Comparing and contrasting the protein embedding landscape with commercially important enzymes haplotypes recovered from meta-genomes.

This research project seeks to investigate and compare haplotype enzymes recovered computationally from metagenomic DNA sequences with their protein embedding landscape counterparts obtained through large language models. The focus of this investigation will be on the aldo/keto reductase (AKR2) family of enzymes.

The primary objective of this study is to explore the similarities and differences between the haplotype-based computational approach and the protein embedding-based approach in capturing the characteristics and functional properties of AKR2 enzymes. By leveraging metagenomic DNA sequences, the computational recovery of haplotype enzymes aims to identify and analyze specific variations within the AKR2 family.

In parallel, the investigation will utilize protein embeddings generated by large language models, such as the ProtT5 Model, to represent the structural and functional properties of AKR2 enzymes. These protein embeddings provide a computer-friendly format that tries to captures information about protein sequences.

By comparing and contrasting the results obtained from the two approaches, this research aims to gain insights into the strengths and limitations of each method. It also seeks to assess the effectiveness of protein embeddings in representing and characterizing AKR2 enzymes when compared to the computational recovery of haplotype enzymes from metagenomic DNA sequences.

Metagenomics
Metagenomics, perhaps better described as “environmental genomics” is the study of DNA sequences of organisms that co-exist in a particular environment; such as skin or soil. In such environments, microbes live in dense communities with a complex network of biological interactions and have typically adapted to produce enzymes -- substances responsible for “interesting” biochemical reactions -- to fill a niche in the system, such as breakdown of biomass, antimicrobial defence and the ability to thrive in the presence of extreme conditions.

The communities of interest are symbiotic, preventing simple isolation and culturing of individual species thereby rendering traditional single-species genomics techniques unsuitable. Thus, the study of such communities poses significant difficulties, even before looking at DNA. Sampling bias can prevent low-abundance or rare species from appearing in the sample or cause them to appear as noise. Highly-abundant species may also obscure others. The presence of more than one organism in a sample complicates the construction of genomes from sequencing data.

Hansel and Gretle
Hansel and Gretel is an algorithm and data structure designed for the recovery of haplotype enzymes from metagenomic DNA reads. Developed by Dr Sam Nicholls and colleagues at Aberystwyth University the methods can recover the set of distinct co-occurring enzyme variants which exist within the population (https://doi.org/10.1093/bioinformatics/btaa977).

Protein Embeddings
Protein embeddings are a method used to represent structural and functional properties of a protein, mostly from its sequence only, in a machine-friendly format (vector representation). Generating these embeddings is computationally expensive, but once computed they can be leveraged for different tasks, such as sequence similarity search, sequence clustering, and sequence classification.

ProtT5 Model is an example of one of these embeddings and it is available for download here (https://www.uniprot.org/help/embeddings). The paper describing how these model are made can be found here (https://arxiv.org/pdf/2007.06225.pdf).

Protein embeddings serve as a valuable method for representing the structural and functional properties of proteins in a format that is easily processed by computers. These embeddings are typically generated from the protein's amino acid sequence and offer a compact, vector-based representation. Although the computational cost of generating protein embeddings can be high, once computed, they can be leveraged for various tasks, including sequence similarity search, sequence clustering, and sequence classification.

One notable example of protein embeddings is the ProtT5 Model, which is available for download at (https://www.uniprot.org/help/embeddings). The model is described in detail in the associated research paper, which can be found at (https://arxiv.org/pdf/2007.06225.pdf). The ProtT5 Model utilizes the T5 (Text-to-Text Transfer Transformer) architecture, originally developed for natural language processing tasks. In the context of proteins, the model is pre-trained on a vast dataset of amino acid sequences, enabling it to learn representations that effectively capture the underlying characteristics of proteins.