Machine Learning Applied to the Optimization of Plastic-Degrading Enzymes

Romeu Fernandes PG45861

Projeto em Bioinformática

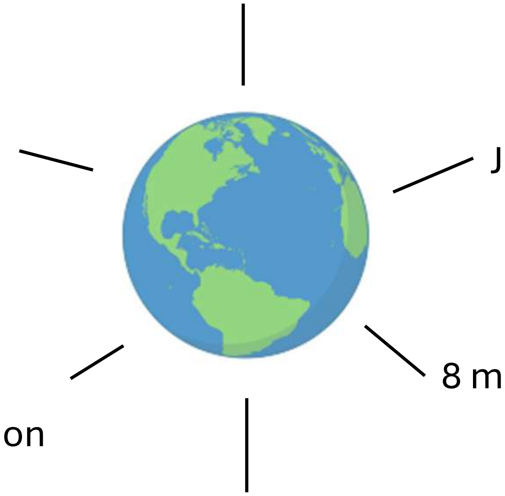
04 de junho de 2025

Supervised by:

João Carneiro

Pedro Soares

Global Plastic Crisis



380 million tons of plastic produced per year

Microplastics found in human blood, placenta, and lung tissue

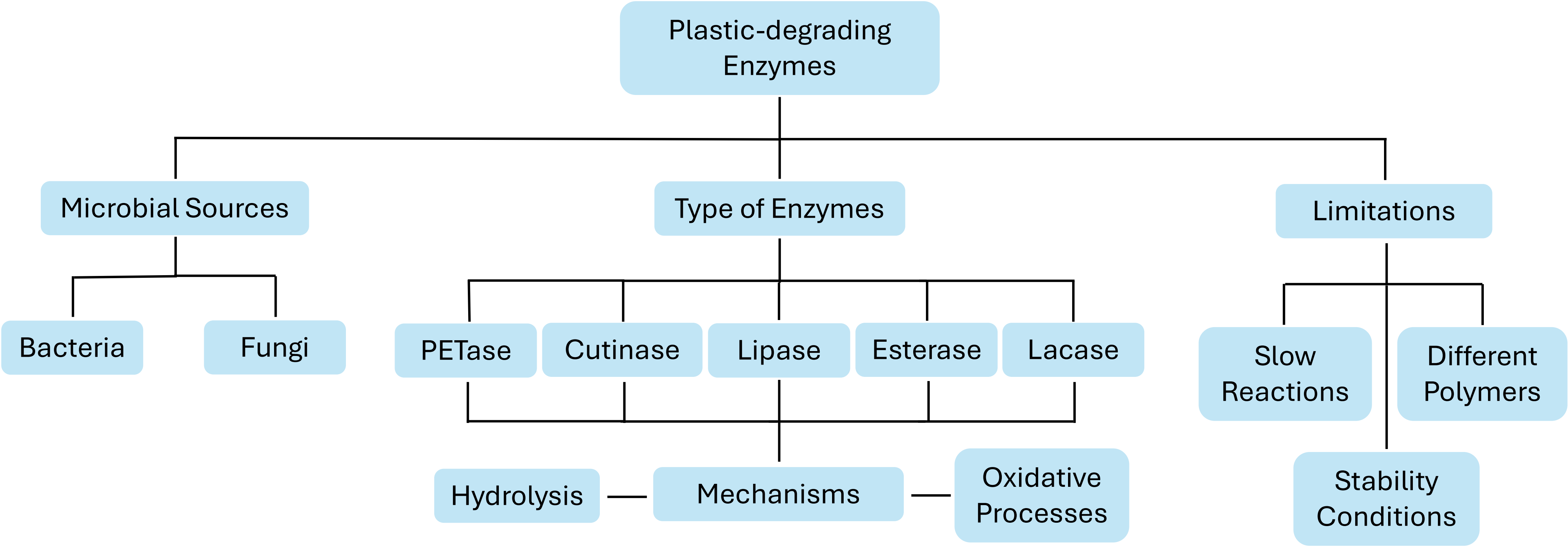
Just 9% of discarded plastic is recycled

6% of oil consumption is directed to plastic production

8 million tons of plastic ends in the ocean annually

Marine garbage patches all around the world

Enzymatic Solutions for Plastic Degradation



Plastic-degrading Enzymes

Microbial Sources

Bacteria

Fungi

Type of Enzymes

PETase

Cutinase

Lipase

Esterase

Lacase

Hydrolysis

Mechanisms

Oxidative Processes

Limitations

Slow Reactions

Different Polymers

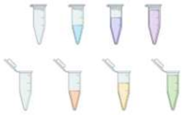
Stability Conditions

ML Approaches in Enzyme Engineering

Before ML

Classical Approach

Intensive and time-consuming research work



Try and error approach
Doesn't guarantee results

Slow process
Takes years to optimize and the optimal solution can never be reached

Few Variants Analysed

After ML

Data Oriented Strategy

ML models capable of predicting enzyme performance

AGTCCCTGAATCGATGCTACATGCAAG
AACTGATCCTAGATCGACGATCACGTG
AAACCCTTAGATGACGGTCAGCCGCAA
GCGGAATTGGCGACATAACAAGTACTG
AGTCGGCGTTCGCCTAACCGCAGTATT
TTCCAGGCCCTTCTGCCGAGCTGATGC
AGCCACAATCGGGCTGAGGGGTCTTGG
AAACGGGTGGGGTAACGCCGTGATCGC

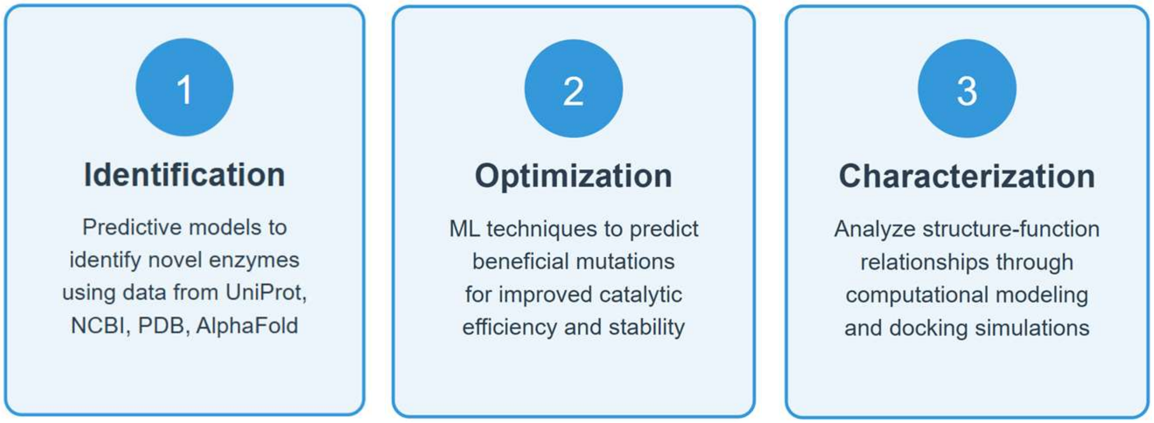
Enzyme sequence is enough to collect vital data

Faster process
Can take up until few months to optimize but everything is automated

Thousands of Variants Analysed

Main Objective

Develop and implement a machine learning framework for:
Identification, optimization and characterization of plastic-degrading enzymes



1

Identification

Predictive models to identify novel enzymes using data from UniProt, NCBI, PDB, AlphaFold

2

Optimization

ML techniques to predict beneficial mutations for improved catalytic efficiency and stability

3

Characterization

Analyze structure-function relationships through computational modeling and docking simulations

FOCUS:
Increased catalytic efficiency and substrate specificity

Specific Aims

Database Development and Knowledge Integration

- Enhance Plastizyme Database
- Integrate UniProt, NCBI Protein, PDB and AlphaFold data
- Standardize and preprocess enzyme sequence and structural data

Predictive Model Development

- Train ML models to predict enzyme functionality
- Predict enzyme function based on 3D structural features using deep learning models
- Optimize function and stability predictions for protein engineering applications

Structure-Function Relationship Analysis

- Clarify structure-function relationships via computational modelling
- Use AlphaFold3 data for precise structural representations
- Implement automated HADDOCK workflows

Integrated Computational Pipeline Development

- Develop a pipeline combining sequence analysis, structural prediction, docking and ML
- Incorporate docking parameters and biding affinity data into ML models

Specific Aims

Performance Validation and Optimization

- Evaluate models using Orange Data Mining software and suitable validation metrics
- Optimize parameters to maximize accuracy in detecting improved enzyme activity

Biotechnological Application Assessment

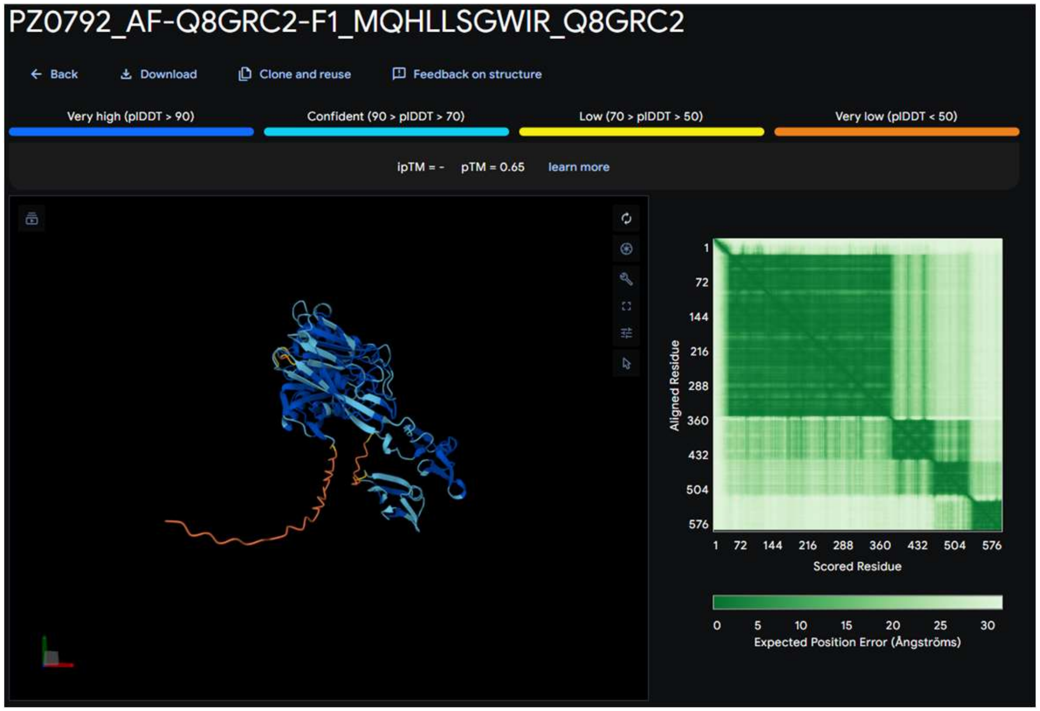
- Evaluate potential enzyme candidates for practical applications in industrial plastic recycling operations and environmental bioremediation
- Prioritize enzyme variants based on predicted efficiency, stability and feasibility for scaled production

AlphaFold3 to HADDOCK

STEP 1 – Prediction

- Used AF3 to predict the structure of 125 enzymes
- All structures had high confidence scores
- The format output was .cif

Protein structure example



PZ0792_AF-Q8GRC2-F1_MQHLLSGWIR_Q8GRC2

ipTM = - pTM = 0.65

AlphaFold3 to HADDOCK

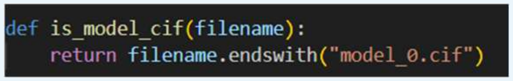
STEP 1 – Prediction

- Used AF3 to predict the structure of 125 enzymes
- All structures had high confidence scores
- The format output was .cif

Protein structure example

STEP 2 – Conversion

- Format mismatch was a challenge
- Created custom code to convert .cif files to .pdb
- Processed all 125 structures



```
def is_model_cif(filename):
    return filename.endswith("model_0.cif")
```

Implemented function

STEP 3 – HADDOCK

- Provide .pdb files to HADDOCK for docking
- Using HADDOCK output as Prodigy input to predict binding affinity
- Work in progress

Attempted *vs* Successful Approaches

Attempted

- ✗ Automated Haddock Workflow by manipulation of AptaCom code
- ✗ Docking of plastic polymers with plastic-degrading enzymes
- ✗ Prodigy binding affinity as machine learning predictor

Successful

- ✓ AlphaFold’s structure predictions for 125 enzymes
- ✓ Automated file format conversion (.CIF to .PDB and .SDF to .PDB)
- ✓ Haddock docking and Prodigy troubleshooting

Design a streamlined machine learning algorithm maintaining consistency with the initial project plan

Next Computational Pipeline Steps

- **HADDOCK Simulations** - Automated docking workflows to model enzyme-substrate interactions with various plastic polymers;
- **Feature Engineering & ML Models** - Advanced feature extraction from structural and sequence data for EnzyNet (3D features) and UniRep (sequence embeddings);
- **Orange Data Mining** - Systematic evaluation of predictive performance using established metrics and advanced visualization tools;
- **Database Enhancement** - Further development of the Plastizyme database with integrated data from UniProt, NCBI Protein, PDB, and AlphaFold.

Integrated Pipeline

Development of a seamless computational workflow combining all analyses into a unified prediction system

Acknowledgements

This research was supported by Portuguese national funds through the Foundation for Science and Technology (FCT) within the scope of UIDB/04050/2020 (UMinho CBMA – https://doi.org/10.54499/UIDB/04050/2020).

Universidade do Minho
Escola de Engenharia

cbma
centre of **molecular** and
environmental **biology**