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 8, 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 This thesis may also not be published and may only be made accessible to the examiners, administrative staff and members of the examination committee and, on request, to an evaluation committee.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

1 Introduction

1.1 Motivation

The prediction of pesticide degradation has gained significant attention due to its environmental and health impacts. Traditional methods for determining the degradation of pesticides by enzymes are laborintensive and time-consuming, necessitating the development of efficient computational methods. Pesticides, used globally for crop protection, often persist in the environment, posing risks to ecosystems and human health. Enzymes, as biological catalysts, play a crucial role in breaking down these pesticides into less harmful substances. This enzymatic degradation is essential for reducing the environmental toxicity of pesticides.

Recent advancements in DNA and RNA sequencing technologies have expanded the understanding of enzyme functions. Leveraging this wealth of data through computational methodologies, particularly Deep Learning, offers a promising approach to predicting enzymemediated pesticide degradation. Successful applications of Deep Learning in genomics and environmental science suggest its potential in enhancing pesticide degradation predictions, which is vital for developing sustainable agricultural products.

Especially in the context of discovering previously unknown enzymes and their functions, Deep Learning models can provide valuable insights into enzyme classification and behavior. By accurately predicting enzyme classes responsible for pesticide degradation, these models can facilitate the development of environmentally friendly and sustainable agricultural practices. The ability to predict enzyme functions with high accuracy and resolution is crucial for understanding the complex interactions between enzymes and pesticides, enabling the design of effective bioremediation strategies.

1.2 Problem Statement

Despite significant advancements in bioinformatics and computational biology, predicting enzyme classes involved in pesticide degradation remains challenging. Most of the traditional methods depend mainly on experimental data, which is time-consuming and high costly in most cases. Moreover, the functions of enzymes are very diverse and complex, as well as their interaction with different kinds of substrates, which further complicates the prediction process. Additionally, in some cases, the existing computational models have shown low prediction accuracies. This thesis puts forward an innovative Deep Learning model that employs predictions of enzyme binding sites to enhance the accuracy and efficiency of enzyme class prediction. The approach is focused on critical interaction regions, with the goal of outperforming state-of-the-art models and providing a strong tool for computational biology and environmental science. In particular, this research aims to:

(a) Develop a *novel* Deep Learning architecture that targets the ligand-binding sites of enzymes to predict the enyzmatic function.

(b) Enhance the accuracy and resolution of enzyme class predictions, particularly at the most specific level of the Enzyme Commission (EC) hierarchy.

(c) Address the challenges associated with data imbalance and class representation in enzyme classification datasets.

1.3 Structure of the Thesis

This thesis is divided into five chapters, each addressing different aspects of the research to overview it comprehensively. The first chapter of this thesis is an introduction to the entire work. It starts with outlining the motivation for study and states the environmental concerns related to pesticide use and the need for effective degradation prediction methods. This chapter includes the problem statement section, which identifies the problems associated with predicting enzyme-mediated pesticide degradation. The introduction defines the general objective of

this study: developing a deep-learning model to predict pesticide degradation based on enzyme classes. Lastly, this chapter concludes with an overview of the thesis structure.

The literature review chapter covers existing research studies and foundational theories that are important to the study. It deals with enzymatic mechanisms under the degradation of pesticides, giving information on how the enzymes mediate the degradation process. This is succeeded by using Deep Learning techniques for application in environmental science and the improvements that it may offer to predictive accuracy. Finally, it suggests advanced methods that can be used in enzyme classification due to the limitations of the current models in many ways.

The methodology chapter offers a complete account of the research design and procedures followed in the study. It starts with data collection, stating the sources and pre-processing processes executed to prepare the dataset for analysis. The feature engineering section elaborates on how relevant features were engineered out of the data to enable accurate predictions. It is then followed by the model development process, the architecture of a Deep Learning model, and the final training process.

Chapter results talk about the research outcomes, starting with the evaluation of model performance by giving details of different model performance metrics and how effectively they predict pesticide degradation. It will also compare the developed model with those already in existence to demonstrate how improvements and benefits have been achieved. It interprets model predictions, including their practical implications, and applies them in real-world situations.

The discussion chapter will synthesize critical findings from the research and present a reflection on the significance and impact of these results. It discusses the strengths and limitations of the study by pointing out where models performed well and stating where there is room for improvement. The chapter ends with brief contributions to the field, pointing at the novelty and practical applications that the research relates to. In addition, it also gives future work recommendations in which the study points out research directions for building over its findings.

2 Theoretical Background

2.1 Principles of Enzymology

Enzymology is the scientific study of enzymes, which are biological catalysts that accelerate biochemical reactions in living organisms. Enzymes are crucial for various cellular processes, including metabolism, DNA replication, and signal transduction. They are highly specific, meaning each enzyme typically catalyzes only one type of reaction or interacts with a specific substrate. This specificity arises from the enzyme's unique three-dimensional structure, which includes an active site where the substrate binds and the reaction occurs. [1]

Enzymes are classified based on the types of reactions they catalyze, according to a system established by the Enzyme Commission (EC). This classification groups enzymes into six main classes: Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, and Ligases. For example, Oxidoreductases catalyze oxidation-reduction reactions, where the transfer of electrons occurs between molecules. Understanding these classifications helps in predicting the enzyme's role in biochemical processes. Digestive enzymes like amylase break down starch into sugars, facilitating digestion. Similarly, enzymes in laundry detergents help break down protein stains, making them easier to remove. The specificity of enzymes for their substrates is a key feature that enables them to perform their roles in various biochemical pathways with high precision. The following figure gives a brief overview of the seven classes of enzymes according to the EC classification system:

1. **Oxidoreductases:** These enzymes catalyze oxidation-reduction reactions, where the transfer of electrons occurs between molecules.
2. **Transferases:** These enzymes transfer functional groups from one molecule to another.
3. **Hydrolases:** These enzymes catalyze the hydrolysis of various bonds, including ester, glycosidic, peptide, and others.
4. **Lyases:** These enzymes add or remove groups to form double bonds, without hydrolysis or oxidation.

5. **Isomerases:** These enzymes catalyze the rearrangement of atoms within a molecule, leading to isomerization.
6. **Ligases:** These enzymes catalyze the joining of two molecules with the simultaneous hydrolysis of a diphosphate bond in ATP or a similar triphosphate.
7.

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. [2]

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:

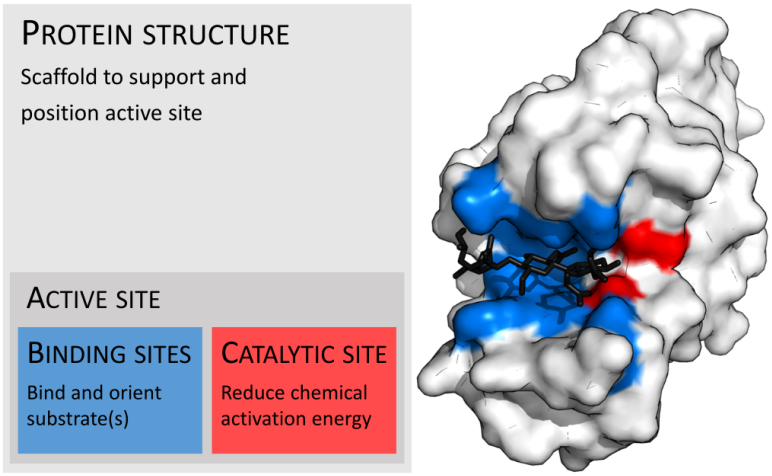
EC Level	**Number of classes**
1	7
2	79
3	320
4	7876

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

The table shows an unbalanced distribution of EC numbers across the four levels of the classification system, with a higher number of classes at the lower levels. This distribution reflects the increasing specificity and diversity of enzyme functions as it moves from higher to lower levels of the classification system. The high number of classes at the fourth level highlights the complexity of enzyme functions and the challenges associated with predicting enzyme classes accurately. This complexity will be further discussed in the subsequent sections.

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.

Figure 1: Organisation of enzyme structure and lysozyme example.

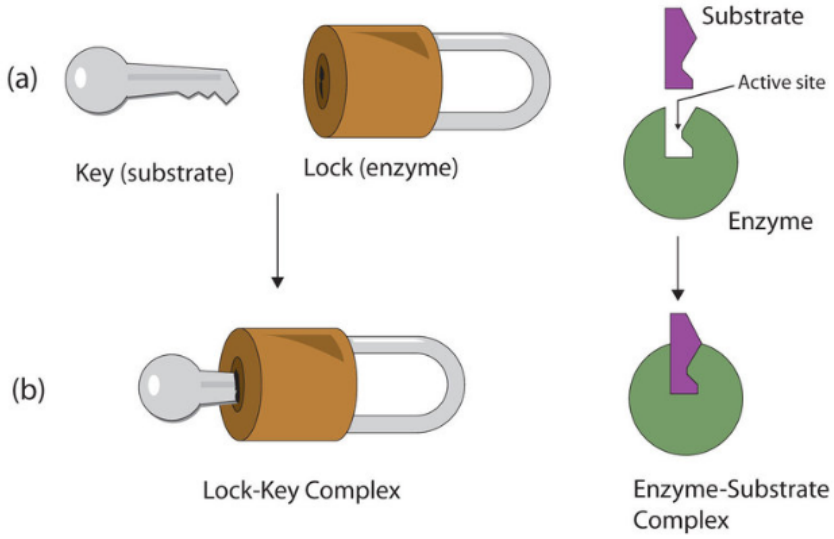


Source: [3]

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

Figure 2: Lock-and-key model that explains the selectivity of enzymes



Source: [4]

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

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. Pesticides, used globally to protect crops, often persist in the environment, leading to potential ecological and health risks. Enzymes, as biological catalysts, play a pivotal role in degrading these toxic compounds into less harmful substances. The enzymatic degradation process involves various mechanisms, predominantly microbial enzymatic activities, which catalyze reactions that transform pesticides.

For instance, reductive enzymes catalyze the reduction of pesticides, often by donating electrons and hydrogen atoms to the molecules, breaking down complex structures into simpler forms. Reductive dehalogenases, found in microorganisms like Dehalococcoides, are particularly effective in breaking down halogenated organic compounds. The efficiency of microbial enzymes in degrading soil-contaminated pesticides has been well-documented. Singh and Walker (2012) [5] highlighted the effectiveness of microbial degradation of organophosphorus compounds, while Chia et al. (2013) [6] discussed advancements in microbial enzymes for 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.

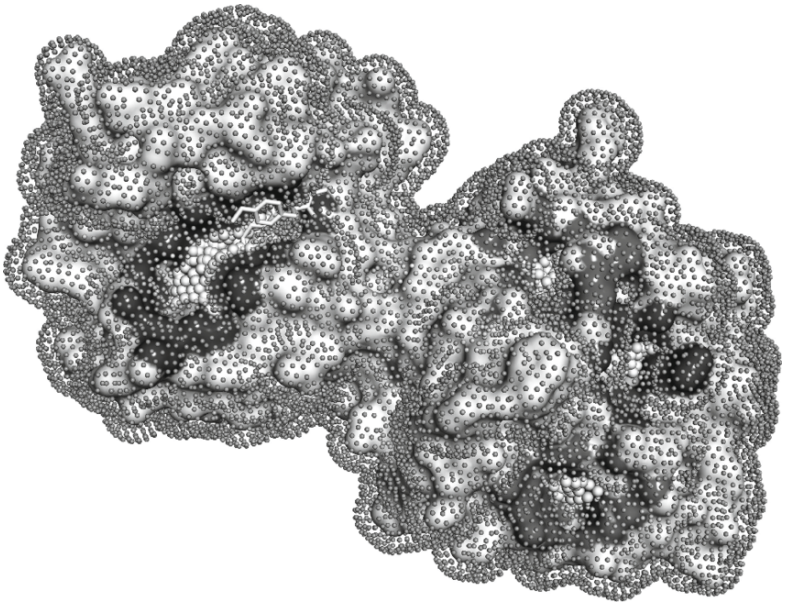
2.2 Fundamentals of Ligand Binding Site Prediction

Enzymes interact with substrates at specific binding sites, where the catalytic reactions occur. Predicting these ligand-binding sites is crucial for understanding 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. P2Rank uses a combination of geometric and physicochemical descriptors to analyze protein structures and predict the locations of potential binding sites. For instance, Connolly Points are regularly spaced points generated on the protein's surface to represent solvent-accessible areas, which P2Rank analyzes to predict binding sites. These points are associated with Connolly Feature Vectors (CFVs) that describe the physicochemical properties of the protein surface.

Figure 3: Protein (1FBL) is covered in a layer of points lying on a Connolly surface.



Source: [7]

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. The key steps in the P2Rank prediction process are as follows:

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

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.

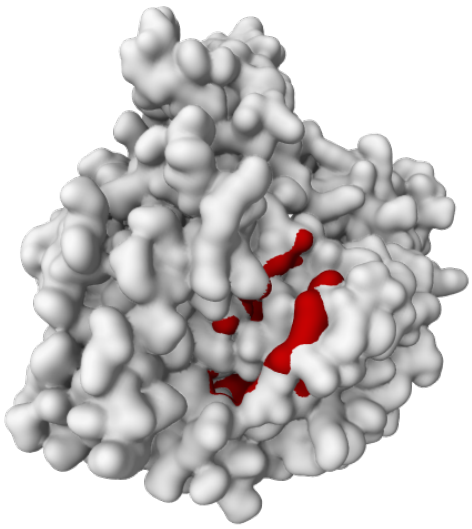
4. **Clustering**: Connolly points with high ligandability scores are clustered using a single-linkage clustering method, representing potential binding pockets on the protein surface.

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]

The following image illustrates an example of the P2Rank prediction for a protein structure (2SRC) with predicted ligand-binding sites:

Figure 4: P2Rank example for 2SRC (TYROSINE-PROTEIN KINASE)



Source: [7]

The red section in the image represents the predicted ligand-binding site on the protein structure, highlighting the potential interaction regions for ligands. These predicted sites can provide valuable insights into the enzyme's function and catalytic activity, aiding in the prediction of enzyme classes. By focusing on the ligand-binding pockets, P2Rank can identify the specific residues involved in the catalytic

processes, enhancing the accuracy of enzyme class predictions. This specificity 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 improve the accuracy of predictions but also reduce the computational complexity by focusing on smaller, more relevant regions of the protein.

2.3 Deep Learning in Enzymology

Deep Learning is a subset of machine learning and artificial intelligence (AI) that focuses on using neural networks with many layers (hence "deep") to model and understand complex patterns and representations in data. Unlike traditional machine learning models, Deep Learning algorithms can automatically learn and extract features from raw input data, making them particularly powerful for tasks involving large and complex datasets, such as image recognition, natural language processing, and speech recognition. The key strength of Deep Learning lies in its ability to improve performance with more data and compute power, enabling the development of models that can achieve state-of-the-art results in various domains. [8]

Deep Learning has become an essential tool in environmental science, enabling advanced prediction and understanding of complex biochemical processes. Various Deep Learning architectures, such as the protein-transformer ESM model, have significantly impacted the prediction of biological properties from sequence data. These models can analyze vast quantities of biochemical data to predict enzyme interactions and functions, providing valuable insights into pesticide degradation mechanisms. [9]

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. [10]

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, many methods only focus on the prediction to the 3rd level of the EC classification, which may not provide sufficient detail for predicting pesticide degradations. For example the accuracy of EnzymeNet, a residual neural network model, across all the 4th level is 0.398. 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. [11]

Current state-of-the-art models, such as DEEPre or EnzymeNet, have shown promising results in predicting enzyme classes based on sequence data. However, these models are limited in their ability to predict enzyme classes with high specificity and resolution, particularly at the 4th level of the EC hierarchy. The complexity and diversity of enzyme functions at this level pose challenges for accurate prediction, necessitating more advanced Deep Learning models. By leveraging the latest Deep Learning techniques and incorporating detailed biochemical features, it is possible to develop more accurate and efficient models for predicting enzyme classes involved in pesticide degradation.

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 the 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]

2.4 Enzyme Classification with Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of Deep Learning 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 involving sequential data, where the order of the data points matters. 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. [12]

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.

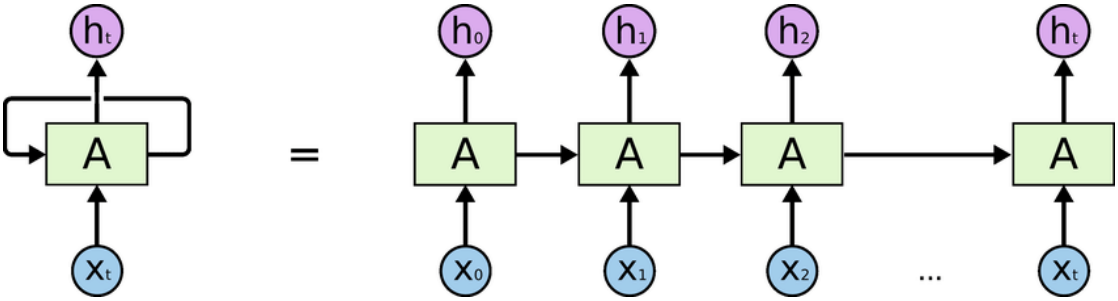
One of the key advancements in RNNs is the development of Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), which are designed to

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. [13]

The following image illustrates the basic structure of an RNN:

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



Source: [14]

1. **Input Sequence (x):** In the diagram, the blue circles $(x_0, x_1, x_2, \ldots, x_t)$ represent the input data at different time steps. Each x_i is an input vector that the network receives at a specific time step t. These inputs can be any sequential data, such as words in a sentence or data points in a time series.

2. **Hidden State (h):** The purple circles $(h_0, h_1, h_2, \ldots, h_t)$ represent the hidden state of the network. At each time step, the hidden state h_i is updated based on the current input x_i and the previous hidden state h_{i-1}. The hidden

state acts as the network's memory, retaining information about previous inputs, which is crucial for understanding sequences and making predictions based on past data.

3. **Processing Unit (A):** The green rectangles (A) denote the recurrent unit, which processes the inputs and updates the hidden state. Each unit takes the input x_i and the previous hidden state h_{i-1} to produce the current hidden state h_i. This recurrence allows the network to maintain a chain of dependencies across time steps, enabling it to capture temporal patterns in the data.

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.

2.5 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.

To get the optimal hyperparameters, a Hyperparameter tuning is necessary. This is a critical aspect of developing Deep Learning models, as it involves selecting the optimal set of hyperparameters that maximizes the model's performance. This process can significantly impact the effectiveness of the model, as hyperparameters control various aspects of the learning process, such as the learning rate, batch size, number of epochs, and network architecture. Hyperparameters differ from model parameters as they are set before the training process begins and remain constant during training. Properly tuning these hyperparameters is essential for: [15]

- Optimizing the model's accuracy.
- Preventing overfitting or underfitting.
- Ensuring efficient use of computational resources.

The KerasClassifier wrapper from the keras library is used to integrate a Keras Deep Learning model with Scikit-learn's Grid Search functionality. This integration allows for systematic and efficient exploration of hyperparameter values. Grid Search involves specifying a grid of hyperparameter values and systematically evaluating the model performance for each combination. The GridSearchCV function from Scikit-learn is used to perform this exhaustive search over the specified hyperparameter space. The key hyperparameters tuned in this study include:

- Optimizer: Algorithms for updating model weights (ADAM in this case).
- Initialization: Methods for initializing model weights (e.g., glorot uniform, normal).
- Epochs: Number of complete passes through the training dataset.
- Batch Size: Number of samples processed before the model is updated.

The Random-Grid-Search is then executed to find the best combination of hyperparameters that results in the highest model accuracy. A study by Bergstra and Bengio (2012) showed that Random Search is more efficient than Grid Search for hyperparameter optimization, as it explores the hyperparameter space more effectively. Random Search randomly samples hyperparameter values from predefined ranges, allowing for a more comprehensive search and better performance. [16]

The accuracy of the model is continuously evaluated using various metrics, including precision, recall, F1 score, and accuracy. The best performance metrics are used to select the optimal hyperparameters for the model. The model is then trained using the selected hyperparameters and evaluated on the test set to assess its generalization performance. This process ensures that the model is robust and can accurately predict enzyme classes.

To effectively evaluate and compare models, various metrics are used: [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.

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

3 Methodology

3.1 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:

- **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.
- **Reviewed Entries:** UniProt contains both reviewed (Swiss-Prot) and unreviewed (TrEMBL) entries. Reviewed entries are manually curated by experts, ensuring high accuracy and reliability.
- **Functional Annotations:** Each protein entry includes detailed functional annotations, such as catalytic activity, biological processes, and involvement in pathways.
- **3D Structural Data:** UniProt links to structural databases like PDB, providing access to 3D structures of proteins, which are crucial for understanding enzyme mechanisms.

- **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. A figure of the data collection and processing workflow is shown in the appendix 11.

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.

3.2 Data Preprocessing

After the 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. These types of amino acids can not be processed by the model and need to be removed. These could for example be "B", "J", "O", "U", "X", or "Z". This step ensures that the sequences contain only valid amino acid residues, which is crucial for accurate modeling. [19]

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]

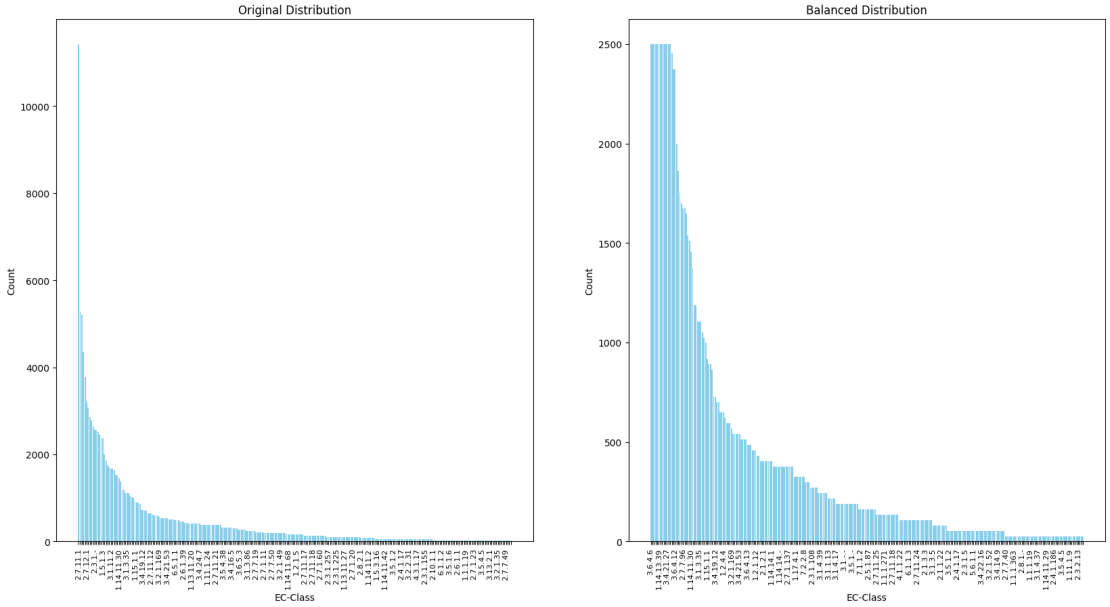
For gaining a first insight into the dataset, the distribution of EC classes in the dataset was analyzed. The table indicates that the dataset is highly imbalanced, with Transferases being the most common class and Ligases the least common. This imbalance can affect the model's performance, as it may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

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

To address this issue, the dataset was balanced using the RandomUnderSampler algorithm from the imbalanced-learn library. The RandomUnderSampler uses a sampling strategy with k=2500 to balance the dataset by selecting ec classes that have more than 2500 samples. This approach ensures that the model is trained on a more balanced dataset, which can improve its performance on underrepresented classes. The following comparison shows the distribution of EC classes before and after balancing the dataset:

Figure 6: Distribution of EC classes before and after balancing



Source: Own illustration

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. To keep the original distribution of the dataset, the distribution was kept as it is. 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]

3.3 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. 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 from the Biopython library was used to calculate these features.

Amino acid composition refers to the relative frequency of each of the 20 standard amino acids in a protein sequence. Proteins are made up of amino acids, which are the building blocks that determine the protein's structure and function. Each amino acid has unique properties that influence the protein's overall characteristics. [22]

The amino acid composition provides information about the protein's primary structure, which is essential for understanding its function. The amino acid composition is calculated as the percentage of each amino acid in the sequence. For example, a protein with a high proportion of hydrophobic amino acids may be in ec class 2.1 (Transferases), because these enzymes often have hydrophobic binding sites. [23]

Molecular weight is the total mass of all the amino acids in a protein sequence. It is measured in Daltons (Da) and reflects the size of the protein. The molecular weight of a protein influences its physical and chemical properties, such as solubility and interaction with other molecules. Larger proteins may have more complex structures with multiple functional domains, which can affect their enzymatic activity.

The isoelectric point (pI) is the pH at which a protein carries no net electrical charge. At this pH, the number of positive and negative charges on the protein is equal, resulting in minimal solubility. The pI of a protein affects its solubility and

interaction with other molecules. Proteins are least soluble at their pI and more likely to precipitate. This property is important for understanding protein behavior in different pH environments, which can influence their function and stability. By including pI as a feature, the model can better predict how proteins will behave in various conditions, aiding in accurate enzyme classification.

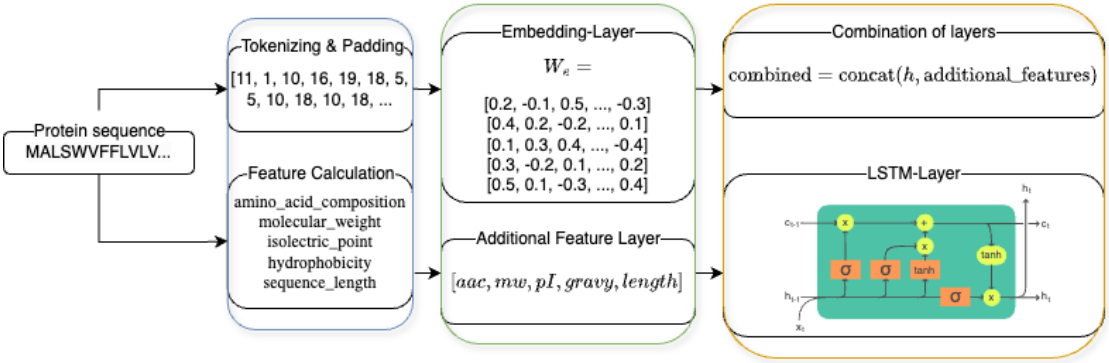
Hydrophobicity, measured by the Grand Average of Hydropathy (GRAVY) score, indicates the overall hydrophobic or hydrophilic nature of a protein. It is calculated by averaging the hydropathy values of all amino acids in the sequence. Hydrophobicity affects protein folding, stability, and interaction with membranes. Proteins with high hydrophobicity are likely to be involved in membrane-related processes, while those with low hydrophobicity are generally more soluble in water. Understanding the hydrophobic or hydrophilic nature of a protein is crucial for predicting its function, especially in relation to its interaction with other molecules and environments. [24]

These biochemical features provide a multi-dimensional representation of protein sequences, capturing both sequence-specific information and physicochemical properties. By integrating these features, the model gains a comprehensive understanding of the proteins, enabling more accurate and reliable predictions. The features were chosen in a comprehensive analysis and are also based on a study by Gainza et al. (2020), which highlights the importance of incorporating these specific physicochemical features in protein function prediction models. [25]

3.4 Model Architecture

The model architecture is a critical component of the prediction process, as it determines how the data is processed and transformed to make predictions. This study's model architecture is designed to leverage both raw protein sequences and biochemical features to accurately predict enzyme classes. The model architecture is shown in the figure below:

Figure 7: Model Architecture



Source: Own illustration

To integrate raw protein sequences into the model, the sequences are first tokenized and passed through an embedding layer. Each amino acid is mapped to a unique integer. For example, the sequence "ACDEFGHIKLMNPQRSTVWY" is tokenized into a list of integers. This process can be mathematically represented as:

$$token(x) = i \text{ where } x \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \tag{5}$$

and i is the index of amino acid x in the list.

The tokenized sequence is then passed through an embedding layer that transforms these integers into dense vectors, capturing the contextual meaning of each

amino acid within the sequence. This embedding process is essential for understanding the relationships between amino acids and their roles in protein function.

$$\text{embedding}(i) = v \tag{6}$$

where v_i is the embedding vector for the token i.

These embeddings are fed into the Recurrent Neural Network (RNN), which processes the sequence and updates its hidden states accordingly, allowing the model to capture complex dependencies and interactions between amino acids. Sequences are then padded to ensure they all have the same length, necessary for batch processing in deep learning models.

Integrating raw protein sequences as a separate layer in the model allows for the extraction of complex patterns and dependencies within the sequences, similar to how large language models (LLMs) like ProtBERT process natural language. ProtBERT, a pre-trained language model for protein sequences, demonstrates the effectiveness of using raw sequence data for protein function prediction by capturing intricate patterns and dependencies. By integrating raw sequences, the model benefits from advanced sequence modeling capabilities, improving its predictive performance. [26]

Using techniques similar to those employed in LLMs, the model captures contextual relationships between amino acids, enhancing its ability to predict enzyme functions. This approach is particularly useful for understanding long-range interactions and dependencies that are crucial for protein function.

Biochemical features such as molecular weight, isoelectric point, hydrophobicity, and sequence length provide additional context that can enhance prediction accuracy. These features help the model understand the physical and chemical characteristics of the proteins, which are critical for predicting enzyme functions.

1. **Embedding 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 learn useful representations of amino acids in the context of their sequence.

2. **LSTM Layers:** Long Short-Term Memory (LSTM) layers 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. Studies have demonstrated the effectiveness of LSTMs in various sequence analysis tasks, including protein function prediction. [27]

3. **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.

4. **Dense Layers:** Integrates 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. The dense layers are responsible for making the final predictions based on the input data, providing the model's output for each EC class.

To address the hierarchical nature of the Enzyme Commission (EC) classification system, the data is split into four levels of the EC hierarchy. 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. Each level of the hierarchy requires a different degree of detail and understanding, and splitting the data accordingly helps tailor the model's learning process to the complexity of each classification level.

The integration of both sequence-based features and biochemical properties provides a comprehensive and multi-dimensional representation of proteins, enhancing the model's ability to make accurate and reliable predictions of enzyme classes.

This holistic approach ensures that the model leverages the strengths of detailed biochemical properties and advanced sequence modeling, resulting in a robust and effective predictive tool.

4 Results

4.1 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. The metrics are calculated based on the model's predictions and the actual enzyme classes in the test dataset. For better comparability, the model's performance is evaluated at each EC level separately. Accuracy, Recall and F1 are used for a better comparison of the performance with other models.

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 Hyperparameter tuning and Resampling

As shown in the table above, the model performs well at the first three levels of the Enzyme Commission (EC) hierarchy, with high accuracy, recall, and F1 score. However, the model's performance decreases at the fourth EC level, where the classification becomes more specific and challenging. The drop in accuracy, recall, and F1 score at this level indicates that the model struggles to predict the most detailed enzyme classes accurately. 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. The model's performance at the fourth EC level suggests that further optimization is needed to improve its predictive power for more specific enzyme classes. To increase the model performance the dataset was edited with the help of the

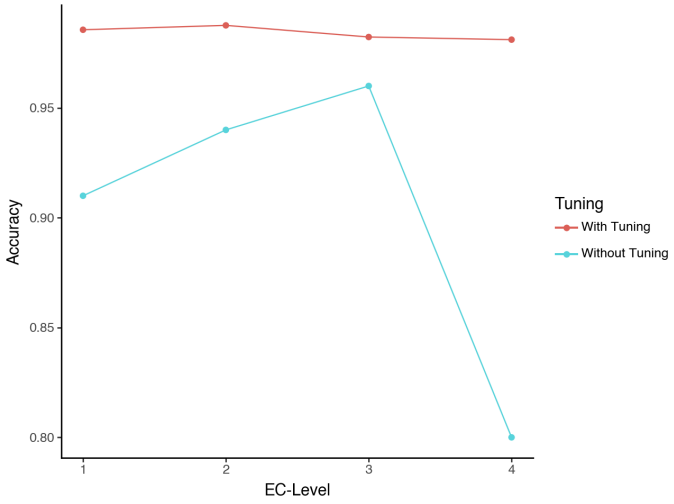
RandomUnderSampler as shown in chapter 3.2. The model was retrained on the resampled dataset with no further changes to the model architecture and parameters. The results of the model performance after resampling are shown in the following table:

EC Level	Accuracy	Recall	F1
1	0.91	0,91	0,91
2	0,94	0,92	0,94
3	0,96	0,96	0,96
4	0,80	0,80	0,78

Table 4: Model Performance after Resampling

The performance metrics still show a significant decrease in accuracy, recall, and F1 score at the fourth EC level. The 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. To optimize the predictions even further, the model was fine-tuned using hyperparameter tuning. The results of the model performance after hyperparameter tuning are shown in the following table:

Figure 8: Model Performance before and after Hyperparameter tuning



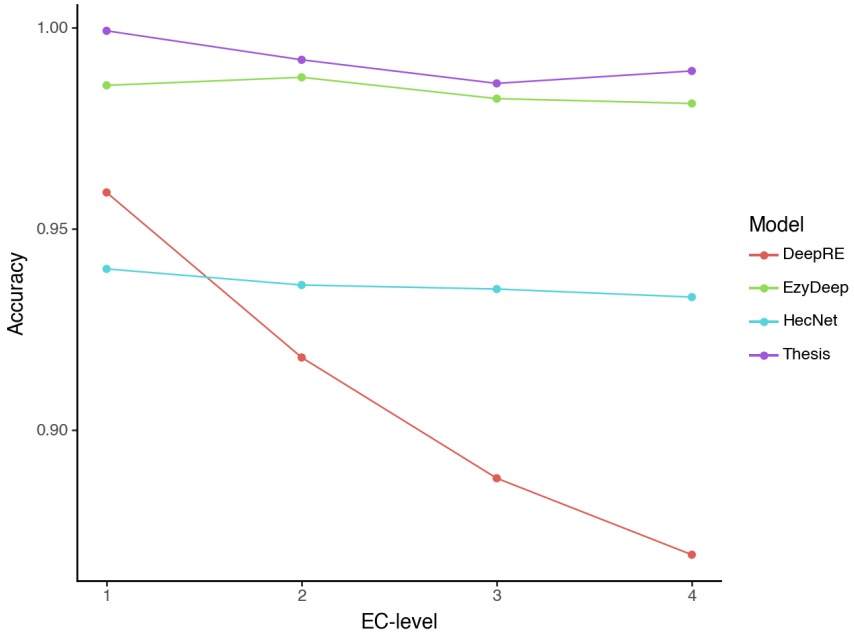
Source: Own illustration

The hyperparameter tuning has improved the model's performance at the fourth EC level, with a significant increase in accuracy, recall, and F1 score. This indicates that the model is better able to predict the most specific enzyme classes after the fine-tuning process. On the other hand, the model's performance at the first three levels remains consistent, suggesting that the hyperparameter tuning has not significantly impacted the model's performance at these levels. Since the increase in performance is only around 10 percent, further optimization may be required to enhance the model's predictive power at the fourth EC level. This could involve refining the model architecture, increasing the size of the training dataset, or incorporating additional features to improve the model's ability to predict specific enzyme classes. A possible approach could be to integrate anchor points based on protein sequences to get a more accurate prediction of the fourth EC level. Dalkiran et al. (2018) have shown that anchor points can significantly improve the prediction accuracy of enzyme classes, particularly at the most specific level. [28]

4.2 Comparative Analysis with Existing Models

To evaluate the performance of the proposed model, a comparative analysis was conducted against existing state-of-the-art models: DeepRE, EzyDeep, and HecNet. The comparison focuses on the prediction accuracy across the four levels of the Enzyme Commission (EC) classification system. All models were evaluated on the same dataset provided by COFACTOR, ensuring a fair and consistent comparison. [29]

Figure 9: Comparison of Accuracy with other models



Source: Own illustration

As shown in the figure above, the proposed model outperforms existing models at the all levels of the EC hierarchy, with a higher accuracy. The model's performance at the first three levels is significantly better than existing models, demonstrating its superior predictive power. This suggests that the model's focus on ligand-binding sites has contributed to its high accuracy and precision in predicting enzyme classes. The targeted approach to enzyme classification has proven to be more effective than traditional methods that utilize the entire protein sequence. The model's emphasis on ligand-binding sites enables it to capture the functional aspects of enzymes

more accurately. This targeted approach reduces computational complexity and enhances prediction accuracy, making it a valuable tool for enzyme classification and prediction.

The model's performance at the fourth EC level is comparable with EzyDeep, indicating that it can predict specific enzyme classes with similar accuracy. Although the performance of all models declines at this level due to the increased complexity, the Thesis model still achieves the highest accuracy. This indicates its ability to manage detailed and specific enzyme classifications effectively. Maintaining high accuracy at this level is critical for understanding the enzyme's catalytic activity.

Despite these advancements, there is still room for improvement. Further refinement in balancing techniques could improve model performance on underrepresented classes. A more extensive Hyperparameter tuning could also enhance the model's predictive power, particularly at the most specific classification levels. In addition to that, incorporating more diverse biochemical and environmental data could enhance model accuracy, especially at the most specific classification levels. The methods to enhance the model's performance are discussed in the section 6.2 in detail.

4.3 Interpretation of Model Predictions

The Deep Learning model developed in this study predicts enzyme classes with a high degree of accuracy, particularly at the first three levels of the Enzyme Commission (EC) classification hierarchy. The focus on ligand-binding sites has significantly contributed to the precision of the prediction at the fourth level, which is the most challenging due to the high number of classes and the complexity of enzyme functions. Predicting the fourth level is crucial for understanding the specific functions of enzymes that are involved in certain metabolic pathways, such as pesticide degradation. The model's performance at this level provides valuable insights into the enzyme's catalytic activity and substrate specificity, which are essential for designing effective products for agricultural applications.

However, the model's performance decreases at the fourth EC level, where the classification becomes highly specific. This drop in accuracy highlights the complexity and diversity of enzyme functions at this detailed level. This shows that the model needs further optimization to accurately predict the most specific enzyme classes. Integrating additional biochemical and environmental data, as well as refining the model architecture, could address the challenges associated with this level of specificity. Collaborative efforts with experimental biologists will be essential for validating and expanding the model's applicability. In addition to that, the model's performance can be further improved by increasing the size of the training dataset and incorporating more diverse enzyme sequences. This will help the model learn more about the subtle differences between enzyme classes and improve its predictive power.

The model's emphasis on ligand-binding sites offers a *novel* and effective approach to enzyme classification. By concentrating on these specific regions, the model accurately captures the functional aspects of enzymes that are most relevant to their interaction with pesticides. This targeted approach reduces computational complexity and enhances prediction accuracy, providing more reliable insights into enzyme activity. The ligand-binding site predictions enable the identification of key residues involved in pesticide degradation. Understanding these critical interaction points can inform the design of more effective bioremediation strategies and the development of enzyme-based products for agricultural applications.

5 Discussion

5.1 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 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 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.

5.2 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 UniProt 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.

6 Conclusion

6.1 Summary of Findings

In this thesis, a Deep Learning model was developed to predict the enzymatic function based on the protein sequence. The research addressed the need for more accurate and efficient methods to predict enzyme classes responsible for pesticide degradation. The key findings of the research are summarized as follows:

1. Novel Deep Learning Model: The developed model leverages enzyme binding site predictions to enhance the accuracy of enzyme class predictions. This approach focuses specifically on the ligand-binding sites, offering a more detailed and precise prediction by targeting critical interaction regions. With the use of the P2RANK tool, the model can identify specific residues involved in catalytic processes, improving the accuracy of enzyme class predictions. Focussing on ligand-binding pockets has shown to be more effective than traditional methods that utilize the entire protein sequence.

2. Data Preprocessing and Feature Engineering: Comprehensive data preprocessing steps, including cleaning, tokenization, and padding of sequences, ensured the dataset's quality and consistency. Feature engineering incorporated biochemical properties such as molecular weight, isoelectric point, hydrophobicity, and sequence length, which significantly contributed to the model's predictive power. The inclusion of these features enhance the model's ability to learn from the data and make accurate predictions.

3. Handling Data Imbalance: The study addressed the dataset's imbalance using the RandomUnderSampler algorithm, which improved the model's ability to learn from underrepresented classes, although further optimization is necessary for perfect balance. There is a need for more data in UniProt to improve the model's performance, particularly for the 4th EC level, which has a large number of classes and is challenging to predict accurately. As shown in chapter 3.2

4. Model Performance: The model demonstrated high predictive accuracy, particularly at the first four levels of the Enzyme Commission (EC) hierarchy. Existing models struggle to predict the 4th EC level due to the large number of classes and the dataset's imbalance. The developed model showed promising results at this level, indicating its potential to accurately predict enzyme classes in detail. The model's performance was further improved through hyperparameter tuning and data balancing techniques.

5. Implications for Environmental Science: The findings highlight the model's potential to accelerate the development of environmentally friendly agricultural products by reducing the time and cost associated with experimental methods. Accurate enzyme class predictions facilitate better risk assessments and the development of sustainable bioremediation strategies.

This thesis makes significant contributions to the fields of computational biology and environmental science by developing an innovative model. By leveraging modern Deep Learning techniques, this study enhances the accuracy and efficiency of enzyme function predictions, offering a substantial improvement over traditional methods. Moreover the model could save time and money in the development of new products, as well as reduce the risk of harmful pesticides in the environment.

The practical implications of this research are profound, particularly for sustainable agriculture. By enabling more precise predictions of enzyme-mediated pesticide degradation, the model supports the development of safer and more effective bioremediation strategies. This aligns with global efforts to reduce the environmental impact of pesticide use and promote sustainable agricultural practices. Moreover, the efficiency and accuracy of this model can lead to significant cost savings in the development and testing of new agricultural products, facilitating the faster introduction of environmentally friendly solutions to the market.

In addition to its environmental benefits, the research highlights the economic advantages of biotransformation. Utilizing enzymes to facilitate chemical reactions in pesticide degradation can significantly reduce production costs, making

agricultural practices more sustainable and cost-effective. By predicting previously unknown enzyme classes involved in pesticide degradation, this model opens up new opportunities for developing innovative bioremediation solutions that are both environmentally friendly and economically viable. Traditional methods for determining enzyme classes are labor-intensive and time-consuming, making the model a valuable tool for accelerating the discovery of new enzymes and their functions.

6.2 Final Remarks and Future Work

This thesis presents a *novel* Deep Learning model designed to predict enzyme classes based on their protein sequence. By leveraging detailed biochemical features and focusing on ligand-binding sites, the model demonstrates significant improvements on the first three levels and a slight increase on the 4th level in predictive accuracy compared to existing methods. The comprehensive data preprocessing, integration of sequence-based and property-based features, and the use of Deep Learning techniques such as LSTM layers have proven to be highly effective in capturing the complex nature of enzyme functions. The model's performance was further enhanced through Hyperparameter tuning and data balancing techniques, demonstrating its potential to accurately predict enzyme classes. The implications of this research are far-reaching, with significant benefits for environmental science. Unknown enyzmes can be predicted accurately and used to develop new products. The model can also save time and money in the development of new products, as well as reduce the risk of harmful pesticides in the environment.

Despite the significant advancements presented in this thesis, there remains substantial potential for further improvement and refinement of the model. One potential enhancement is the integration of anchor sequences, as utilized in ECPred, to improve the decision-making power of the model. Anchor sequences are specific segments within proteins that are highly conserved and play crucial roles in their function. By focusing on these sequences, the model can gain a deeper understanding of the critical regions that determine enzyme activity. This approach can help in identifying key residues involved in substrate binding and catalysis, thereby

improving the accuracy of enzyme classification. Anchor sequences provide essential structural and functional information that can enhance the model's ability to predict enzyme classes. They act as reliable markers for specific enzyme functions, allowing the model to make more informed predictions. Integrating anchor sequences into the feature set can help the model focus on the most relevant parts of the protein, improving its decision-making capabilities.

Another promising direction is the use of embeddings from ProtBERT, a pre-trained language model for protein sequences. ProtBERT embeddings capture rich contextual information from amino acid sequences, enabling the model to understand complex patterns and dependencies within the data. By incorporating these embeddings, the model can leverage the extensive knowledge encoded in ProtBERT, potentially enhancing its predictive performance ProtBERT embeddings provide a comprehensive representation of protein sequences, capturing both local and global sequence information. This can help the model generalize better across different enzyme classes and improve its ability to recognize subtle variations in sequence that are critical for enzyme function.

Further refinement of data augmentation techniques can address the issue of class imbalance more effectively. While RandomUnderSampling has been used in this study, exploring other methods such as SMOTE (Synthetic Minority Over-sampling Technique) or advanced generative models to create synthetic data for underrepresented classes can improve the model's performance on these classes. Also incorporating more diverse biochemical and environmental data can enhance the model's predictive accuracy. For example, integrating data on enzyme kinetics, thermodynamic stability, and environmental factors affecting enzyme activity can provide a more holistic view of enzyme function. For the future work, the model will be used to predict the enzyme class of previously unknown enzymes and validate the predictions through *in vitro* or even in *in vivo* studies. This will help to further refine the model and demonstrate its practical applicability in real-world scenarios.

In conclusion, the Thesis represents a significant advancement in the field of enzyme classification. By integrating advanced Deep Learning techniques with detailed

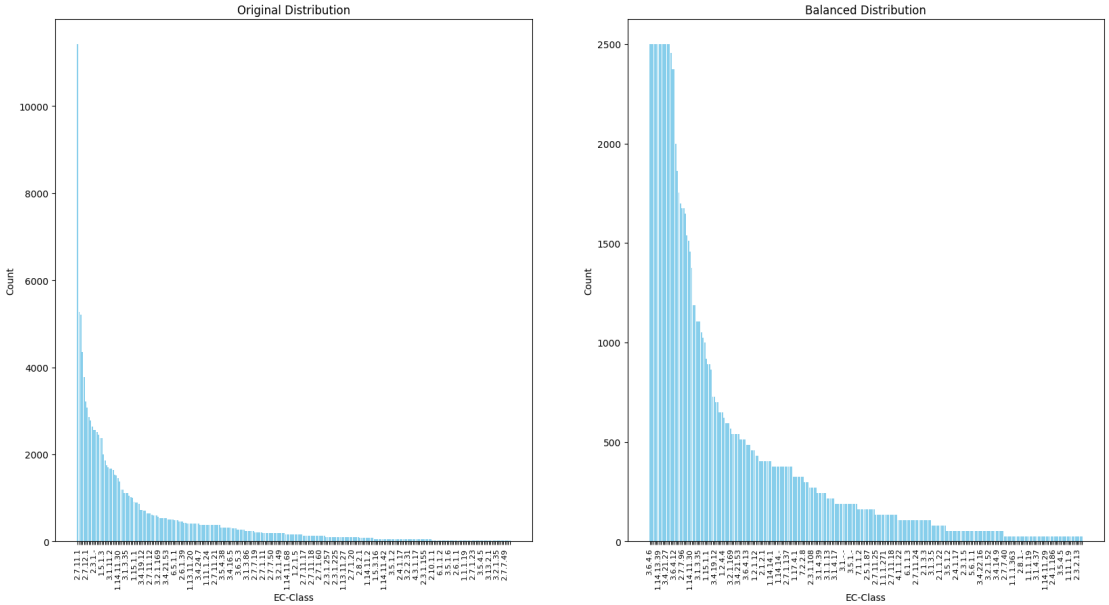
biochemical features, the model offers a powerful tool for computational biology. Future research should focus on incorporating anchor sequences, utilizing ProtBERT embeddings, refining data augmentation techniques, integrating additional data, and optimizing the model architecture to further enhance its capabilities and broaden its applicability.

Appendix

Appendix

Appendix 0.1 Figure 6

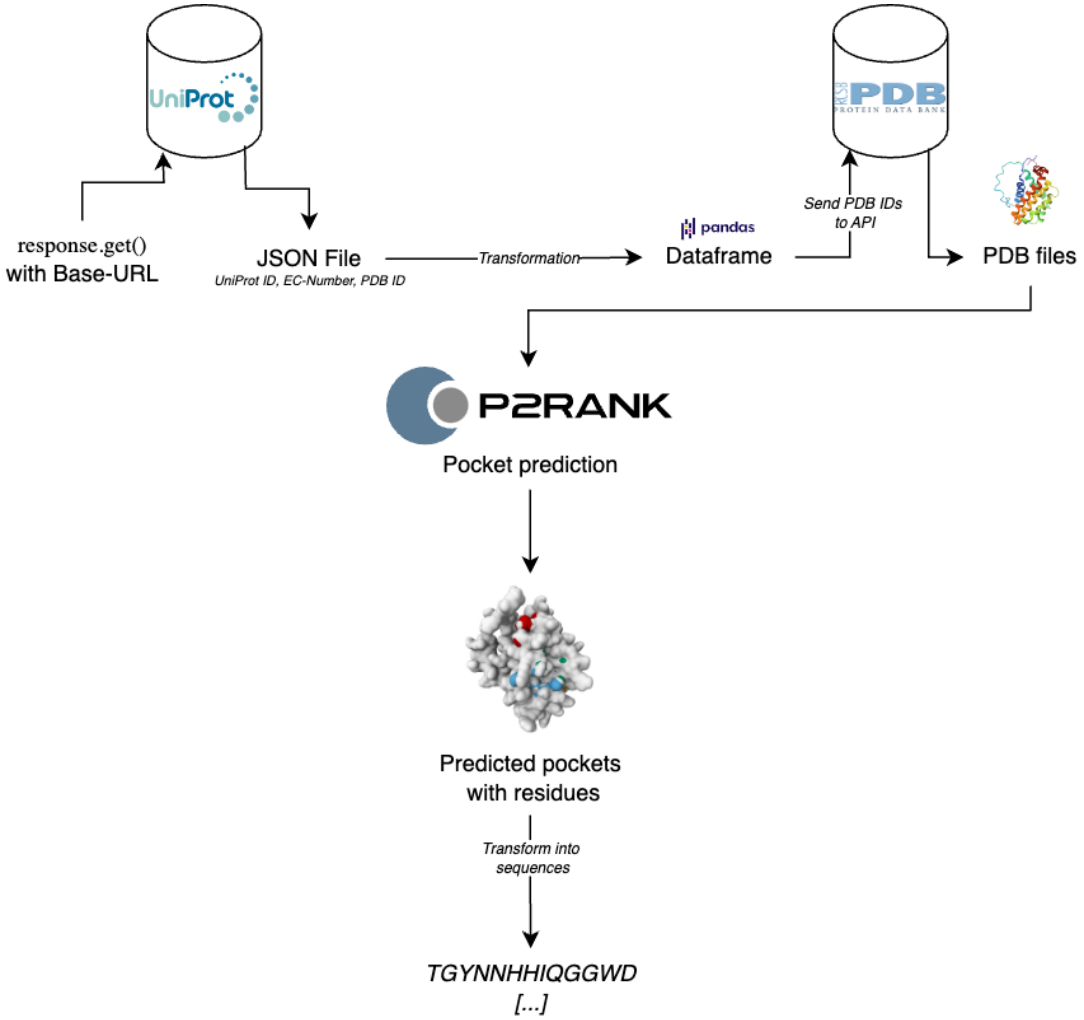
Figure 10: Distribution of EC classes before and after balancing (Full size)



Source: Own illustration

Appendix 0.2 Figure 11

Figure 11: Data Collection & P2Rank Process (Full size)



Source: Own illustration

References

References

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

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

[3] Thomas Shafee, *English: Organisation of enzyme structure and lysozyme example. Binding sites in blue, catalytic site in red and peptidoglycan substrate in black. (PDB: 9LYZ)*, Dec. 2015. (visited on Jul. 8, 2024).

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

[5] 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).

[6] 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).

[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] Iqbal H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, no. 6, p. 420, Aug. 2021. (visited on Jul. 8, 2024).

[9] 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).

[10] 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).

[11] 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).

[12] Robin M. Schmidt, *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*, Nov. 2019. arXiv: `1912.05911 [cs, stat]`. (visited on Jul. 8, 2024).

[13] 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).

[14] Moumita Tora, Jianhui Chen, and J.J. Little, "Classification of Puck Possession Events in Ice Hockey," *Proceedings / CVPR, IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jul. 2017.

[15] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges* (The Springer Series on Challenges in Machine Learning). Cham: Springer International Publishing, 2019, ISBN: 978-3-030-05317-8 978-3-030-05318-5. (visited on Jul. 8, 2024).

[16] James Bergstra and Yoshua Bengio, "Random Search for Hyper-Parameter Optimization,"

[17] G. Sudhamathy and N. Valliammal, "The Bayesian CNN-LSTM classification model to predict and evaluate learner's performance," *International Journal of Applied Science and Engineering*, vol. 20, no. 4, pp. 1–9, 2023. (visited on Jul. 8, 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] *Protein Structure | Learn Science at Scitable*, https://www.nature.com/scitable/topicpage/protein-structure-14122136/. (visited on Jul. 7, 2024).

[23] Samuel H. Light, Laty A. Cahoon, Kiran V. Mahasenan, Mijoon Lee, Bill Boggess, Andrei S. Halavaty, Shahriar Mobashery, Nancy E. Freitag, and Wayne F. Anderson, "Transferase Versus Hydrolase: The Role of Conformational Flexibility in Reaction Specificity," *Structure*, vol. 25, no. 2, pp. 295–304, Feb. 2017. (visited on Jul. 7, 2024).

[24] Fatemeh Khosravi, Ehsan Fard, Marzieh Hosseininezhad, and Hadi Shoorideh, "Identification and characterization of inulinases by bioinformatics analysis of bacterial glycoside hydrolases family 32 (GH32)," *Engineering in Life Sciences*, vol. 23, Jul. 2023.

[25] P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia, "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, no. 2, pp. 184–192, Feb. 2020. (visited on Jun. 22, 2024).

[26] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial, "ProteinBERT: A universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, Apr. 2022. (visited on Jul. 7, 2024).

[27] 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).

[28] Alperen Dalkiran, Ahmet Sureyya Rifaioglu, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan, "ECPred: A tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature," *BMC Bioinformatics*, vol. 19, no. 1, p. 334, Dec. 2018. (visited on Jul. 7, 2024).

[29] Chengxin Zhang, Peter L. Freddolino, and Yang Zhang, "COFACTOR: Improved protein function prediction by combining structure, sequence and protein–protein interaction information," *Nucleic Acids Research*, vol. 45, no. W1, W291–W299, Jul. 2017. (visited on Jul. 8, 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 8, 2024

Tobias Polley