# Enzyme_Stability Prediction

**DATA SCIENCE PROJECT**

GitHub

Michalina Hulak

Piotr Ciechowicz

Cezary Gruszka

# Introduction

In this Data Science project, we aim to predict enzyme stability using machine learning and deep learning models. Enzyme stability is an important factor in determining the performance and shelf life of industrial enzymes, which are used in a variety of applications including food processing, pharmaceuticals, and biofuels.

# Goal

Our goal is to build a predictive model that can accurately predict enzyme stability based on their pH and protein sequence.

To achieve this, we explored the data. Then, we preprocessed the data, including feature engineering and handling missing values. We also performed data visualization and statistical analysis to better understand the data.

Next, we trained and evaluated various machine learning models, such as random forests, and neural networks and we choosed the best performing model for our predictions.

# Data and feature extractions

"Feature extraction" is a technique used in machine learning to extract relevant features (i.e., informative aspects) from raw data in order to improve the performance of a model.

| seq_id | protein_sequence | pH | data_source | tm |
|---|---|---|---|---|
| 0 | AAAAKAAALALLGEAPEVVDIWLPAGWRQPFRVFRLERKGDGVLVG... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 75.7 |
| 1 | AAADGEPLHNEEERAGAGQVGRSLPQESEEQRTGSRPRRRRDLGSR... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 50.5 |
| 2 | AAAFSTPRATSYRILSSAGSGSTRADAPQVRRLHTTRDLLAKDYYA... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 40.5 |
| 3 | AAASGLRTAIPAQPLRHLLQPAPRPCLRPFGLLSVRAGSARRSGLL... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 47.2 |
| 4 | AAATKSGPRRQSQGASVRTFTPFYFLVEPVDTLSVRGSSVILNCSA... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 49.5 |

Raw data

**k-mer**

using k-mer in protein sequence involves identifying and counting all possible subsequences (k-mers) of length k in a protein sequence. Each k-mer is treated as a feature, and its frequency or occurrence is used to represent the protein sequence as a feature vector.

| seq_id | protein_sequence | pH | tm | protein_length | protein_length_normalized | kmer_3_AAA | kmer_3_AAK | ... | kmer_3_CMY | kmer_3_WMH | kmer_3_CWP | kmer_3_MCM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AAAAKAAALALLGEAPEVVDIWLPAGWRQPFRVFRLERKGDGVLVG... | 7.0 | 75.7 | 341 | 0.038212 | 3 | 1 | ... | 0 | 0 | 0 | 0 |
| 1 | AAADGEPLHNEEERAGAGQVGRSLPQESEEQRTGSRPRRRRDLGSR... | 7.0 | 50.5 | 286 | 0.031957 | 1 | 0 | ... | 0 | 0 | 0 | 0 |
| 2 | AAAFSTPRATSYRILSSAGSGSTRADAPQVRRLHTTRDLLAKDYYA... | 7.0 | 40.5 | 497 | 0.055954 | 4 | 1 | ... | 0 | 0 | 0 | 0 |

**amino acid**

using amino acids involves identifying and counting the occurrences of each of the 20 standard amino acids (represented by their one-letter codes) in the protein sequence. Each amino acid is treated as a feature, and its frequency or occurrence is used as a feature vector.

| seq_id | protein_sequence | pH | tm | protein_length | A | C | D | ... | T | V | W | Y | protein_length_normalized |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AAAAKAAALALLGEAPEVVDIWLPAGWRQPFRVFRLERKGDGVLVG... | 7.0 | 75.7 | 341 | 0.131965 | 0.002933 | 0.038123 | ... | 0.041056 | 0.108504 | 0.011730 | 0.008798 | 0.038212 |
| 1 | AAADGEPLHNEEERAGAGQVGRSLPQESEEQRTGSRPRRRRDLGSR... | 7.0 | 50.5 | 286 | 0.097902 | 0.000000 | 0.034965 | ... | 0.041958 | 0.045455 | 0.010490 | 0.010490 | 0.031957 |
| 2 | AAAFSTPRATSYRILSSAGSGSTRADAPQVRRLHTTRDLLAKDYYA... | 7.0 | 40.5 | 497 | 0.100604 | 0.018109 | 0.054326 | ... | 0.060362 | 0.060362 | 0.006036 | 0.032193 | 0.055954 |

# Data and feature extractions

"Feature extraction" is a technique used in machine learning to extract relevant features (i.e., informative aspects) from raw data in order to improve the performance of a model.

| seq_id | protein_sequence | pH | data_source | tm |
|---|---|---|---|---|
| 0 | AAAAKAAALALLGEAPEVVDIWLPAGWRQPFRVFRLERKGDGVLVG... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 75.7 |
| 1 | AAADGEPLHNEEERAGAGQVGRSLPQESEEQRTGSRPRRRRDLGSR... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 50.5 |
| 2 | AAAFSTPRATSYRILSSAGSGSTRADAPQVRRLHTTRDLLAKDYYA... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 40.5 |
| 3 | AAASGLRTAIPAQPLRHLLQPAPRPCLRPFGLLSVRAGSARRSGLL... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 47.2 |
| 4 | AAATKSGPRRQSQGASVRTFTPFYFLVEPVDTLSVRGSSVILNCSA... | 7.0 | doi.org/10.1038/s41592-020-0801-4 | 49.5 |

Raw data

**ESM (Evolutionary Scale Modeling)**

ESM (Evolutionary Scale Modeling) model is a deep learning model designed to predict protein properties based on their amino acid sequences. It transform amino acid sequences into a tensor that can be used as input to the model. We used: model_name = 'esm2_t33_650M_UR50D'

| | seq_id | embedding | tm | protein_length_normalized | pH_normalized | isoelectric_point_normalized |
|---|---|---|---|---|---|---|
| 0 | 0 | [tensor(0.0705), tensor(-0.1038), tensor(-0.00... | 75.7 | 0.038212 | 0.556049 | 0.259066 |
| 1 | 1 | [tensor(0.0653), tensor(-0.0039), tensor(-0.04... | 50.5 | 0.031957 | 0.556049 | 0.137651 |
| 2 | 2 | [tensor(-4.7959e-05), tensor(-0.0171), tensor(... | 40.5 | 0.055954 | 0.556049 | 0.626681 |

# Results

## k-mer

| | Model | MSE_train | MSE_val | MSE_test | MAE_train | MAE_val | MAE_test | R^2_train | R^2_val | R^2_test |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | RandomForest | 14.024573 | 91.734142 | 89.817641 | 2.625676 | 6.733020 | 6.712207 | 0.905334 | 0.372586 | 0.350883 |
| 1 | CatBoostRegressor | 15.797424 | 75.840777 | 75.214817 | 2.924441 | 6.501065 | 6.440109 | 0.893367 | 0.481288 | 0.456418 |

## amino-acid

| | Model | MSE_train | MSE_val | MSE_test | MAE_train | MAE_val | MAE_test | R^2_train | R^2_val | R^2_test |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | RandomForest | 9.959821 | 57.940179 | 60.131315 | 2.225169 | 5.607841 | 5.700010 | 0.934890 | 0.618301 | 0.587339 |
| 1 | ElasticNetCV | 118.333692 | 116.415428 | 117.043314 | 8.284225 | 8.202961 | 8.325385 | 0.226416 | 0.233077 | 0.196770 |
| 2 | XGBRegressor | 1.817231 | 59.491060 | 63.727129 | 0.244239 | 5.641808 | 5.787173 | 0.988120 | 0.608084 | 0.562662 |
| 3 | CatBoostRegressor | 4.993440 | 55.364115 | 58.387872 | 1.400873 | 5.536607 | 5.654937 | 0.967356 | 0.635271 | 0.599303 |
| 4 | LGBMRegressor | 4.203817 | 58.723742 | 60.195176 | 1.016494 | 5.661891 | 5.720341 | 0.972518 | 0.613139 | 0.586900 |
| 5 | RidgeCV | 111.819371 | 108.546811 | 112.397011 | 8.203759 | 8.075496 | 8.299346 | 0.269002 | 0.284914 | 0.228656 |

## ESM

| | Dataset | MAE | MSE | R^2 |
|---|---|---|---|---|
| 0 | Train | 5.023776 | 45.969743 | 0.699482 |
| 1 | Val | 5.281133 | 48.819488 | 0.678386 |

```
Test MAE: 5.3416
Test MSE: 52.2098
Test R^2: 0.6417
```

# Conclusions

1. For the models trained on k-mer features, both RandomForest and CatBoostRegressor showed moderate performance. However, there is a clear indication of overfitting as the models performed significantly better on the training set compared to the validation and test sets. The R-squared values dropped considerably from the training set to the validation and test sets, suggesting limited generalization ability.

2. In contrast, the models trained on amino acid features exhibited better performance overall. RandomForest, XGBRegressor, CatBoostRegressor, and LGBMRegressor achieved lower MSE and MAE values on the validation and test sets compared to the k-mer models. These models also demonstrated higher R-squared values, indicating better fit and higher explained variance. However, ElasticNetCV and RidgeCV models performed poorly, with high MSE and MAE values and low R-squared values.

3. Regarding the issue of overfitting, there is evidence of overfitting in the k-mer models and amino acid models as indicated by the large discrepancy between the performance on the training set and the validation/test sets.

4. The PyTorch model (ESM) shows promising performance for predicting protein thermostability. The relatively low MAE and MSE values and the moderate R^2 score suggest that the model captures a significant portion of the variability in the data and makes accurate predictions.