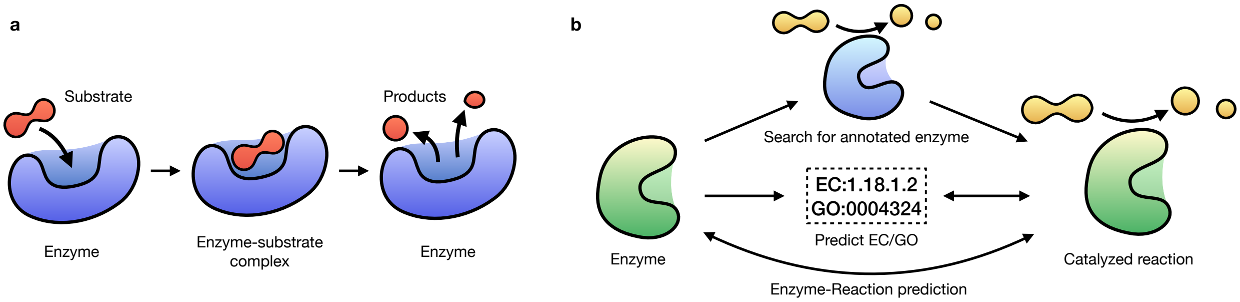
Dataset documentation and intended uses:

1. Introduction

The current methodologies for enzyme annotation primarily rely on established databases and classifications such as KEGG Orthology (KO), Enzyme Commission (EC) numbers, and Gene Ontology (GO) annotations, each with its specific focus and methodology. For instance, the EC system categorizes enzymes based on the chemical reactions they catalyze, providing a hierarchical numerical classification. KO links gene products to their functional orthologs across different species, whereas GO offers a broader ontology for describing the roles of genes and proteins in any organism.



Despite their widespread use, these systems have notable limitations. The EC classification, while widely used, sometimes groups vastly different enzymes under the same category or subdivides similar ones excessively, based on the substrates they interact with—leading to ambiguities in enzyme function characterization. GO annotations, although comprehensive, frequently lack specificity in defining enzyme functions and suffer from an underdeveloped database structure. Similarly, KO tends to categorize based on gene or protein families rather than specific functions, potentially assigning different identifiers to proteins with identical functions.

Given these challenges, we propose a novel benchmark and a new enzyme-reaction dataset to learn enzymes more accurately by focusing on their catalyzed reactions directly rather than solely on gene family or human-assigned function types. Our approach also leverages machine learning techniques—graph representation learning and protein language models—to analyze enzyme reaction data, providing a more nuanced understanding of enzyme functionality. This method aims to overcome the limitations of current annotation systems by offering a clearer, more consistent categorization of enzymes based on their biochemical roles, which could significantly enhance both academic research and industrial applications in enzyme technology.

2. Collection and Workflow

2.1. Overview

Our study utilizes a comprehensive dataset compiled from the SwissProt and Rhea databases. SwissProt, a curated subset of the UniProt database, has been selected for its high-quality, human-derived functional annotations of protein sequences. This section of UniProt is particularly valuable for its expert-reviewed entries, which ensure reliable and accurate functional data, making it ideal for our analysis. Rhea is employed for its precise mapping from enzymes to specific catalyzed functions, offering detailed descriptions of biochemical reactions.

2.2. Data Collection

The SwissProt and Rhea dataset are downloaded on January 8, 2024, and includes data entries up to this date, providing the most recent and comprehensive data available for our study. We selectively exclude water molecules and unspecific functional groups that could mask the true molecular structures. Conversely, we remove metal ions, gas molecules, and other small molecules because of their potential to bind to proteins, a characteristic that presents a valuable learning feature for our model. To this end, the total dataset comprises 178,463 positive enzyme-reaction pairs, including 178,327 unique enzymes and 7,726 unique reactions.

2.3. Comparison

There are two datasets related to the enzyme-reaction prediction task. The first one is from ESP, which used GO annotation database for UniProt dataset, lay emphasis on the substrate binding to the enzyme. The ESP dataset contains 18,351 enzyme-substrate pairs with experimental evidence for substrate binding, contains 12,156 unique enzymes and 1,379 unique molecules. The other dataset is from EnzymeMap, which used as training set in CLIPZyme. The EnzymeMap dataset is a high-quality dataset of atom mapped and balanced enzymatic reaction, with enzyme information from BRENDA. This dataset contains 46,356 enzyme driven reactions, including 16,776 distinct reactions and 12,749 enzymes.

3. Impact and Challenge

We introduce Reactzyme, a new benchmark for enzyme-reaction prediction. Unlike previous methods that rely on protein sequence or structure similarity or provide EC/GO annotations to predict reaction, our approach directly evaluates the mapping between enzymes and their catalyzed reactions. These enzyme-reaction prediction methods are able to handle protein with novel reactions and to discover proteins that catalyze unreported reactions. Reactzyme thus provides a robust framework for advancing our understanding of enzyme functionality and expanding the known repertoire of biochemical reactions.

We evaluate the performance of several baselines on the Reactzyme dataset, including MAT, UniMol for reactions, as well as ESM, SaProt for enzymes, with additional GNN encodings. While the baselines demonstrate competitive results on time- and enzyme-similarity-based splits, the reaction-similarity-based split remains particularly challenging. This difficulty may arise from the presence of many unseen reactions in the test set of the reaction-similarity-based split. One potential avenue for improvement is to explore contrastive learning techniques to address this challenge. However, we acknowledge that this remains an open problem for researchers in our community to tackle.

The Reactzyme benchmark facilitates the evaluation of models working with protein and molecule representations, which requires a comprehensive understanding in both modalities. Models demonstrating high performance in enzyme-reaction prediction can be further leveraged for protein function prediction and enzyme discovery. This includes identifying key enzymes in biosynthesis and discovering potent enzymes for degrading emerging pollutants, for these reactions that have not been previously found in enzymes.

URL to website/platform:
Our dataset can be downloaded via https://zenodo.org/records/11494913

URL to Croissant metadata:
The dataset is neither image or text/audio, thus it is hard for us to provide Croissant metadata, but our dataset can be found via https://zenodo.org/records/11494913

Author Statement:
We bare all responsibility in case of violation of rights, etc., and we confirm the data license.

Hosting, Licensing, and Maintenance plan:
We will be hosting and maintaining the dataset. The dataset and baselines will be uploaded on GitHub later with the correspondence email: Chenqing.hua@mail.mcgill.ca. The general content of the dataset will not be updated, but we will update if errors are found. The dataset is constructed with MIT license.