FHDW

Fachhochschule
der *Wirtschaft*

Bachelorthesis

A Deep Learning Approach for Predicting Pesticide Degradation Based on Enzyme Classes

Examiner:
Prof. Dr. Thomas Ströder
Fethi Temiz

Author:
Tobias Polley
100853

Düsselthaler Str. 20
40211 Düsseldorf

BFWC321B
Cyber Security

Submitted on:
July 5, 2024

Restriction Notice

This work contains confidential information about the company Bayer AG. The disclosure of the contents of this work (even in part) is prohibited. No copies or transcripts - not even in digital form - may be made. This thesis may also not be published and may only be made accessible to the examiners, administrative staff and members of the examination committee and, on request, to an evaluation committee. Persons who gain access to this thesis undertake not to disclose any information concerning the company Bayer AG to third parties via the contents of this thesis and all its appendices. Exceptions require the written permission of the company Bayer AG and the author.

The thesis or parts thereof may be subjected to a plagiarism check by the FHDW using a plagiarism software provider. The blocking notice is therefore not effective in the event of a plagiarism check.

Contents

List of Figures

List of Tables

Listing Directory

1 Introduction

1.1 Motivation

In recent years, the prediction of pesticide degradation has gained significant attention due to its environmental and health impacts. Traditional methods for determining the degradation of enzymes are labor-intensive and time-consuming. Consequently, there is a growing need for computational methods that can efficiently and accurately predict the degradation behavior of pesticides.

The pesticide degradation can be facilitated by enzymes, which are biological catalysts that accelerate chemical reactions. Enzymes play a crucial role in breaking down pesticides into harmless byproducts, reducing their toxicity and environmental impact. Understanding the enzymatic mechanisms involved in pesticide degradation is essential for developing sustainable and environmentally friendly agricultural products.

Recent advancements in DNA and RNA sequencing technologies have led to an explosion of information about new organisms and their enzymes. These developments are crucial for the industry and, in particular, for Bayer, as they provide a wealth of data that can be leveraged to enhance the prediction and understanding of enzymatic functions involved in pesticide degradation. This influx of sequence data from diverse ecosystems, including soil, offers new opportunities to identify enzymes that play significant roles in environmental processes.

The applications of advanced computational methodologies, such as Deep Learning for predicting these enzymatic functions, have the potential to significantly enhance the development of environmentally friendly and safe agricultural products at Bayer Crop Science. This would also reduce the time and cost of testing existing and new products, as well as the risk of developing products that are harmful to the environment. Despite significant advancements in bioinformatics and computational biology, predicting enzyme classes remains challenging and fraught with uncertainties. Traditional methods rely heavily on experimental data, which can be resource-intensive and time-consuming. Moreover, the vast diversity of enzyme functions and their complex interactions with various substrates add layers of difficulty to accurate predictions. Thus, there is a pressing need for computational tools that leverage modern machine learning techniques to enhance the prediction accuracy of enzyme-related models.

This thesis introduces a novel Deep Learning model that leverages enzyme binding site predictions to enhance the accuracy and efficiency of enzyme class prediction. Unlike existing models, the approach specifically targets enzyme binding sites, offering a more detailed and accurate prediction by focusing on the critical interaction regions.

By addressing this research question, the study seeks to contribute to the fields of computational biology and environmental science, providing a tool that can accurately predict enzymatic functions and their behavior in pesticide degradations. In addition, this research aims to outperform existing models in predicting enzyme classes responsible for pesticide degradation, thereby enhancing the accuracy of enzyme classification predictions. The findings of this study could have significant implications for the development of environmentally friendly and sustainable agricultural products, as well as the reduction of harmful pesticides in the environment.

1.2 Structure of the Thesis

This thesis is structured into five chapters, each addressing different aspects of the research and providing a comprehensive overview of the study. The first chapter sets the stage for the entire thesis. It begins by outlining the motivation behind the research, highlighting the environmental concerns related to pesticide use and the need for effective degradation prediction methods. The problem statement section identifies the challenges associated with predicting enzyme-mediated pesticide degradation. The introduction section defines the main objective of the study, which is to develop a Deep Learning model to predict pesticide degradation based on enzyme classes. Finally, this chapter provides an overview of the structure of the thesis.

The literature review chapter delves into existing research and foundational theories relevant to the study. It covers enzymatic mechanisms involved in pesticide breakdown, offering insights into how enzymes facilitate the degradation process. Additionally, it explores the application of Deep Learning techniques in environmental science, emphasizing their potential to enhance predictive accuracy. The chapter concludes with a discussion of the limitations of current models and the need for more advanced approaches to enzyme classification.

The methodology chapter provides a detailed description of the research design and procedures followed in this study. It begins with the Data Collection, specifying the sources and preprocessing steps to prepare the dataset for analysis. The feature engineering section discusses how relevant features were extracted from the data to calculate accurate predictions. The chapter then explains the model development process, including the architecture of the Deep Learning model and the final training process.

The results chapter presents the outcomes of the research. It begins with an evaluation of the model's performance, highlighting key metrics and the effectiveness of the model in predicting pesticide degradation. A comparative analysis with existing models is included to demonstrate the improvements and advantages of the developed model. The chapter also interprets the model predictions, offering insights into the practical implications of the findings and how they can be applied in real-world scenarios.

The discussion chapter summarizes the key findings of the research, reflecting on the significance and impact of the results. It discusses the strengths and limitations of the study, acknowledging areas where the model performed well and identifying potential areas for improvement. The chapter concludes with an overview of the contributions to the field, highlighting the novelty and practical applications of the research. Additionally, it provides recommendations for future work, suggesting directions for further research to build on the findings of this study.

2 Related Work and State of the Art

2.1 Role of Enzymes in Environmental Pesticide Degradation

The prediction of pesticide degradation and the identification of enzymatic functions involved is a critical area of research with significant implications for environmental sustainability and agricultural practices. As the use of pesticides continues to be a global necessity for crop protection, understanding the mechanisms by which these chemicals are broken down in the environment is essential.

The degradation of pesticides in the environment is a complex process that occurs through various mechanisms, predominantly by microbial enzymatic activities. Enzymes catalyze reactions that transform toxic pesticide compounds into less harmful substances. The most common enzymatic mechanisms involved in pesticide degradation include hydrolytic and oxidative reactions.

For example, reductive enzymes catalyze the reduction of pesticides, in most cases by donating electrons and hydrogen atoms on the molecules. Such a reduction may well break complex structures that facilitate the conversion of pesticides into simpler forms that are much less toxic. For instance, reductive dehalogenases are believed to be important in breaking down halogenated organic compounds.

The application of microbial enzymes in bioremediation strategies has significantly improved the degradation of pesticides in contaminated soils. This approach leverages the natural capability of microbes to detoxify pollutants through enzymatic reactions. Studies have established the efficiency of microbial enzymes in degrading soil-contaminated pesticides. For instance, Singh and Walker [1] highlighted the effectiveness of microbial degradation of organophosphorus compound, while Chia et al. [2] discussed advancements and applications of microbial enzymes in enhancing biodegradation processes.

Understanding these enzymatic mechanisms is crucial for predicting the enzyme classes responsible for pesticide degradation. By analyzing enzyme-pesticide interactions, it is possible to identify specific enzyme classes involved in the degradation processes. This knowledge can inform the development of more accurate predictive models for pesticide degradation, facilitating better risk assessments and environmental management strategies. Advanced computational methods, such as Deep Learning, can further enhance these predictive models by accurately identifying and classifying enzymes based on their interaction with pesticides, leading to more efficient and targeted development of new products.

2.2 Deep Learning Techniques in Environmental Science

Deep Learning has been an essential tool in environmental science, enabling advanced prediction and understanding complex biochemical processes. There are several Deep Learning architectures such as the protein-transformer ESM model, which has made a significant impact on predicting biological properties from sequence data. [3]

In the context of pesticide degradation and enzyme classification, such models can analyze large quantities of available biochemical data to make predictions about enzyme interactions and functions. Several deep learning architectures have been applied in enzyme classification and prediction tasks, from which valuable insights into the mechanism of pesticide degradation can be obtained.

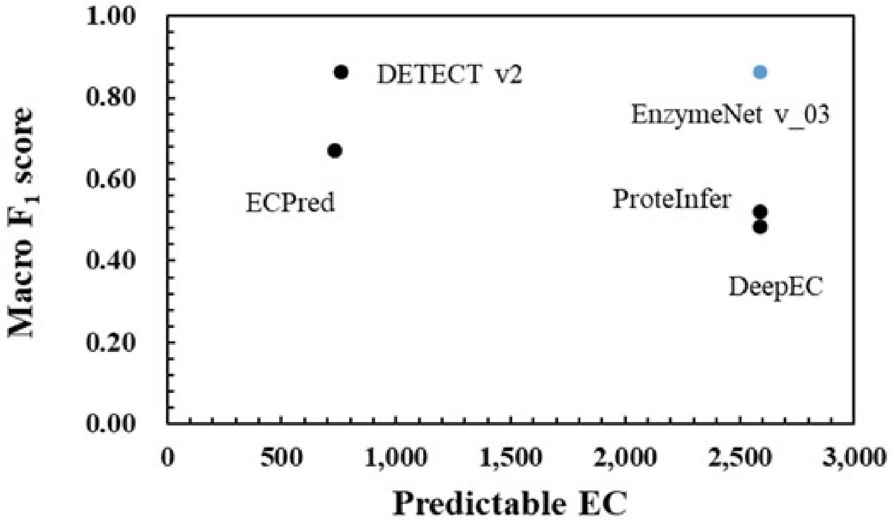
For instance, the DEEPre model applies deep learning to predict EC numbers based on raw sequence data. Such models apply convolutional and sequential feature extraction techniques, leading to significant improvements in prediction accuracy over methods in current use. In this respect, such models may play a key role in predicting the pesticide biodegradation pathways and help to make environmental risk assessment more precise and fast. [4]

The DeEPn model is one of the examples when EC classification has been done using a deep neural network for enzymes being classified into their functional classes, including all seven EC classes. This model has shown high precision and accuracy and, hence could become an essential tool for environmental scientists interested in understanding and predicting enzyme-mediated degradation processes.

The proper classification of enzymes through DeEPn can help predict potential candidates for bioremediation, among other applications related to the environment. [5]

Despite the advances made by these models, there is still a need for new approaches to further improve the accuracy of sequence based predictions. Traditional models often rely on pre-defined features and limited datasets, which can restrict their performance and generalizability. In addition to this, the existing methods only focus on the prediction to the 3rd level of the EC classification, which may not provide sufficient detail for predicting pesticide degradations.

Figure 1: Macro F1 score for different models and EnzymeNet



Source: Watanabe et al. (2023)

For example the accuracy of EnzymeNet, a residual neural network model, across all the sub-subclasses is 0.398. In addition to that there is no score for the 4th level. The following picture shows the macro F1 score for diffrent models and EnzymeNet, which is the best one, but still not good enough. Therefore, there is a need for more advanced deep learning models that can predict enzyme classes with higher accuracy and resolution, enabling more precise predictions of pesticide degradation pathways. [6]

By contrast, the proposed approach leverages the deep learning tool p2rank to analyze the interactive parts of enzymes, focusing on the ligand-binding sites and the specific amino acids involved. This method can potentially provide a more detailed and accurate prediction of enzyme classes responsible for pesticide degradation, enhancing our understanding of the biodegradation pathways and mechanisms involved. Furthermore, the emphasis on ligand-binding pockets allows for a more nuanced analysis compared to traditional methods that utilize the entire protein sequence. By concentrating on these critical interaction sites, which are crucial for protein functions, P2Rank can identify the specific residues that are directly involved in the catalytic processes. This specificity could not only improves the accuracy of predictions but also reduce the computational complexity by focusing on smaller, more relevant regions of the protein. [7]

3 Theoretical Background

3.1 Principles of Enzymology

Enzymology is the scientific study of enzymes, which are biological catalysts that accelerate biochemical reactions in living organisms. These macromolecules are essential for various cellular processes, including metabolism, DNA replication, and signal transduction. The understanding of enzyme structure, function, and kinetics is crucial for developing applications in biotechnology, medicine, and environmental science. [8]

Enzymes are classified based on the types of reactions they catalyze, according to a system established by the Enzyme Commission (EC). This classification system groups enzymes into six main classes, each with specific types of reactions they facilitate:

1. **Oxidoreductases:** These enzymes catalyze oxidation-reduction reactions, where the transfer of electrons occurs between molecules. Examples include dehydrogenases and oxidases.
2. **Transferases:** These enzymes transfer functional groups from one molecule to another. Examples include kinases, which transfer phosphate groups.
3. **Hydrolases:** These enzymes catalyze the hydrolysis of various bonds, including ester, glycosidic, peptide, and others. Examples include proteases and lipases.
4. **Lyases:** These enzymes add or remove groups to form double bonds, without hydrolysis or oxidation. Examples include decarboxylases and dehydratases.
5. **Isomerases:** These enzymes catalyze the rearrangement of atoms within a molecule, leading to isomerization. Examples include racemases and epimerases.
6. **Ligases:** These enzymes catalyze the joining of two molecules with the simultaneous hydrolysis of a diphosphate bond in ATP or a similar triphosphate. Examples include synthetases and carboxylases.

For example, the enzyme tripeptide aminopeptidase has the EC number "3.4.11.4", where the first digit (3) represents the class (Hydrolases in this case), the second digit (4) represents the subclass (hydrolases that act on peptide bonds), the third digit (11) represents the sub-subclass (Hydrolases that cleave off the amino-terminal amino acid from a polypeptide), and the fourth digit (4) represents the serial number of the enzyme within the sub-subclass (Hydrolases that cleave off the amino-terminal end from a tripeptide). This systematic classification allows researchers to identify enzymes based on their catalytic activities and biochemical properties. [9] The distribution of EC numbers across the six classes is not uniform, with hydrolases being the most abundant class, reflecting the importance of hydrolysis in biological processes. The following figure shows the distribuion of EC numbers across all four levels of the classification system:

EC Level	**Number of classes**
1	7
2	79
3	320
4	7876

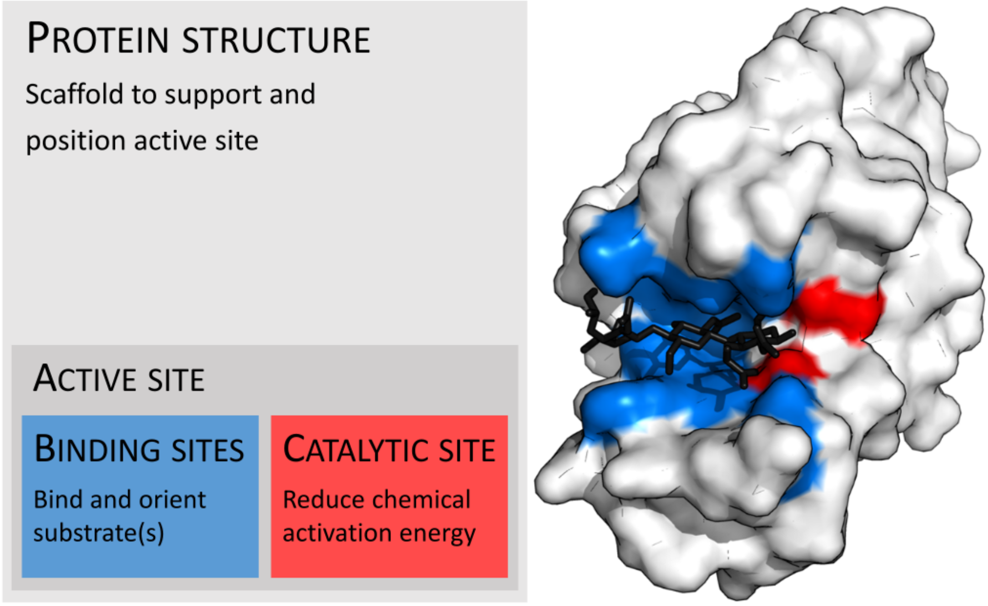
Table 1: Distribution of EC numbers across different levels of the classification system.

Enzymes are not only classified based on their catalytic activities but also based on their biological functions. The three-dimensional (3D) structure of enzymes is fundamental to their function. Enzymes are composed of one or more polypeptide chains that fold into specific shapes to form the active site. The active site is where substrate molecules bind and undergo a chemical reaction. The enzyme structure serves as a scaffold to support and correctly position the active site for optimal catalytic activity.

- **Protein Structure:** The overall structure of the enzyme provides the framework that supports and positions the active site. This structure is critical for the enzyme's stability and functionality. The enzyme's polypeptide chains fold into a unique 3D shape, creating a specific environment for the active site.

- **Active Site:** The active site includes two critical regions: binding sites and the catalytic site. The binding sites (highlighted in blue) are regions where

Figure 2: Organisation of enzyme structure and lysozyme example.


PROTEIN STRUCTURE
Scaffold to support and position active site
ACTIVE SITE
BINDING SITES
Bind and orient substrate(s)
CATALYTIC SITE
Reduce chemical activation energy

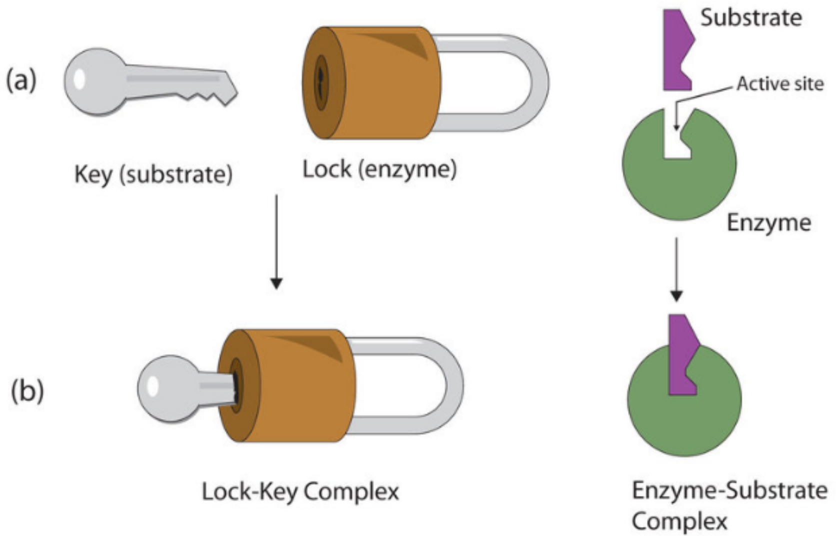
Source: Thomas Shafee, CC BY 4.0 via Wikimedia Commons

substrates bind to the enzyme. These sites ensure that the substrates are properly oriented for the reaction. The catalytic site (highlighted in red) is the region where the chemical reaction occurs. The catalytic site often contains amino acids with specific functional groups that participate directly in the reaction, reducing the activation energy required for the reaction to proceed.

The Key-Lock Principle, first proposed by Emil Fischer in 1894, is a model for understanding the specificity of enzyme-substrate interactions. According to this principle, the enzyme (lock) has a specific active site shape that only fits a particular substrate (key). This model emphasizes the specificity of enzyme-substrate interactions and how enzymes are highly selective for their substrates. This principle is fundamental to understanding enzyme function and the mechanisms of catalysis. The specificity of these interactions is crucial for predicting enzyme activities because it determines the substrates that can bind to the enzyme and undergo catalysis.

The precise arrangement of amino acids in the active site allows enzymes to be

Figure 3: Lock-and-key model that explains the selectivity of enzymes



Source: [10]

highly specific for their substrates, facilitating efficient catalysis. This specificity is a key feature that enables enzymes to perform their roles in various biochemical pathways with high precision. Understanding the structure-function relationship of enzymes is essential for predicting their activities.

3.2 Fundamentals of Ligand Binding Site Prediction

As mentioned earlier, enzymes interact with substrates at specific binding sites, where the catalytic reactions occur. Predicting these ligand-binding sites is crucial for understanding the enzyme function. Several computational methods have been developed to predict ligand-binding sites from protein structures, including geometric, physicochemical, and machine learning-based approaches.

One approach is P2Rank, a machine learning-based tool designed for the rapid and accurate prediction of ligand binding sites from protein structures. It employs a combination of geometric and physicochemical descriptors to analyze protein structures and predict the locations of potential binding sites. P2Rank uses a random forest algorithm, an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.

The tool focuses on the interactive parts of enzymes, particularly the ligand-binding sites and the specific amino acids involved. This detailed analysis allows for accurate predictions of enzyme classes and their associated degradation pathways. P2Rank's ability to quickly and accurately predict binding sites makes it a valuable tool for drug discovery and environmental bioremediation applications.

P2Rank leverages local chemical neighborhood features near the protein surface to infer potential binding sites for ligands. Here is an overview of the key steps involved in the P2Rank prediction process:

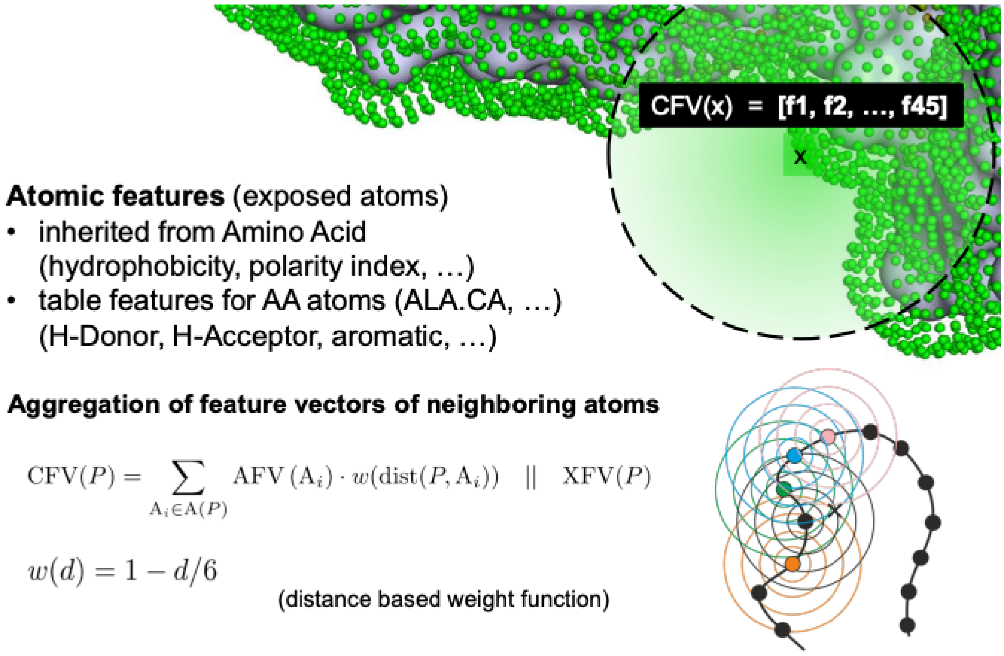
1. **Generation of Connolly Points**: Connolly Points are regularly spaced points generated on the protein's Connolly surface, representing the solvent-accessible surface area of the protein. These points are generated using a numerical algorithm that ensures even spacing, typically with a solvent radius of 1.6 Å.

2. **Calculation of Feature Descriptors**: Atomic Feature Vectors (AFVs) are calculated for each solvent-exposed heavy atom in the protein, describing various physico-chemical properties such as hydrophobicity, aromaticity, and more. These properties are projected onto the Connolly points using a distance-weighted approach, creating Connolly Feature Vectors (CFVs) for each point. The image shows Connolly Points (green dots) on the protein's surface, where each point is associated with a Connolly Feature Vector (CFV).

    a) Atomic Features: Features are inherited from the amino acid, including properties like hydrophobicity and polarity index. Additional features for AA atoms include H-Donor, H-Acceptor, and aromaticity.

    b) Aggregation of Feature Vectors: The CFV for each Connolly point is calculated by aggregating the AFVs of neighboring atoms using a distance-based weight function $w(d) = 1 - d/6$.

Figure 4: Calculation of feature vectors for neighboring atoms in P2Rank.



Atomic features (exposed atoms)

- inherited from Amino Acid (hydrophobicity, polarity index, ...)
- table features for AA atoms (ALA.CA, ...) (H-Donor, H-Acceptor, aromatic, ...)

Aggregation of feature vectors of neighboring atoms

$$\mathrm{CFV}(P) = \sum_{\mathrm{A}_i \in \mathrm{A}(P)} \mathrm{AFV}\,(\mathrm{A}_i) \cdot w(\mathrm{dist}(P, \mathrm{A}_i)) \quad || \quad \mathrm{XFV}(P)$$

$$w(d) = 1 - d/6$$

(distance based weight function)

Source: Radoslav Kriván and David Hoksza

3. **Ligandability Prediction**: A Random Forest classifier is used to predict the ligandability score for each Connolly point, indicating the likelihood that a point is part of a ligand-binding site.

4. **Clustering**: Connolly points with high ligandability scores are clustered using a single-linkage clustering method, representing potential binding pockets on the protein surface.

5. **Ranking**: Each predicted pocket is assigned a score based on the cumulative ligandability scores of its constituent points, helping prioritize the most likely binding sites for further analysis or docking studies.

P2Rank's approach can significantly enhance the accuracy of predicting enzyme-mediated degradation of pesticides by providing detailed insights into the binding interactions at the molecular level. This integration of deep learning and enzyme analysis forms a robust framework for developing bioremediation strategies and understanding the environmental fate of various pollutants. [7]

3.3 Introduction to Recurrent Neural Networks

RNNs (RNNs) are a class of artificial neural networks designed to recognize patterns in sequences of data such as text, genomes, handwriting, and spoken words. Unlike traditional feedforward neural networks, RNNs have connections that form directed cycles, allowing information to persist. This makes them particularly powerful for tasks that involve sequential data, where the order of the data points matters. [11]

RNNs are designed to process sequences of data by maintaining a memory of previous inputs. This memory allows RNNs to make use of information from earlier in the sequence to influence the current processing step, which is essential for understanding context in sequential data. The fundamental difference between RNNs and traditional neural networks is the presence of loops in the network that enable the persistence of information across time steps.

RNNs are designed to process sequences of data by maintaining a memory of previous inputs. This memory allows RNNs to make use of information from earlier in the sequence to influence the current processing step, which is essential for understanding context in sequential data. The fundamental difference between RNNs and traditional neural networks is the presence of loops in the network that enable the persistence of information across time steps.

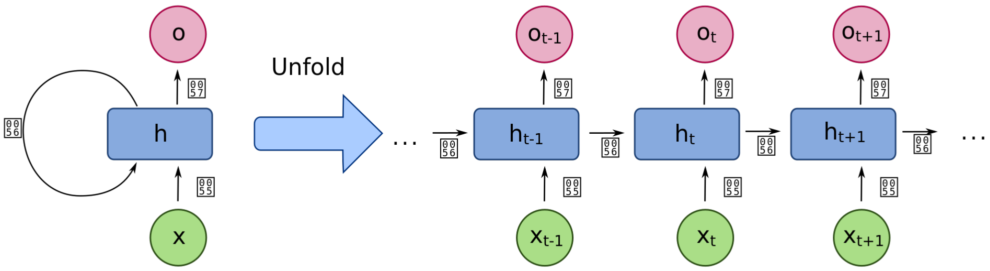
The basic structure of an RNN includes an input layer, a hidden layer with recurrent connections, and an output layer. At each time step, the hidden layer receives the input data and its own previous state, allowing it to retain and process information from previous steps in the sequence.

One of the key advancements in RNNs is the development of Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), which are designed to overcome the limitations of traditional RNNs, such as the vanishing gradient problem. These architectures use gating mechanisms to control the flow of information, making it easier to capture long-term dependencies in data.

In the context of bioinformatics, RNNs, particularly LSTMs and GRUs, are extensively used for sequence analysis tasks such as protein secondary structure prediction, gene expression analysis, and more. They are effective because they can handle the sequential nature of biological data and capture dependencies that span over long sequences. LSTM networks are a type of RNN that can learn long-term dependencies. They incorporate memory cells that can maintain their state over long periods. LSTMs have three main gates (input gate, forget gate, and output gate) that regulate the flow of information into and out of the memory cell, thus enabling the network to remember important information for longer durations. [12]

The following image illustrates the basic structure of an RNN:

Figure 5: A diagram for a one-unit RNN.



Source: fdeloche, CC BY-SA 4.0 via Wikimedia Commons

1. Input Sequence (x): The green circles represent the input data at different time steps (x_{t-1}, x_t, x_{t+1}).

2. Hidden State (h): The blue rectangles represent the hidden state of the network. At each time step, the hidden state (h) is updated based on the current input and the previous hidden state (h_{t-1}, h_t, h_{t+1}).

3. Output Sequence (o): The pink circles represent the output of the network at each time step (o_{t-1}, o_t, o_{t+1}).

The recurrent connection (arrow looping back) in the hidden state allows information to persist across time steps, enabling the network to maintain context and capture dependencies in the sequence data.

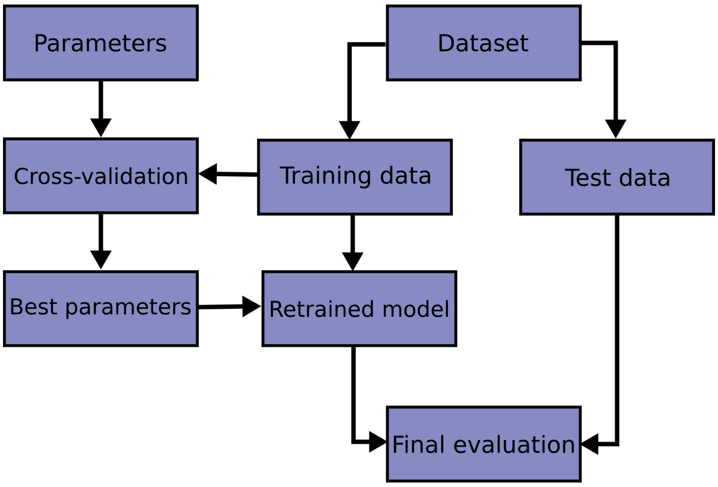
In this study, RNNs are employed for predicting the enzyme class based on the amino acid sequences of a ligand binding site. The sequential nature of the amino acid sequences makes RNNs well-suited for this task, as they can capture the dependencies and patterns in the data that are crucial for predicting enzyme classes accurately. Especially for complex and long sequences, RNNs, particularly LSTMs, are effective in learning the underlying structure and relationships in the data.

3.4 Evaluation of Deep Learning Models

When developing a deep learning model, it is crucial to compare its performance with other models and make necessary adjustments. This process ensures that the selected model is not only the best fit for the current dataset but also generalizes well to new data. By evaluating multiple models, it is possible to identify which model architecture and hyperparameters yield the best performance. This can be done through adjusting hyperparameters such as learning rate, batch size, and network depth to optimize model performance. Techniques like dropout, L2 regularization, and batch normalization can also be used to prevent overfitting and improve model generalizability. To access the generalizability of the model, it is essential to evaluate its performance on unseen data, typically through a validation set or cross-validation. Cross-Validation is a widely used technique for evaluating model performance by splitting the dataset into multiple subsets, training the model on different subsets, and testing it on the remaining data. This process helps assess the model's performance across different data partitions and provides a more robust estimate of its generalization capabilities.

The best-known cross-validation method is k-fold. he data is divided into k equally sized folds. The model is trained k times, each time using a diffrent fold as the validation set and the remaining $k-1$ folds as the training set. Cross-validation helps in understanding the variability of model performance and reduces the risk of overfitting. Repeating cross-validation multiple times, known as repeated cross-validation, can further enhance reliability. [13]

Figure 6: Cross-validation workflow for evaluating deep learning models.



Source: scikit-learn Documentation

Grid Search is a systematic way of tuning hyperparameters to find the optimal set that maximizes model performance. It involves defining a grid of possible hyperparameter values and exhaustively training and evaluating the model for each combination. For each combination of hyperparameters, the model is trained using the training data. The performance is than evaluated with a cross-validation for each hyperparameter combination. The hyperparameters that yield the best performance are selected as the optimal set. Grid Search is a powerful tool for fine-tuning deep learning models and optimizing their performance. [14]

To effectively evaluate and compare models, various metrics are used: [15] [16] [17]

Accuracy: Accuracy measures the proportion of correctly predicted instances out of the total instances. It is a straightforward metric but may not be reliable for imbalanced datasets.

$$Accuracy = \frac{\text{Total Number of Predictions}}{\text{Number of Correct Predictions}} \tag{1}$$

Precision: Precision indicates the accuracy of positive predictions, reflecting how many predicted positive instances are actually positive.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

Recall: Recall measures the model's ability to identify all relevant instances, showing the proportion of actual positives correctly identified.

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

F1 Score: The F1 Score is the harmonic mean of precision and recall, providing a balanced measure when there is an uneven class distribution.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4 Methodology

4.1 Data Collection

Data collection is a crucial step in model building, as the quality and relevance of the data directly impact the model's performance.. The data for this thesis was collected from UniProt, a comprehensive resource that provides detailed protein sequence and functional information. UniProt or Universal Protein Resource is a central protein sequence, and annotation database. It is widely accepted as comprehensive and provides high-quality data, which makes it a must to perform bioinformatics and computational biology. UniProt pools in much valuable information: experimental findings, different kinds of analyses, and literature information and hence provides rich and reliable sources of further research. Researchers can receive high-quality reviewed entries with 3D structural data and catalytic properties in place, which makes the data reliable and applicable for predicting enzyme functions. [18]

Some of the main features of UniProt are:

1. Comprehensive Protein Data: UniProt contains a vast collection of protein sequences, functional annotations, and cross-references to other databases, making it a valuable resource for protein research.

2. Reviewed Entries: UniProt contains both reviewed (Swiss-Prot) and unreviewed (TrEMBL) entries. Reviewed entries are manually curated by experts, ensuring high accuracy and reliability.

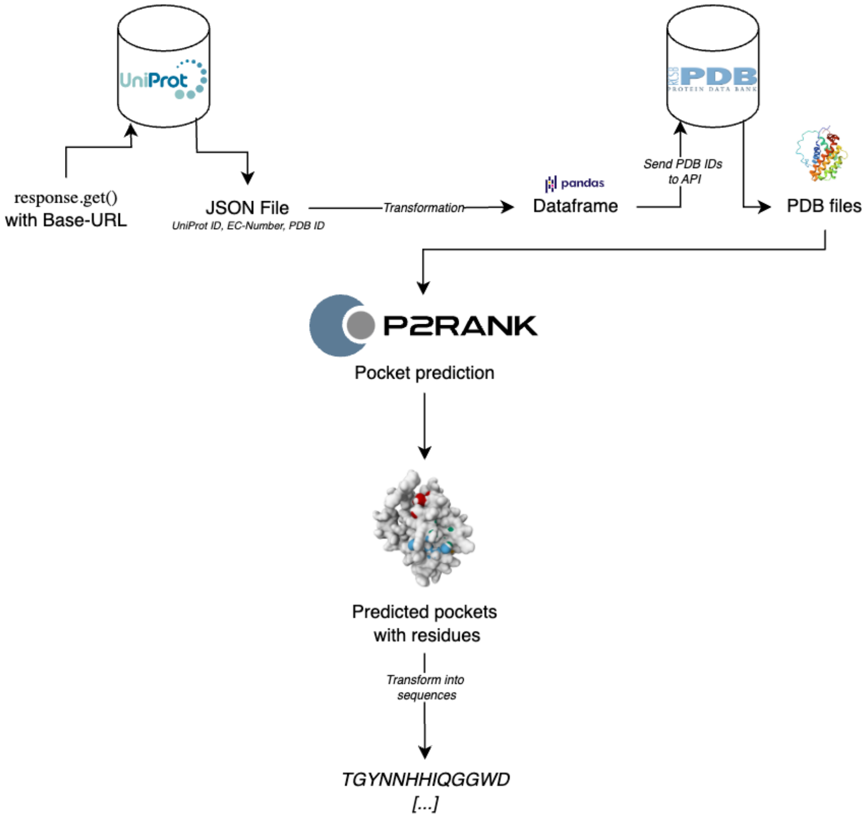
3. Functional Annotations: Each protein entry includes detailed functional annotations, such as catalytic activity, biological processes, and involvement in pathways.

4. 3D Structural Data: UniProt links to structural databases like PDB, providing access to 3D structures of proteins, which are crucial for understanding enzyme mechanisms.

5. Cross-references: Extensive cross-references to other databases (e.g., PDB, BRENDA, Reactome) enhance the richness of the data.

For this study, UniProt was chosen due to its high-quality data, extensive coverage of protein information, and user-friendly interface. The data collection process involved querying UniProt for enzyme entries with 3D structural data and catalytic activity annotations, extracting relevant information, and preprocessing the data for model development. The data retrieval process utilized the UniProt REST API to download protein data that met specific criteria. The criteria included reviewed entries with both 3D structural data and catalytic properties. The following figure illustrates the data collection process:

Figure 7: Data Collection & P2Rank Process



Source: Own illustration

1. API Request: The script constructs a query to the UniProt REST API to

retrieve reviewed protein entries with specified fields and criteria.

2. Data Retrieval: Data is retrieved and transformed into a pandas DataFrame for further processing.
3. Data Filtering: The DataFrame is filtered to retain entries with non-null EC numbers and PDB codes.
4. PDB Download: The PDB files corresponding to the protein entries are downloaded from the PDB database using the PDB IDs.
5. P2Rank Prediction: The P2Rank workflow is applied to the PDB files to predict interactive site residues.
6. Data Integration: The interactive site predictions are integrated with the protein sequences for further analysis.

4.2 Data Preprocessing

After data collection, the next step is to preprocess the data to make it suitable for model training. Data preprocessing involves several critical steps to prepare the dataset for the prediction model. These steps include data retrieval, cleaning, transformation, integration, and normalization. The goal of data preprocessing is to ensure that the data is clean, consistent, and suitable for training the model.

After collecting the sequences for every enzyme based on the Ligand-Binding-Site prediction, the sequences are cleaned by removing any non-standard amino acids, special characters, or gaps. [19] This step ensures that the sequences contain only valid amino acid residues, which is crucial for accurate modeling.

The cleaned sequences are then tokenized and encoded into numerical data for input into the model. This involves converting each amino acid into a unique integer identifier. The sequences are also padded to ensure they all have the same length, which is necessary for batch processing in Deep Learning models. Tokenization and padding allow the model to handle sequences of varying lengths and ensure uniform input size for the neural network. [20]

For gaining a first insight into the dataset, the distribution of EC classes in the dataset was analyzed. The table indicates that the dataset is highly imbalanced, with Transferases being the most common class and Ligases the least common. This imbalance can affect the model's performance, as it may struggle to learn from underrepresented classes. To address this issue, the dataset was balanced using the RandomUnderSampler algorithm from the imbalanced-learn library.

EC Class	Count
Oxidoreductases	23544
Transferases	59022
Hydrolases	41283
Lyases	2376
Isomerases	4617
Ligases	1323
Translocases	1836

Table 2: Distribution of EC classes in the dataset

Taking a closer look at the distribution of EC classes on the second level of the hierarchy, the class 2.7 (Phosphotransferases) is the most common, while the class 2.6 (Acyltransferases) is the least common. This is because Phosphotransferases are involved in a wide range of cellular processes, making them more prevalent in the dataset. The class 2.6, on the other hand, is more specialized and less common in the dataset. After rebalancing the dataset, the distribution of EC classes is more uniform, but still not perfectly balanced. Further optimization may be required, but is beyond the scope of this study and would require additional data in the UniProt database.

Finally, the transformed features are integrated into a single dataset. The data is then normalized to ensure that all features are on a similar scale, which is important for the convergence of deep learning models. Normalization helps in speeding up the training process and achieving better performance. [21]

4.3 Feature Engineering

Feature engineering is a critical step in preparing data for prediction models. This process involves transforming raw data into meaningful features that can improve the performance of the model. In this section, the author describes the feature engineering techniques used in this study, focusing on the processing of protein sequences and the calculation of additional features to enhance the predictive power of the Deep Learning model.

To capture meaningful information from protein sequences, this study used several features derived from the sequences, including amino acid composition, molecular weight, isoelectric point, hydrophobicity, and sequence length. These features provide valuable insights into the physicochemical properties of the proteins, enabling the model to learn patterns that correlate with enzyme functions. The ProteinAnalysis class ferom the Biopython library was used to calculate these features. The following Python code snippet demonstrates the calculation of additional features from the protein sequences:

The first step is to clean the protein sequence shown in chapter 4.1. The sequence itself is used as a feature, and additional features are calculated using the ProteinAnalysis class from Biopython. To convert the cleaned sequences into a format suitable for the model, a tokenizer is used to encode the sequences into numerical data. In the context of protein sequences, each amino acid is mapped to a unique integer. For example, the sequence "ACDEFGHIKLMNPQRSTVWY" is tokenized into a list of integers.

Tokenization involves converting each amino acid into an integer based on its position in a predefined list of valid amino acids. This process can be mathematically represented as:

$$\begin{aligned} token(x) = i \quad \text{where} \quad x \in \{&A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, \\ &W, Y\} \quad \text{and} \quad i \quad \text{is the index of amino acid} \quad x \quad \text{in the list} \end{aligned}$$

The tokenized sequence is then passed through an embedding layer that transforms these integers into dense vectors. This embedding process is essential for capturing the contextual meaning of each amino acid within the sequence: $embedding(i) = v$ where v_i is the embedding vector for the token i. These embeddings are fed into the RNN, which processes the sequence and updates its hidden states accordingly, allowing the model to capture complex dependencies and interactions between amino acids. The sequences are then padded to ensure they all have the same length, which is necessary for batch processing in Deep Learning models.

Recent advancements have demonstrated that sequence-based models, including language models like ESM-1b, can achieve high accuracy in predicting protein functions and properties. For instance, the study by Hu et al. (2022) highlights the potential of protein-sequence based models like ESM-1b in predicting protein function from sequences. [22]

In addition to tokenizing the protein sequences, several biochemical features are calculated to provide a comprehensive representation of the proteins. These features include amino acid composition, molecular weight, isoelectric point, hydrophobicity, and sequence length. The Python code for calculating these features is as follows:

1. **Amino Acid Composition**: The amino acid composition represents the relative frequency of each of the 20 standard amino acids in a protein sequence.

    a) Calculation: It is calculated as the percentage of each amino acid in the sequence.

    b) Relevance: Different proteins have characteristic amino acid compositions that can provide clues about their function and stability. For example, membrane proteins often have higher hydrophobic amino acid content.

    c) Example: A protein with a high proportion of hydrophobic amino acids might be involved in membrane-related processes.

2. **Molecular Weight**: Molecular weight is the total mass of all amino acids in the protein sequence.

    a) Calculation: It is calculated by summing the average atomic masses of the amino acids in the sequence.

    b) Relevance: The molecular weight of a protein can influence its physical and chemical properties, such as solubility and interaction with other molecules.

    c) Example: Enzymes with larger molecular weights may have multiple domains or subunits.

3. **Isoelectric Point**: The isoelectric point is the pH at which the protein carries no net electrical charge.

    a) Calculation: It is determined by calculating the pH at which the positive and negative charges on the amino acids balance out.

    b) Relevance: The pI affects protein solubility and interaction with other molecules. Proteins are least soluble at their pI and more likely to precipitate.

    c) Example: Proteins with a low pI are often found in acidic environments, such as lysosomal enzymes.

4. **Hydrophobicity (GRAVY Score)**: The GRAVY (Grand Average of Hydropathicity) score is a measure of the overall hydrophobic or hydrophilic nature of a protein.

    a) Calculation: It is calculated by averaging the hydropathy values of all amino acids in the sequence.

    b) Relevance: Hydrophobicity influences protein folding, stability, and interaction with membranes.

    c) Example: Transmembrane proteins typically have a high GRAVY score due to their hydrophobic transmembrane regions.

5. **Sequence length**: The sequence length is the total number of amino acids in the protein sequence.
   a) Calculation: It is simply the count of amino acids in the sequence.
   b) Relevance: The length of a protein can indicate its complexity and the number of functional domains.
   c) Example: Longer proteins may have multiple functional domains or be involved in complex regulatory mechanisms.

These biochemical features provide a multi-dimensional representation of protein sequences, capturing both sequence-specific information and physicochemical properties. This feature-set is essential for analyzing the enzymes and predicting their functions accurately. A study by Gainza et al. (2020) demonstrates the importance of incorporating physicochemical features in protein function prediction models, showing that these features enhance the model's performance. For example, the choice of features such as molecular weight and isoelectric point is grounded in their proven relevance to protein function prediction. [23]

4.4 Model Development

In this section, this studys describes the model development process, including data splitting, model architecture, and the rationale behind the chosen methods. The goal is to predict EC classes based on protein sequences using a deep learning approach enhanced by additional biochemical features.

To use diffrent models for each hierachy level, the data is split into four levels of the EC hierachy. The first level represents the broadest classification, while the fourth level provides the most specific classification. This ensures that models are trained and evaluated on appropriately structured data, allowing for predictions at varying levels of specificity.

The model architecture combines sequence-based features with additional biochemical features to enhance prediction accuracy. The architecture consists of an embedding layer, two LSTM layers, and dense layers that integrate additional features.

The embdding layer converts amino acid sequences into dense vector representations, capturing semantic similarities between amino acids. This layer allows the model to handle varying sequence lengths and to learn useful representations of amino acids in the context of their sequence. After that Long Short-Term Memory (LSTM) layers are used to capture long-range dependencies in the sequence data, which is crucial for understanding the functional context of amino acids within the sequence. LSTMs are particularly effective in modeling sequential data due to their ability to remember information for long periods and manage the vanishing gradient problem. LSTMs are particularly well-suited for tasks involving sequence data due to their ability to manage long-term dependencies and their robustness against the vanishing gradient problem. Studies have demonstrated the effectiveness of LSTMs in various sequence analysis tasks, including protein function prediction and other bioinformatics applications.[24] [25]

Biochemical properties such as molecular weight, isoelectric point, hydrophobicity, and sequence length are included to provide additional context that can enhance the prediction accuracy. These features help the model understand the physical and chemical characteristics of the proteins, which are critical for predicting enzyme functions. The concatenation layer combines the output of the LSTM layers with the additional biochemical features, allowing the model to leverage both sequence-based and property-based information. This integration ensures that the model considers both the sequence context and the biochemical properties of the proteins. Finally the dense layers are used to integrate the combined features and produce the final classification output. These layers apply non-linear transformations to the combined features, enabling the model to learn complex patterns and relationships. This model uses an Adam optimizer with a learning rate of 0.001 and sparse categorical cross-entropy loss function, which is suitable for multi-class classification tasks. The model is compiled with the specified optimizer, loss function, and evaluation metrics to prepare it for training.

5 Results

5.1 Model Performance

This chapter presents the performance metrics of the developed model before and after the hyperparameter tuning. The model was evaluated at different EC levels to assess its accuracy, recall, and F1 score. These metrics provide insight into the model's initial performance and highlight areas for potential improvement through hyperparameter tuning.

EC Level	Accuracy	Recall	F1
1	0.94	0,94	0,93
2	0,90	0,90	0,90
3	0,94	0,94	0,93
4	0,75	0,75	0,72

Table 3: Model Performance before Hyperparametertuning

Initial results suggest that the model performs well at higher levels of the EC hierarchy (levels 1 to 3), but there is a notable decrease in performance at the most specific level (level 4). This suggests that there is room for improvement, particularly in fine-tuning the model for more specific classifications. Predicting the 4th EC Level is significantly more challenging than predicting higher levels due to the need to select from approximately 7000 classes 1, increasing the difficulty of accurate prediction. Moreover, the model has to predict the right class from a very unbalanced dataset, which makes it even more challenging.

To adress this issue, the model was fine-tuned using a grid search approach to optimize the hyperparameters. In addition to that the initial dataset was edited with the help of the SMOTE algorithm to balance the dataset. The results of the hyperparameter tuning are shown in the following table. [26] The following table shows the performance of the model after the hyperparameter tuning:

— NEEDS TO BE DONE —

5.2 Comparative Analysis with Existing Models

— NEEDS TO BE DONE —

5.3 Interpretation of Model Predictions

— NEEDS TO BE DONE —

6 Discussion

6.1 Implications of Findings

The findings of this study have substantial implications for both the field of computational biology and the practical application of deep learning models in environmental science. By developing a deep learning model that accurately predicts enzyme classes responsible for pesticide degradation, this research contributes to several critical areas.

Traditional methods of determining pesticide degradation and enzyme classification are often labor-intensive, time-consuming, and expensive. The computational approach presented in this study offers a more efficient alternative. By leveraging deep learning models, the research significantly reduces the time and cost associated with experimental methods. This efficiency can accelerate the development and testing of new agricultural products, ensuring that safer and more effective solutions reach the market faster.

The model developed in this study demonstrates significant improvements in predictive accuracy, particularly for the 4th level of the Enzyme Commission EC classification hierarchy. This enhanced accuracy is crucial for advancing the understanding of enzyme functions and their specific roles in biodegradation processes. Accurately predicting enzyme classes allows for more precise identification of enzymatic pathways involved in pesticide degradation, which is fundamental for developing effective bioremediation strategies. The Deep Learning model can be employed to help predicting unknown enzyme classifications in order to clean up contaminated environments more efficiently and reduce the ecological footprint of agricultural practices. The findings support the creation of more sustainable and environmentally friendly agricultural products, aligning with global efforts to mitigate pollution and protect natural ecosystems. At Bayer Crop Science, the model can be integrated into the product development process to enhance the safety and sustainability of new agricultural products, reducing the environmental impact of pesticide use. Especially in the modern context of increasing environmental awareness and regulatory scrutiny, the model provides a valuable tool for the development of new enzymes and biodegradation pathways.

The findings open several avenues for future research. One potential direction is the refinement of the model to improve performance at the fourth EC level, which remains challenging due to the high specificity and diversity of enzyme functions. Additionally, integrating this model with real-world environmental data can validate its practical applicability and uncover further insights into enzymatic degradation pathways. Collaborative efforts with experimental biologists can enhance the model's accuracy and expand its scope to include a wider range of pollutants and environmental conditions.

Moreover, this novel approach can be further improved to achieve even better results. Current methods often utilize the entire protein sequence and do not focus specifically on the ligand-binding pocket. By concentrating more on these specific pockets, it is possible to enhance the precision of enzyme classification and the prediction of degradation pathways. Future advancements should therefore aim to refine this focus on ligand-binding sites, leveraging detailed structural information to improve predictive accuracy.

6.2 Strenths and Limitations

Strengths:

1. Innovative Approach: The primary strength of this thesis lies in its innovative approach to predicting pesticide degradation by focusing on enzyme classification through deep learning. By integrating advanced computational techniques, this research addresses a gap in the current methodologies used for enzyme function prediction.
2. Comprehensive Methodology: The detailed and methodical approach taken in data collection, preprocessing, feature engineering, and model development ensures the robustness of the study. Each step is meticulously documented,

demonstrating a thorough understanding of the processes involved in developing a predictive model.

3. Utilization of Advanced Tools: The use of state-of-the-art tools such as P2Rank for ligand-binding site prediction and RNNs for sequence analysis highlights the technical sophistication of the study. These tools provide a solid foundation for accurate predictions and demonstrate the potential for further applications in bioinformatics.

4. Significant Performance Improvement: The developed model shows significant improvement in predictive accuracy, particularly at the higher levels of the Enzyme Commission hierarchy. This improvement underscores the effectiveness of combining sequence-based features with additional biochemical features.

5. Environmental and Economic Impact: By enabling more accurate predictions of pesticide degradation pathways, the study contributes to environmental sustainability and cost efficiency. The ability to develop targeted bioremediation techniques and accelerate the development of environmentally friendly agricultural products has far-reaching benefits.

Limitations:

1. Performance at Specific EC Levels: While the model performs well at higher EC levels, there is a notable decrease in performance at the most specific level (level 4). This limitation suggests that the model struggles with the high specificity and diversity of enzyme functions at this level, necessitating further refinement and optimization. One of the reasons is that the glsglsglsUniProt database is still unbalanced and needs to be further improved. In addition to that, it is possible that some enzymes are wrongly classified in the database, which can lead to wrong predictions.

2. Imbalanced Dataset: The initial dataset used in the study is highly imbalanced, with certain enzyme classes being significantly underrepresented. Although techniques such as SMOTE were employed to address this issue, the

imbalance may still affect the model's ability to generalize across all enzyme classes.

3. Focus on Ligand-Binding Pockets: Although the study emphasizes the importance of ligand-binding pockets, current methods still utilize the entire protein sequence, which may dilute the specificity of predictions. Future research should aim to enhance the focus on these pockets to improve predictive accuracy.

4. Generalizability to Real-World Data: The model's performance is primarily evaluated using data from UniProt and PDB, which are well-curated databases. The generalizability of the model to real-world environmental data remains to be validated, as real-world scenarios often involve more complex and noisy data. Although the model needs to be validated in vivo or in vitro, the results are promising and provide a strong foundation for further research.

7 Conclusion

7.1 Summary of Findings

— NEEDS TO BE DONE —

7.2 Contributions to the Field

— NEEDS TO BE DONE —

7.3 Final Remarks and Future Work

— NEEDS TO BE DONE —

Appendix

Appendix

List of References

References

[1] Brajesh K. Singh and Allan Walker, “Microbial degradation of organophosphorus compounds,” *FEMS Microbiology Reviews*, vol. 30, no. 3, pp. 428–471, May 2006. (visited on Jun. 6, 2024).

[2] Xing Kai Chia, Tony Hadibarata, Risky Ayu Kristanti, Muhammad Noor Hazwan Jusoh, Inn Shi Tan, and Henry Chee Yew Foo, “The function of microbial enzymes in breaking down soil contaminated with pesticides: A review,” *Bioprocess and Biosystems Engineering*, vol. 47, no. 5, pp. 597–620, May 2024. (visited on Jun. 6, 2024).

[3] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus, “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, e2016239118, Apr. 2021. (visited on Jun. 24, 2024).

[4] Yu Li, Sheng Wang, Ramzan Umarov, Bingqing Xie, M. Fan, Lihua Li, and Xin Gao, “DEEPre: Sequence-based enzyme EC number prediction by deep learning,” *Bioinformatics*, vol. 34, pp. 760–769, 2017. (visited on May 29, 2024).

[5] *DeEPn: A deep neural network based tool for enzyme functional annotation - Consensus*, https://consensus.app/papers/deepn-network-based-tool-annotation-

semwal/c62a6c023f2d5b2f9348b36f6329daae/?utm_source=chatgpt. (visited on May 29, 2024).

[6] Naoki Watanabe, Masaki Yamamoto, Masahiro Murata, Yuki Kuriya, and Michihiro Araki, "EnzymeNet: Residual neural networks model for Enzyme Commission number prediction," *Bioinformatics Advances*, vol. 3, no. 1, vbad173, Jan. 2023. (visited on Jun. 25, 2024).

[7] Radoslav Krivák and David Hoksza, "P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure," *Journal of Cheminformatics*, vol. 10, no. 1, p. 39, Aug. 2018. (visited on May 29, 2024).

[8] Peter K. Robinson, "Enzymes: Principles and biotechnological applications," *Essays in Biochemistry*, vol. 59, pp. 1–41, Nov. 2015. (visited on Jun. 6, 2024).

[9] "Enzyme Nomenclature. Recommendations 1992," *European Journal of Biochemistry*, vol. 223, no. 1, pp. 1–5, 1994. (visited on Jun. 26, 2024).

[10] Liubov Poshyvailo-Strube, "Modelling and simulations of enzyme-catalyzed reactions," Ph.D. dissertation, Jun. 2015.

[11] "Recurrent neural network," *Wikipedia*, Jun. 2024. (visited on Jun. 27, 2024).

[12] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. (visited on Jun. 20, 2024).

[13] D. Krstajic, L. Buturovic, D. Leahy, and Simon Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of Cheminformatics*, vol. 6, 2014. (visited on Jun. 25, 2024).

[14] James Bergstra and Yoshua Bengio, “Random Search for Hyper-Parameter Optimization,”

[15] “Precision and recall,” *Wikipedia*, Apr. 2024. (visited on May 31, 2024).

[16] “F-score,” *Wikipedia*, May 2024. (visited on May 31, 2024).

[17] “Accuracy and precision,” *Wikipedia*, Apr. 2024. (visited on May 31, 2024).

[18] UniProt Consortium, “UniProt: The universal protein knowledgebase in 2021,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, Jan. 2021.

[19] “A one-letter notation for amino acid sequences (definitive rules),” *Pure and Applied Chemistry*, vol. 31, no. 4, pp. 639–646, Jan. 1972. (visited on Jul. 4, 2024).

[20] Yizhou Dang, Yuting Liu, Enneng Yang, Guibing Guo, Linying Jiang, Xingwei Wang, and Jianzhe Zhao, *Repeated Padding as Data Augmentation for Sequential Recommendation*, Mar. 2024. arXiv: `2403.06372 [cs]`. (visited on Jul. 4, 2024).

[21] Sergey Ioffe and Christian Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, Mar. 2015. arXiv: `1502.03167 [cs]`. (visited on Jul. 4, 2024).

[22] Mingyang Hu, Fajie Yuan, Kevin K. Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding, “Exploring evolution-based & -free protein language models as protein function predictors,” *ArXiv*, vol. abs/2206.06583, 2022. (visited on Jun. 22, 2024).

[23] P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia, “Deciphering interaction fingerprints from protein molec-

ular surfaces using geometric deep learning," *Nature Methods*, vol. 17, no. 2, pp. 184–192, Feb. 2020. (visited on Jun. 22, 2024).

[24] Jiale Liu and Xinqi Gong, "Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction," *BMC Bioinformatics*, vol. 20, 2019. (visited on Jun. 23, 2024).

[25] Yunong Zhang, "Encoder-decoder models in sequence-to-sequence learning: A survey of RNN and LSTM approaches," *Applied and Computational Engineering*, 2023. (visited on Jun. 23, 2024).

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002. arXiv: `1106.1813 [cs]`. (visited on Jun. 27, 2024).

Statement of Independent Work

I hereby declare that I have prepared this work independently. Only the sources and aids expressly named in the work were used. I have marked as such any verbatim or analogously adopted ideas. This thesis has not yet been submitted in the same or a similar form to any examination authority.

Düsseldorf, July 5, 2024

Tobias Polley