Proteolytic enzymes, or proteases, play essential roles in regulating key biological processes by selectively cleaving peptide bonds in protein and peptide substrates. They play important roles in protein activation and degradation, digestion, apoptosis, and pathogen replication. Dysfunction of proteases is associated with numerous diseases, making them critical targets for therapeutic intervention. Therefore, accurate identification of protease cleavage sites is vital for understanding enzyme specificity and for developing novel inhibitors and therapeutic strategies.

Current computational methods for cleavage site identification predominantly rely on sequence-based approaches. Although these methods have produced valuable insights, they suffer from major limitations, particularly their inability to account for the structural and dynamic context of enzyme-substrate interactions. As a result, our understanding of proteolytic mechanisms and enzyme selectivity remains incomplete, and the rational design of protease-targeting therapeutics is impeded.

This project proposes the development of an innovative computational framework for predicting proteolytic cleavage sites by integrating structural modeling of enzyme-substrate complexes with machine learning–based classification. Our approach will leverage advanced protein structure prediction tools (e.g., AlphaFold-Multimer, AlphaFold3) and flexible protein–peptide docking simulations using the CABS coarse-grained model. These initial models will be further refined through molecular mechanics and molecular dynamics simulations to generate accurate representations of enzyme–substrate geometries.

We will develop a machine learning–based scoring function to differentiate between binding patterns characteristic of cleaved and non-cleaved peptides, using a rich set of structural and sequence-derived features. The model will be trained on large datasets of molecular structures derived from experimentally validated cleaved and non-cleaved peptides. Two complementary strategies will be implemented for cleavage site identification: (1) a structure-based approach using geometric criteria, focusing on the spatial relationship between scissile bond atoms and catalytic residues; and (2) a scoring model trained to recognize enzyme-peptide interaction features characteristic for cleavable substrates.

Our preliminary studies have demonstrated the feasibility of this approach in enzyme-substrate systems involving pepsin, renin, and HIV-1 protease. For these enzymes, cleavage sites predicted through coarse-grained docking and structural analysis closely matched experimental observations. Additionally, early classification experiments revealed that cleaved and non-cleaved peptides exhibit distinct binding patterns, supporting the applicability of machine learning based methods in this context.

The proposed methodology represents a paradigm shift from traditional sequence-centric models to an integrated, structure-informed framework. This approach will enable the prediction of cleavage sites in systems previously considered intractable and will provide mechanistic insights into enzyme–substrate recognition. Furthermore, it will facilitate the identification of novel substrate motifs, elucidate multi-residue cooperativity effects, and reveal conformational factors influencing enzymatic specificity—capabilities beyond the reach of current methods. By integrating structural modeling, docking simulations, and machine learning, this project will advance the field of computational enzymology. It will yield a robust and scalable tool for both academic and industrial researchers, supporting high-throughput cleavage site prediction, leading to a rational inhibitor design, and a deeper understanding of protease biology.