# Mapping Protein Language: Exploring Amino Acid Functionalitythrough Machine Learning

Filip Kapral1, Václav Hanžl1, Vojtěch Spiwok1

1 Department of Biochemistry and Microbiology, University of Chemistry and Technology, Prague, Technická 5, Prague 6, 166 28, Czech Republic

Language is a structured system that enables humans to exchange information through words and sentences, where meaning depends on both vocabulary and context. Similarly, proteins can be viewed as a universal "language" with amino acids acting as words whose functions are determined by their sequence and structural environment. Just as natural language contains homonyms – words like bow (a weapon) and bow (to bend) that share spelling but differ in meaning - amino acids such as histidine can adopt distinct roles depending on their context. For instance, histidine may function as a charged surface residue, a metal-chelating site, or a catalytic component in an enzyme.

Advances in machine learning, particularly transformer-based architectures, have revolutionized our ability to decode such biological and linguistic patterns. Originally developed for natural language processing (e.g., translation models like ChatGPT), transformers now power breakthroughs in protein science, including structure prediction tools like AlphaFold, OpenFold, and ESMfold. ESMfold employs a large language model to "read" amino acid sequences, treating each residue as a word embedded in an 1280-dimensional vector. As these vectors pass through the transformer, they evolve to encode not just the identity of the amino acid but also its structural and functional context - mirroring how word meanings shift in sentences.

Building on this, we use ESM-2 (Evolutionary Scale Modeling) to map residue-level functionality across human proteins. Each amino acid is transformed into a vector representation, visualized in 2D and linked to 3D protein structures for deeper analysis. Our goal is to create an interactive atlas of amino acid residues, enabling researchers to explore the diverse roles of residues in proteins. By integrating protein language modeling with structural biology, this work bridges computational and experimental insights, offering a new lens to study protein function.

This work was supported by the Ministry of Education, Youth and Sports (reg. No. LUC24136, LM2023055), ELIXIR CZ and COST (ML4NGP, CA21160).