Protein-protein interaction is an important process, that is crucial for many functions of a cell - such as response to outside stimuli. However, it is notoriously hard to study. While there are some experimental methods to verify potential interaction partners, they usually require at least some suspects. My goal is to fill this void, tapping into the largest protein-related resource available - sequence databases.

Recently, a new theoretical approach to study protein 3D structure and interaction was proposed - "Direct Coupling Analysis" (DCA). It is a statistical method, which for a set of related sequences calculates a model of correlation patterns between corresponding residues in the proteins. Through this, it is possible to score the tendency to co-mutate between given residues, which tends to arise from joint evolutionary pressure - so should be in some form of contact. This way, if we find two positions that are scored highly by DCA, we can expect them to be close in the structure. Additionally, model of coevolutionary pattern given by DCA can be used to check how well a given sequence fits to the already analyzed set. For two groups of sequences representing interacting protein families, we can assign specific interaction partners, by assuming that the correct matching is the one exhibiting the largest coevolution. Finally, with DCA it is possible to assign a total coevolutionary score to two sets of possible interaction partners - which can be used to guide the expensive experimental approach.

My aim is to move DCA to a new level by incorporating, currently omitted, biological informations. Current DCA implementations by design treat amino acids as independent, dissimilar entities - which absolutely doesn't hold true for actual biological reality. Currently DCA struggles with fast evolving sequences - high variability leads to less meaningful results. Introducing this new data into DCA is not a trivial task, but the success will pave the road to new applications, including aforementioned problematic datasets. My bioinformatic background, along with the vast experience of my supervisor and collaborator (Joanna Sulkowska and Martin Weigt, pioneer of protein study using DCA, and one of the authors of the method, respectively) guarantee the success of this project. Additionally, this input data modification will be introduced in a way that reduces the complexity, and thus the computational times, of this problem. Which ties into another goal of this project, improving algorithms efficiency, so that DCA will become a viable method for potential interaction screening of whole proteomic databases.

On a smaller scale, I will use my new version of DCA to look closely at coevolution present in specific interfaces. In particular, I will focus on two classes of proteins, both characterized by non-trivial topology. Here it means a presence of a knot-like fold within their structure. Such proteins, although very conserved structuraly tend to exhibit high sequential variability. This complicates the application of traditional DCA implementations, but can be overcome using additional physico-chemical data I will introduce. Additionally, both types of proteins I want to study can be classified as fusion proteins (so at least two separate molecules which at some point during evolution joined to form one, continuous protein chain), and the interface I will analyze will be the interface within - between the internal, historically separate, domains.

In particular, I will study the internal interface of bidirectional transmembrane transportes, between their internal repeats (the reason why these proteins are assumed to originate from a fusion of a homodimer). In some cases, this fusion resulted in a knot - yet not always. I plan to compare the course of evolution between this two different folds to describe this discrepancy.

Additionally, we have recently found a group of fusion methyltransferases - made up of two, each usually homodimeric, methyltransferases, both of them knotted by themselves. I want to see how do they compare to their unfused counterparts, and predict how, and through what interactions, such a protein can function. And through protein interface - what is the fusion structure like. Finding a doubly knotted protein, as this one could be, would be a breakthrough in the field of protein folding and understanding of protein topology.