ERxn: Predicting enzymatic reactions using attention based deep learning model with structural data

Jonathan Funk (s212697)[1], Eric Bautista Farrerons (s212514)[1], Belkis Chau Zhong (s181024)[1] and Ole Winther[2,3]

1 DTU Bioinformatics, Technical University of Denmark; 2 The Bioinformatics Centre, Department of Biology, University of Copenhagen, 3 DTU Compute, Technical University of Denmark

Introduction

Enzymes are proteins that function as biological catalysts to accelerate chemical reactions. The reactions catalyzed by enzymes are determined by their three-dimensional structures.[1] [2] The low availability of protein structures has been a limiting factor that hampered the design of machine learning algorithms to predict protein functions for many decades [5]. As a consequence, most of the currently available machine learning methods that aim to predict biochemical reactions, for example, the molecular transformer model [6], only rely on protein sequences. However, with the highly accurate protein structure predictions released by AlphaFold2 [4, 7], structural data is now available and we, therefore, propose using both protein sequence and structural data, combined with attention-based deep learning models, to predict the enzymatic reactions.

Methodology

- We construct a new dataset from a recent version of SwissProt where proteins have experimental evidence for the biochemical reactions they catalyze. Using the AlphaFold database [4, 7] we augment our dataset with protein structure predictions.
- We developed a model to classify enzymes into enzyme commission (EC) class based on their protein structures.
- We made sure that all protein classes are present in the training, test, and validation dataset.
- Classes which were over-represented were down-sampled and classes with less than 10 occurrences were discarded.
- Self-supervised pre-training was used during experiments, however, the performance was worse with pre-training.
- We reproduced the molecular transformer [6] and applied transfer learning on biochemical reactions obtained from BRENDA [3].
- Biochemical reactions which could not fully be expressed using simplified molecular input line entry system (SMILES) were discarded. This is the case for example the case when macromolecules or undefined reagents of a certain class were reported.

Our Enzymatic Transformer

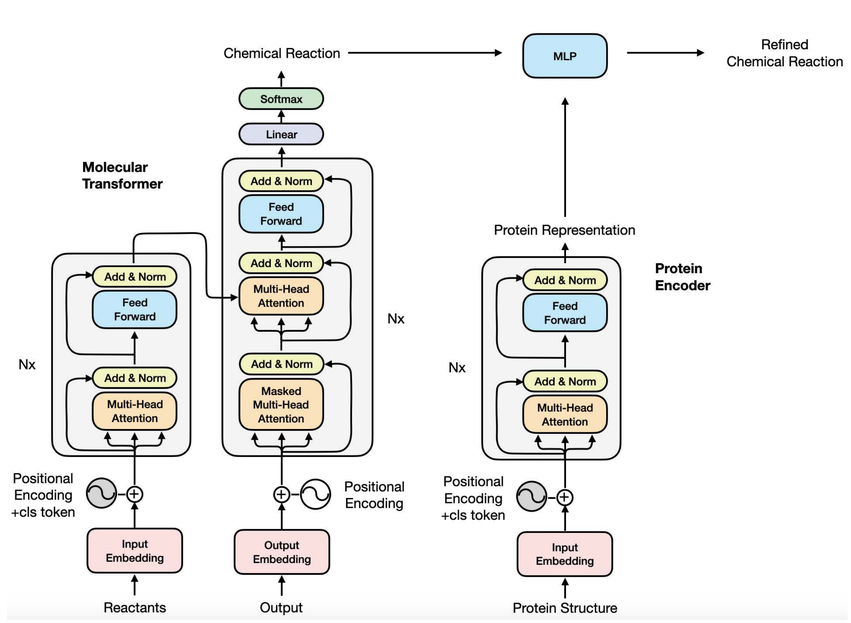

Figure 1: Architecture of our models. On the left is the molecular transformer, which solves chemical equations. On the right is the enzyme encoder, which classifies enzymes into EC group according to their structure. The molecular transformer has encoder and decoder blocks, while the enzyme encoder only consists of encoder blocks. A final multi layer perceptron connects the two individual models.

Datasets visualization

Figure 3 shows the distribution of the number of atoms per protein. A cutoff value of 10,000 atoms was chosen for training. All structures with less atoms were padded to ensure equal sequence length. Figure 2 shows the distribution of enzymes based on their EC numbers. EC class 2 enzymes are the most abundant in the BRENDA dataset. In Figure 4, an example of a reaction from the MIT_mixed_augm dataset and its corresponding SMILES representation is presented.

Figure 2: Distribution of different enzyme classes in the BRENDA datasets.

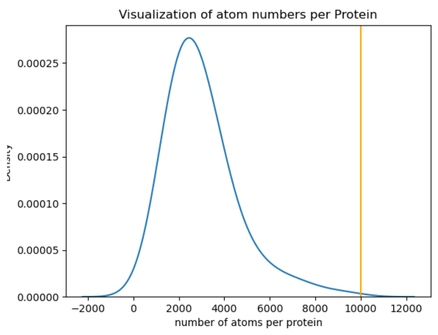

Figure 3: Density plot of the number of atoms per protein. The orange line shows the cutoff value.

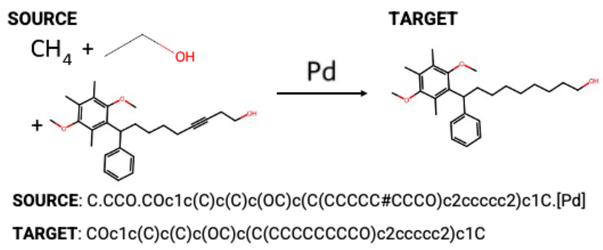

Figure 4: Example of a reaction from the MIT_mixed_augm dataset and SMILES.

Model performance

The accuracy of the original Molecular Transformer is the best among all methods. Both the transfer learning and BRENDA runs decrease drastically the model performance, suggesting insufficiency of training data points.

Table 1: Prediction accuracy achieved on the test set when using the Molecular Transformer model using different datasets. The prediction accuracy values were calculated after choosing the correct output from the top-5 translations.
*MT = Molecular Transformer

Method	Dataset	Accuracy
Original MT*	MIT_mixed_augm	0.942
Implemented MT*	MIT_mixed_augm	**0.942**
Implemented MT*	BRENDA	0.022
Transfer Learning	MIT_mixed_augm & BRENDA	0.175

Performance for each enzyme class

Table 2: Confusion matrix of the validation set on the final EnzymeEncoder model.

EC	1	2	3	4	5	6	7	accuracy
1	288	41	26	19	2	6	4	0.618
2	63	1209	134	41	25	22	16	0.818
3	26	55	488	32	12	5	4	0.639
4	19	38	43	297	12	9	2	0.704
5	19	47	17	12	195	3	0	0.756
6	34	72	45	18	10	409	12	0.899
7	17	16	11	3	2	1	113	0.748

Visualization of results

Using principal component analysis we can analyze how different enzyme classes are represented by the enzyme encoder.

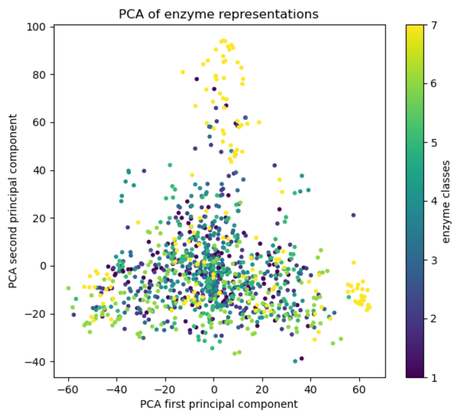

Figure 5: Principal component analysis of enzyme encoder representations, grouped by EC number.

When applying transfer learning using biochemical reaction data the rules of chemistry don't change, but the shape of the involved molecules differ widely. In enzymatic reactions researchers are particularly interested in the stereo- and regio-selectivity of reactions.

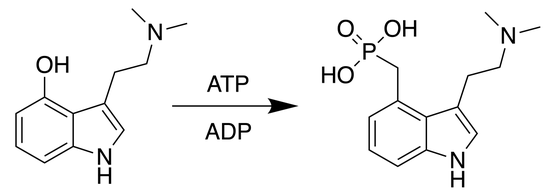

Figure 6: example reaction

Discussion

Structural data with attention based deep learning model provides a 70% accuracy of enzyme class. The molecular transformer was attempted to be fine tuned with transfer learning of the BRENDA dataset, but the accuracy dropped drastically (76.7%). This is most probably due to the small dataset from BRENDA. In the future, the data size might be increased by augmentation to improve the model performance. This study shows a good application of using structural data for enzymatic reaction prediction and this approach might be extended in the future with reaction conditions inclusive for predicting the enzyme yield.

References

[1] G. J. Babbitt PC. Understanding enzyme superfamilies: chemistry as the fundamental determinant in the evolution of new catalytic activities. *Journal of Biological Chemistry*, 272(49):30591–30594, 1997.

[2] R. A. Jensen. Enzyme recruitment in evolution of new function. *Annual review of microbiology*, 30(1):409–425, 1976.

[3] L. Jeske, S. Placzek, I. Schomburg, A. Chang, and D. Schomburg. Brenda in 2019: a european elixir core data resource. *Nucleic acids research*, 47(D1):D542–D549, 2019.

[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[5] S. Z. Matthew P Jacobson, Chakrapani Kalyanaraman and B. Tian. Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. *Trends in biochemical sciences*, 39(8):363–371, 2014.

[6] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572â1583, 2019. doi: 10.1021/acscentsci.9b00576.

[7] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.