DeepFold: Discovering Thermostable PETase Variants for Industrial Usage

Gupta, Aarav (School: Issaquah High School)

This project's purpose was to find alternatives to the PET-plastic degrading enzyme 'PETase' from vast amounts of existing protein data that function at industrial temperatures (~70°C) for an industrial systematic degradation of PET plastics. To screen at this scale, a four-stage computational pruning pipeline was developed in which about 1.2 billion unique proteins from different extremophilic environments and general protein collections went through the pipeline to find suitable PETase candidates which function at industrial temperatures. The first stage consisted of a Hidden Markov Model (a statistical model that calculates the probability of other data being related to data the model is trained on), trained on data from existing PETase variants to retain enzymes with high probability of being PETase variants. The second stage was a structure-aware SaProt model trained on PETase data (using Contrastive Learning due to limited amounts of unique PETase data) to find candidates that functioned like PETase, coupled with scoring head for structural validation and external validation for generalization beyond training distribution. The third stage contained FoldX, which used structure-aware analysis to measure each candidate's ability to withstand high temperatures, and DeepSTABp, which used sequence-based analysis to measure each candidate's denaturing threshold, both working together to provide high-confidence high temperature-functioning candidates. The final filter used AutoDock to calculate how well each PETase candidate would bind to the PET plastic structure. Through this process, one novel enzyme was discovered to function like PETase and function at near-industrial temperatures, along with the development of a revolutionary computational pipeline.

Awards Won: