Biological code of knots – identification of knotted patterns in biomolecules via AI approach

This project is concerned with a few seemingly unrelated research areas: proteins, knots, and artificial intelligence. Proteins are basic building blocks of living organisms, which play many important biological roles. There are 21 amino acids, which are fundamental building blocks of proteins. Each protein is a chain made of a few hundred, or even a few thousand 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 (which can be regarded as a long word made of 21 types of letters) must somehow determine the three-dimensional native structure of the protein, as well as its function. Understanding of such relations is one of the aims of structural biology; as it turns out, it is also one of the greatest challenges of modern science.

Because proteins are long chains, one may suspect that they may also form knots – analogously to knots formed on a rope, phone cable, etc. For many years it had been believed that, because of their complicated structure, knots cannot be formed in a native state of a protein. However, it turns out that it is not true – such knots have been found, and the Principal Investigator of this project is one of the world leaders studying such structures. It should be stressed that information whether a given chain is knotted is more fundamental than details of its three-dimensional shape – so this is one reason why research devoted to knotted proteins is so important. Nonetheless, independently of many successes in this research direction, various issues concerning knotted proteins – in particular what are biological functions of knots – still remain to be understood. It should also be stressed that this research direction is inherently related to mathematical knot theory, whose aim is to describe properties and to classify knots as mathematical objects.

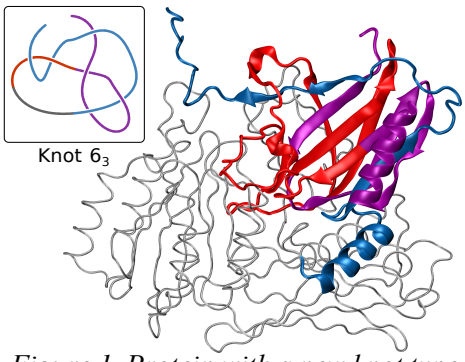

Figure 1. Protein with a new knot type identified by our team, based on the data from the AlphaFold database.

Chances to achieve the goals listed above are constrained by the amount of available data. Identification of three-dimensional protein structure is costly and time-consuming. The Principal Investigator of this project has already shown that a few percent of all proteins are knotted proteins – on one hand this is a large number, but on the other hand it is still not sufficient to answer most important questions about the role of knots in proteins. However, in the last year the amount of data crucial for structural biology has significantly increased, as a result of ground-breaking developments based on methods of artificial intelligence: the AlphaFold 2 (AF 2) program by Google, based on artificial intelligence algorithms, predicted structure of more than 400000 proteins from 20 different genomes, which exceeds the amount of experimental data. Furthermore, a preliminary analysis for a human organism, conducted by our team, has shown that AF 2 predicts knots, also of types not identified before. In this project we are going to broaden analysis of knotted proteins to include results found by AF 2, and to create new artificial intelligence algorithms for analysis of such proteins. We hope that such an approach will result in discoveries, whose importance will be analogous to the importance of the results of AlphaFold in comparison to traditional methods of protein structure prediction.

More precisely, in this project we are going to answer, among others, the following fundamental questions: what types of knots may be formed in proteins, what amino acid sequences can be responsible for knotting (and also how Alpha Fold predicted knotted structures without using examples of other knotted proteins), is it possible to engineer an amino acid sequence in order to create an “artificial” knotted protein. To this end we will conduct a vast, topological and biological overview of AlphaFold data, and we will also use methods of artificial intelligence in order to determine a knotting code, structure prediction, and engineering of artificial knotted proteins. Moreover, we plan to show that methods of artificial intelligence can be used in the topological analysis of various other entangled structures (such as lassos, theta-curves, cysteine knots), which will be accompanied by a rigorous mathematical analysis.

This project will be conducted by an international, interdisciplinary team, consisting of experts in artificial intelligence and machine learning (the Czech team); mathematicians, whose main task will be to apply knot theory to model open chains (Slovenian team); and experts in the analysis of entangled proteins (the Polish team). It should be stressed, that methods of artificial intelligence are currently being actively developed and used in various disciplines. Therefore, it is likely that, in near future, the methods developed in this project will find applications in analysis of other biological problems (e.g. related to chromosomes or RNA), will help in drug design (as one of the proteins from SARS-CoV-2 has a slipknot topology), or in design of stable materials based on nontrivial polymer topology.