Applications of Machine Learning in Enzyme Design

Jonah Nichols

May 20, 2022

1 Introduction

Plastic is an incredibly versatile and cheap material. It is used in nearly every industry, from packaging to consumable goods to medicine to research [**geyer2017production**]. Estimates for global plastic production range from 335 MT to 380 Mt [**drzyzga2019plastic**, **geyer2017production**, **leal2019overview**]. This quantity is high due to its high versatility and durability. This comes at a great cost at the end of its lifespan.

Plastic waste management can be done in a few different ways, including but not limited to placing it into the ocean, placing it into the landfill, or burning it. The more labor intensive solution is to burn the plastic to help regain some of the energy lost from its production. This comes with the harmful fumes which can wreak havoc amongst various organs and body systems [**agnes2016environmental**]. Placing it into the ocean and landfill are both quick and easy, yet have vast ecological impacts for those living in those environments.

Plastic placed into the ocean usually comes from the land, with an estimated 4.8 to 12.7 million MT of plastic entering the ocean in 2010 alone [**jambeck2015plastic**]. This causes massive ecological harm to the sea life living in the ocean. The harm comes from microplastics, which biota in the sea eat.

Bioremediation is a more viable solution for plastic waste management. While it is true that bioremediation is slower and incredibly plastic specific, it's important to remember that

2 Methods

2.1 Data Acquisition

RheaDB was utilized to find ester hydrolysis reactions according to the Enzyme Classification (EC) standards. From selected reactions, Rhea's cross references were utilized to obtain sequences for the various chemical reactions from UniprotKB. For every sequence-reaction pair, the Morgan Fingerprint Difference of the reaction was created using RDKit's built-in fingerprinting function. Due to a lack of known directional data, unknown directional enzymes were assumed to catalyse the left-to-right version of the reaction. Next the fingerprint was converted into a PyTorch tensor for import into the machine learning algorithm.

For every sequence, a one-hot encoding of the sequence was built to create a suitable output for the reaction. Each amino acid was represented by an integer. These sequences were then converted into PyTorch tensors for input into the machine learning algorithm paired up with its corresponding reaction fingerprint. These pairs were split randomly into a ratio of 75% training and 25% test datasets

2.2 Neural Network Design

Given inexperience in the subject, the neural network was designed to a very low degree of confidence. The input layer was the reaction fingerprint previously generated. Three repetitions of linear layers followed by ReLU activation layers expanded the fingerprint out to 7192 bits, corresponding to the longest sequence found. A final linear layer presented the output of the model.

2.3 Training and Analysis

For training, the inputs and outputs were batched to 64 samples and shuffled to limit overfitting. The L1 Loss function was used to determine the error of the model. Stochastic

gradient descent was utilized to optimize the model. The training sequence was run for 100 epochs to get an optimized model.

Due to time constraints, the model's learning capabilities and pattern recognition were not determined.

3 Results

Of the 9890 sequences queried from UniprotKB with the enzyme class of 3.1.x.x, 19084 reaction-sequence pairs were generated and used for training in the machine learning model.

In a visual examination of the output the neural network gave, it became apparent that the neural network had not learned much of the data. The output given was to be a one-hot encoded sequence in number form, to be converted back to a proper sequence. Unfortunately the output received was a sequence of numbers between 12-14 and then a gradual descent to 7's, presumably the "pad" character created to sync

4 Discussion

Future work on the subject is key to enhancing our ability to create and use enzymes to catalyze reactions. The most arbitrary extension of this work is in rewriting the featurization of the enzymes to more resemble other machine learning projects. Rather than represent each amino acid as a number, represent it as a 21-bit tensor. Each bit would represent a different amino acid being turned on.

Another more laborious route for the project is to move towards a 3D-based approach opposed to the current sequence-based approach. This route is more difficult due to the conversion from a structure to a one-dimensional input for a neural network. Before the fasta-approach branch it was intended as a structure-based approach but processing times were too long for a limited time scale.

Another another pathway to take would be returning the project to a more typical enzyme

design process. This would involve completely refactoring all of the work done so far, but may offer better end results. The process, described briefly, is beginning with the simulated transition state using density-functional theory (DFT) and building up the enzyme around said transition state.

Regardless, continuing work on enzyme design is irrefutably important. Creating enzymes from scratch has possibilities ranging from new antibiotics to ecological benefits.