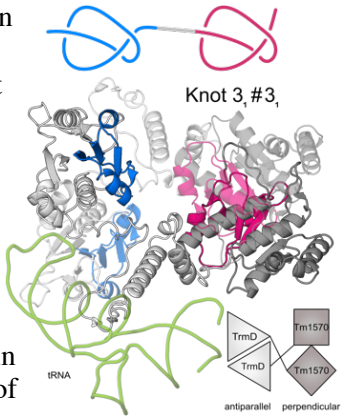
Classification of entangled proteins with intra-chain bonds and influence of entanglement on misfolding

This project is concerned with a few seemingly unrelated research areas: entanglement, protein folding and misfolding, as well as artificial intelligence (AI) and big data approach. A newly developed AI tool AlphaFold has recently led to the ground-breaking achievements in structural biology, predicting 3-dimensional structures of proteins with almost identical quality as in experiment. AI databases AlphaFold (by Google) and ESMFold (by Meta/Facebook) have predicted together 800 million protein structures, which we will use in this project to correlate entanglement with biological function, likelihood for misfolding, and to characterize other features of entangled proteins.

Proteins are basic building blocks of living organisms and play many important biological roles. Each protein is a chain made of a few hundred, or even a few thousand amino acids (there 21 basic types of such amino acids). In appropriate conditions such a chain attains some particular three-dimensional shape, which is called the native structure; formation of such a shape is necessary so that a protein can perform its biological function. It follows that a sequence of amino acids in a protein must somehow determine its three-dimensional native structure, as well as its function. To conduct biological function, a protein however has to fold to its native shape; if it does not fold properly (so called misfolding occurs), a protein cannot perform its function.



Because proteins are long chains, one may suspect that they may also get entangled and form e.g. knots, analogously to knots formed on a rope or a phone cable. For many years it had been believed that, because of their complicated structure, entanglement does not appear in native states of proteins. However, it turns out that it is not true – entangled proteins have been found, and the Principal Investigator of this project is one of the world leaders studying such structures. Recently, based on 200 million structures predicted by AlphaFold, the PI has discovered and classified 700.000 new potentially knotted proteins. Moreover, the PI determined the first doubly knotted protein, which is shown in the Figure. How this topology influences folding, stability and likelihood of misfolding are new timely questions, which need to be answered.

There are other types of entangled structures in proteins discovered by the PI, such as lassos, links, theta-curves, etc. Lassos, which are the most common ones, consist of two geometric components: a loop formed by the protein backbone (and closed by a (non)-covalent contact between two residues), and the piercing of that loop (one or more times) by one terminus of a protein. Lassos have been known for 10 years, however due to insufficient data set we still do not understand some of their fundamental aspects: how complex lasso motifs could be, or what are their functions in proteins. It was shown that proteins with non-covalent lassos can escape from a typical pathway of removing unfolded proteins, however it has never been thoroughly investigated how lasso and knots motifs influence likelihood of misfolding.

It is believed that protein misfolding is the primary cause of Alzheimer's disease, Parkinson's disease, Huntington's disease, and other diseases. Some knotted proteins were already point out as involved in Parkinson's disease. On the other hand, our preliminary data based on proteins from E. Coli proteome, which were investigated experimentally, shows that there is clear correlation between misfolding and entanglement. Based on all proteins from PDB we found that 61% of proteins (based on E.Coli) which are able to escape from degradation pathway are entangled (most of them form lassos). In this project, based on 850 million structures deposited in AlphaFold and ESMFold databases, we are going to thoroughly extend this analysis.

More precisely, in this project we are going to address the following fundamental issues: discover and classify all types of lasso, links, theta curves and other new types of motifs in protein structures predicted by AlphaFold and ESMFold. These results will broaden our knowledge about protein complexity, and can lead to identification of correlations between biological function and entanglement. Our preliminary data based on AlphaFold shows that lasso motifs occur more frequently in allosteric proteins than non-allosteric proteins and occur more in enzymes than non-enzymes. Clearly, evolutionary selection pressures must enrich them for some functional benefit. Based on AlphaFold, we have already identified 700.000 potentially knotted proteins, which we will use to investigate correlations between entanglement and misfolding. We also aim to conduct massive numerical simulations to predict likelihood of protein folding/misfolding for proteins from different organisms, that can be directly investigate experimentally. Since we are the first group in the word which proved experimentally existence of doubly knotted proteins, e.g. TrmD-TM1570, we will also be able to establish theory of their folding; moreover, we will show how such topology influences misfolding and will verify it experimentally.