1. Dataset Name

EzyPredict Dataset
Created by : Simran Kaur Aulakh in Sept - October 2023.

2. Description:

A dataset containing protein level information for all enzymes in the Yeast8(https://github.com/SysBioChalmers/yeast-GEM) *Saccharomyces cerevisiae* metabolic model. The dataset contains :

- UniProtID - unique protein identifiers from the UniProt database (https://www.uniprot.org/)
- Peptide statistics for each protein (https://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/)
- ESM2 encodings for each protein sequence (using https://github.com/facebookresearch/esm)
- Enzyme Commision labels (up to level 2) for each protein (also from UniProt)

3. Purpose & Motivation:

The purpose of this dataset is to train a machine learning model to predict the function (Enzyme Commission number up to level 2) of an enzyme based on its sequence and physicochemical properties. This small dataset was created as an exercise during a course and is not a research-scale or deployment scale dataset or implementation fit for the purpose of prediction enzyme function in general. Much larger datasets are required for that.

4. Collection Process:

The dataset was created by sourcing ORF (Open Reading Frame) IDs from the Yeast8 model, converting these to UniProtIDs, then using sequences of each UniProtID to :

- Fetch peptide biophysical and biochemical data using the pepstats API (https://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/). Only the following were retained :
    - Isoelectric Point
    - Fraction of amino acids classified as each of the following categories :
        - Tiny
        - Small
        - Aliphatic
        - Aromatic
        - Non-polar
        - Polar
        - Charged
        - Basic
        - Acidic

- Fetch Enzyme Commission (EC) numbers for functional annotation using a Uniprot download file (“uniprot2EC_uniprotkb_download_2023_09_07.tsv”). Only the first 2 levels of EC numbers were considered.
- Convert protein sequences into ESM2 Encodings: Protein sequences were encoded using the ESM2 model,transforming them into fixed-size vectors representing the sequence. Peptides longer than 1024 amino acids were discarded for this training exercise.

Raw data : all raw data files created during the preprocessing were saved and are also provided in the repository

5. Dataset Size:

- Total Entries: 944
- Features: 1292 (1291 inputs + 1 output)
- Reduced Features :

6. Data Columns:

- UniProtID: A unique identifier for each protein sourced from UniProt.
- Pepstats Features:
    - Isoelectric Point
    - Tiny
    - Small
    - Aliphatic
    - Aromatic
    - Non-polar
    - Polar
    - Charged
    - Basic
    - Acidic
- Encodings: 1028 columns that correspond to the vector of encodings derived from the ESM2 model
- EC number: The Enzyme Commission number denoting the function of the protein.

Missing Information : The dataset has been filtered during the creation process such that any UniProtIDs with missing information in any of the columns has been left out. So there are no missing values in this dataset.

7. Potential Biases:

The dataset only contains enzymes from the yeast *Saccharomyces cerevisiae* for the purpose of simplicity such that it can be run on a local computer for the purpose of an exercise in learning how to create an ML model. It is not generalisable to other organisms.

8. Ethical Review:

The dataset contains no data from humans.

All databases and software accessed for the creation of the dataset are under open licences, free for usage by anyone in the public. All sources have been cited.

9. Usage:

This dataset can be used for:

- Training machine learning models to predict enzyme function of *S.cerevisiae* enzymes as a learning exercise
- Studying correlations between protein sequence properties and their function.

This dataset should not be used for:

- Extrapolating the results to other organisms
- High-confidence prediction of enzyme function in *S.cerevisiae*
- Predictions of non enzymatic protein function

10. Maintenance:

The dataset can be updated as new protein sequences get added to UniProt or when updates are made to the pepstats tool or the ESM2 model.
All data collection scripts are included in the repository.

11. Access and Distribution:

The dataset can be accessed through the project's Github repository : https://github.com/simranolak/EzyPredict. It's a private repository so access will be provided upon request.

12. Licences:

UniProt - Creative Commons (CC BY 4.0 LIcence) (https://www.uniprot.org/help/license)
Pepstats - Creative Commons (CC0 Licence) (https://www.ebi.ac.uk/licencing)
ESM2 - under MIT licence - open for any copying, modification and use without any liability or warranty on the part of facebook research.
(https://github.com/facebookresearch/esm/blob/main/LICENSE)