REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Biomedical Sciences Program

DISCRIMINATION OF ACIDIC AND ALKALINE ALPHA/BETA HYDROLASE ENZYMES THROUGH DATA-DRIVEN MACHINE LEARNING APPROACHES

Maryam Ahmed Abed ALGHABAWI

Master's Thesis

Supervisor

Asst. Prof. Dr. Nurcan VARDAR YEL

İstanbul, 2025

DISCRIMINATION OF ACIDIC AND ALKALINE ALPHA/BETA HYDROLASE ENZYMES THROUGH DATA-DRIVEN MACHINE LEARNING APPROACHES

Maryam Ahmed Abed ALGHABAWI

Biomedical Sciences

Master’s Thesis

ALTINBAŞ UNIVERSITY

2025

The thesis titled DISCRIMINATION OF ACIDIC AND ALKALINE ALPHA/BETA HYDROLASE ENZYMES THROUGH DATA-DRIVEN MACHINE LEARNING APPROACHES prepared by MARYAM AHMED ABED ALGHABAWI and submitted on 22/12 /2025 has been **accepted unanimously** for the degree of Master of Science in Biomedical Science.

Asst. Prof. Dr. Nurcan VARDAR YEL

Supervisor

Thesis Defence Committee Members:

Asst. Prof. Dr. Nurcan VARDAR YEL	Department of Medical Services and Techniques, Altınbaş University	_____
Asst. Prof. Dr. İpek ADA ALVER	Department of Medical Services and Techniques, Altınbaş University	_____
Asst. Prof. Dr. Meryem Menekşe Kılıç	Department of Medical Services and Techniques, İstanbul Topkapı University	_____

I hereby declare that this thesis meets all format and submission requirements for a Master's thesis.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Maryam Ahmed Abed ALGHABAWI

Signature

DEDICATION

I dedicate this thesis first and foremost to Almighty Allah, whose mercy, wisdom, and guidance have granted me the strength and perseverance to complete this work.

With profound respect and sincere appreciation, I would like to express my deepest gratitude to my supervisor, Assist. Prof. Dr. Nurcan Vardar Yel, for her continuous guidance, invaluable feedback, and exceptional academic support throughout the course of this research. Her commitment to scientific rigor and her encouragement at every stage of this work have been truly instrumental in its completion.

My heartfelt thanks extend to the faculty members and staff of the university for providing an encouraging academic environment and for their kind cooperation whenever needed.

I am deeply grateful to my family for their endless love, patience, and moral strength, and to my husband Humam, whose unwavering support and motivation have accompanied me every step of the way.

ABSTRACT

DISCRIMINATION OF ACIDIC AND ALKALINE ALPHA/BETA HYDROLASE ENZYMES THROUGH DATA-DRIVEN MACHINE LEARNING APPROACHES

ALGHABAWI, Maryam Ahmed Abed

M.Sc., Biomedical Science, Altınbaş University

Supervisor: Asst. Prof. Dr. Nurcan VARDAR YEL

Date: 12/2025

Pages: 67

Alpha/beta-hydrolases form a broad and versatile enzyme superfamily that includes lipases, esterases, carboxylesterases, and amidases key catalysts in both cellular metabolism and industrial biotechnology. Although they share a similar α/β fold and catalytic triad, members of this family often exhibit very low sequence identity, which makes it challenging to classify them solely based on sequence similarity. This challenge becomes even greater in extremophilic hydrolases that have adapted to survive and function in very acidic or alkaline environments.

In this study, we aimed to classify α/β-hydrolase enzymes originating from acidophilic and alkaliphilic microorganisms by using a machine learning approach. A total of 403 bacterial enzymes were analysed based on their amino acid composition and physicochemical characteristics, without relying on secondary structure information. Several supervised learning models were tested, including Decision Tree, Random Forest, eXtreme Gradient Boosting, and Support Vector Machine. Among them, the Random Forest model showed the highest performance, reaching 76% accuracy with an AUC value of 0.90 ± 0.03. Alanine, lysine, and the grand average of hydropathicity (GRAVY) emerged as the most influential features in distinguishing acidic and alkaline enzymes. Overall, the results suggest that integrating sequence-derived and chemical descriptors through machine learning can effectively separate acidic from alkaline α/β-hydrolases. The findings not only shed light on

molecular adaptation mechanisms to extreme pH but also provide a computational basis for guiding the discovery and engineering of new extremozymes for industrial applications such as bioremediation, detergent design, and biomining.

Keywords: Alpha/Beta-hydrolase, Machine learning, Enzyme Classification, Extremophiles, Acidophilic Enzymes, Alkaliphilic Enzymes.

TABLE OF CONTENTS

Pages

ABSTRACT vi

LIST OF TABLES..... x

LIST OF FIGURES..... xi

ABBREVIATIONS..... xii

1. INTRODUCTION 1

2. GENERAL INFORMATION ……..…….3

2.1 GENERAL CHARACTERISTICS OF ACIDOPHILES AND ALKALIPHILES…3

2.2 ACIDIC AND ALKALINE ENZYMES: CHARACTERISTICS AND SIGNIFICANCE.……4

2.2.1 pH-dependent Activity and Stability…….....5

2.2.2 Industrial and Environmental Applications……..…..6

2.3 ALPHA BETA HYDROLASES ……..…..……...7

2.3.1 Structural Features of Alpha/Beta Hydrolases…….....……8

2.3.2 Biological Functions and Diversity of Alpha/Beta Hydrolases ……..…..…...10

2.4 MACHINE LEARNING IN ENZYME CLASSIFICATION..……..…..…...……...14

2.4.1 Overview of Machine Learning Techniques…….……...…...…….…16

2.4.2 Previous Applications of ML in Enzymology ……..…….....…...17

2.5 RATIONALE AND OBJECTIVES OF THE STUDY……....……...……...….....19

2.5.1 Gaps in Current Classification Methods …...……..……19

2.5.2 Study Goals and Expected Contributions…..……...……20

3. METHODOLOGY…….……22

3.1 THE DATASET..……..…….……..22

3.2 MACHINE LEARNING ANALYSIS…….…..…….……..32

3.2.1 Cross-Validation..……..…….……..……..32

3.2.2 Comparison of Different Algorithms ….…….……..32

3.2.2.1 Random forest (RF)..……..…….…….……..33

3.2.2.2 Decision trees (DTs)..……..…….……..33

3.2.2.3 Extreme gradient boosting (XGBoost)..……..…….……..34

3.2.2.4 Support vector machines (SVMs)..……..……...…34

3.2.3 Hyperparameter Tuning..……..……..……..…….34

3.2.4 Performance Evaluation of Machine Learning Models…….…….35

4. RESULTS……37

4.1 DATA COLLECTION AND CROSS-VALIDATION PROCEDURES……..……37

4.2 FEATURE IMPORTANCE ANALYSIS……..……....…40

4.3 CORRELATION MATRIX…….…….……..…44

5. DISCUSSION AND CONCLUSION……46

REFERENCES 49

LIST OF TABLES

Pages

Table 3.1: Acidic α/β-hydrolases Screened From Different Bacterial Sources……..….23

Table 3.2: Alkaline α/β-hydrolases Screened From Different Bacterial Sources.……..….28

Table 4.1: Accuracy of Models Using 5-Fold Cross-Validation with a 70:30 Train-Test Split……..…….……..….37

Table 4.2: Comparative Results of SHAP and Feature Importance.……...….42

LIST OF FIGURES

Pages

Figure 2.1: Schematic Representation of the Canonical α/β-hydrolase Fold 10

Figure 4.1: a, Confusion Matrix for RF used in the Prediction of Acidic Bacterial α/β Hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; b, ROC curve with 5-fold Cross-validation with the RF Algorithm 38

Figure 4.2: a, Confusion Matrix for DT Used in the Prediction of Acidic Bacterial α/β Hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; b, ROC Curve with 5-fold Cross-Validation Using the DT Algorithm..... 39

Figure 4.3: a, Confusion Matrix for SVM Used in the Prediction of Acidic Bacterial α/β Hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; b, ROC Curve with 5-fold Cross-validation Using the SVM Algorithm. 39

Figure 4.4: a, Confusion Matrix for XGBoost Used in the Prediction of Acidic Bacterial α/β Hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; b, ROC Curve With 5-fold Cross-validation Using the XGBoost Algorithm. 40

Figure 4.5: Feature Importance Results for Random Forest Algorithm..... 41

Figure 4.6: SHapley Additive exPlanations (SHAP) Analysis 43

Figure 4.7: Feature Correlation Matrix Depicting Variable Interrelationships 44

ABBREVIATIONS

AAC	:	Amino Acid Composition
AUC	:	Area Under the Curve
DT	:	Decision Tree
FN	:	False Negative
FP	:	False Positive
GBDT	:	Gradient-Boosted Decision Tree
GRAVY	:	Grand Average of Hydropathicity
P	:	Precision
R	:	Recall
RF	:	Random Forest
ROC	:	Receiver Operating Characteristic
SVM	:	Support Vector Machine
TF	:	True Positive
TN	:	True Negative
UniProt	:	The Universal Protein Knowledgebase
XGBoost	:	eXtreme Gradient Boosting

1. INTRODUCTION

Extremophilic bacteria adapted to survive in extremely acidic and alkaline conditions are known as acidophiles and alkaliphiles, respectively (Gupta et al., 2014; Kumar et al., 2022). A class of microorganisms known as acidophiles, or acid-loving organisms, are organisms that thrive in acidic conditions; they typically grow best in environments with a pH value below 5.0 (Johnson & Quatrini, 2020; Seckbach, 2000; Mirete et al., 2017). Microorganisms that have a pH optimum for growth of less than pH 3 are specifically termed 'acidophiles' (Johnson, 2007; Johnson & Aguilera, 2016; Ghosh & Banerjee, 2022). To grow at low pH, acidophiles must maintain a pH gradient of several units across the cellular membrane while producing ATP via the influx of protons through the F0F1 ATPase (Matin, 2009; Booth, 1985). Recent genomic and biochemical analyses have revealed that acidophiles share distinctive structural and functional characteristics, including a reversed membrane potential, highly impermeable cell membranes, and a predominance of secondary transporters. Once protons enter the cytoplasm, additional mechanisms are required to alleviate the effects of lowered internal pH (Slonczewski et al., 2009; Mirete et al., 2017). Within the group of iron-oxidizing prokaryotes, acidophilic microorganisms include *Thiobacillus ferrooxidans, Sulfobacillus acidophilus*, and *Acidianus brierleyi*, which are mesophiles, moderate thermophiles, and extreme thermophiles, respectively (Johnson, 1998; Kanekar and Kanekar, 2022). *Sulfolobus shibitae, T. caldus*, and *T. thiooxidans* are notable members of the sulfur-oxidizing prokaryote group (Johnson & Hallberg, 2008; Hallberg & Johnson, 2003). *Thermoplasma acidophilum, Alicyclobacillus* spp., and *Sf. acidocaldarium* are classified as moderate to extreme thermophiles, while bacteria like *Acidophilium* spp., *Acidobacterium capsulatum*, and *Acidomonas methanolica* are classified as mesophiles in the heterotrophic prokaryote group (Johnson, 1998; Hallberg & Johnson, 2003). While some acidophilic microbes are naturally occurring, others are the result of human or industrial activity. Typical habitats for them include volcanic soils, sulfur hot springs, and acid mine drainage (Johnson & Aguilera, 2016; Razia et al., 2023; Sharma et al., 2016). Furthermore, drainage water from ore mines, metal ore piles, coal waste piles, and coal mines contains acidophilic organisms (Dave & Tipre, 2012; Johnson, 1995; Hallberg & Johnson, 2003). Acidophilic bacteria and archaea have been found to include a wide range of enzymes, such as cellulases, acid phosphatases, amylolytic, xylanolytic, and proteolytic enzymes, as well as maltose-binding proteins (Sharma et al., 2012; Ghosh & Banerjee, 2022). Many of these

enzymes are active at both low pH and high temperatures, especially those derived from thermoacidophiles. Because of these qualities, they are useful in a variety of commercial settings, such as the feed, bread, fruit juice, and starch sectors (Sharma et al., 2016; Sharma et al., 2012; Kour et al., 2019). Acidophiles maintain intracellular pH stability through several mechanisms, including highly impermeable cell membranes, active proton extrusion, cytoplasmic buffering, and organic acid degradation, enabling them to tolerate extreme acidity (Mirete et al., 2017; Matin, 2009; Slonczewski et al., 2009; Booth, 1985). As a result, certain animals have not evolved adaptations for acid stability in their intracellular proteins. But in some acidophiles, such as *Acetobacter aceti*, the acidic cytoplasm forces nearly every protein in the genome to evolve acid resistance mechanisms (Trček et al., 2015; Qiu et al., 2021; Menzel & Gottschalk, 1985).

In this study, enzymes from the acidic and alkaline α/β hydrolase families were distinguished by analysing their amino acid sequences and specific physicochemical properties. This was achieved by analysing both acidic and alkaline α/β hydrolase enzymes using a machine learning technique, focusing on their primary structures and physicochemical properties without considering secondary structural changes. This approach enabled a more detailed understanding of how these enzymes can be categorized based on their fundamental properties.

2. GENERAL INFORMATION

2.1 GENERAL CHARACTERISTICS OF ACIDOPHILES AND ALKALIPHILES

Acidophiles and alkaliphiles represent two interesting groups of extremophilic microorganisms that have learned to live under conditions which would destroy most forms of life. Optimal acidophile growth occurs at very low pH, typically below 3, whereas alkaliphiles enjoy living at the other end of the scale, growing best at pH values above 9. In an effort to maintain a stable internal environment, both have evolved elaborate systems that allow them to function efficiently in surroundings that would quickly denature normal cellular components (Satyanarayana et al., 2016; Tütüncü & Vardar-Yel, 2022).

Acidophiles are usually found in acid mine drainage, geothermal vents, and volcanic fields; such places usually have a rich amount of metal ions and sulfur compounds. The usual threats that such an environment may pose to acidophiles include proton invasion into the cells. Several defense strategies have been developed to help the acidophiles cope with these conditions. Unusually impermeable proton cell membranes exist owing to unique structure lipid features, which serve as effective barriers. Simultaneously, the protons are continuously expelled by the proton pumps such as H^+-ATPases to keep the cytoplasmic pH near neutrality, while potassium uptake mechanisms play an important role in creating a positive internal charge that helps repel protons. In some species, amino acid decarboxylation reactions and cytoplasmic buffering compounds provide additional protection. Being specifically stable in low-pH environments, the enzymes of acidophiles are often rich in acidic amino acids, which help them maintain proper folding and solubility when under extreme acidity conditions (Tiquia-Arashiro & Rodrigues, 2016).

Alkaliphiles, on the other hand, are organisms that live in highly basic environments such as soda lakes, alkaline soils, and industrial effluents. Their challenge is essentially the opposite: instead of avoiding proton invasion, they must capture enough protons to sustain vital processes. They fulfil this need through elaborate ion-exchange systems-most notably, Na^+/H^+ and K^+/H^+ antiporters-capable of controlled proton uptake. The surfaces of the cells of alkaliphiles often contain acidic polymers that attract and bind protons, serving to preserve the proton motive force necessary for ATP synthesis. Some alkaliphiles adjust their metabolism by the production of organic acids, lowering the pH of the cytoplasm locally. In

addition, their ATP synthase enzymes have adapted to operate efficiently even when the external proton concentration is very low (Mamo & Mattiasson, 2015).

Acidophiles and alkaliphiles, though living in contrasting environments, have the same ultimate aim-to maintain internal pH homeostasis. Acidophiles minimize proton entry, while alkaliphiles maximize proton capture. In both cases, however, balance is achieved through a combination of membrane adaptation, ion transport, and biochemical buffering. These remarkable physiological solutions highlight the flexibility of microbial life and its ability to push the boundaries of biochemistry. From an applied point of view, these organisms are much more than scientific curiosities. Acidophiles are valued in biomining and metal recovery owing to their ability to oxidize iron and sulfur compounds, while alkaliphiles are appreciated for their stable enzymes, which remain active within detergents, food processing, and paper industries. Their robustness and catalytic properties, arising from the same physiological bases of their existence, make them promising candidates for applied purposes requiring hardy biological catalysis under adverse chemical conditions (Tütüncü & Vardar-Yel, 2022).

2.2 ACIDIC AND ALKALINE ENZYMES: CHARACTERISTICS AND SIGNIFICANCE

The efficacy of enzymes is highly dependent on their environment pH. The ionization of residues surrounding the catalytic site of the enzyme plays an important role. The members of the α/β-hydrolases superfamily have adapted themselves well in highly pH-variable conditions. As a result, two broad classes of enzymes followed from this adaptation—the acidic hydrolases and alkaline hydrolases. Nature's capability of adapting molecules into opposite pH conditions can be witnessed from biocatalysts. Acid hydrolases act effectively in low pH/proton-rich regions, while alkaline hydrolases act effectively in highly basic/hydroxide-rich regions (Robinson, 2015). This amazing biochemical flexibility has drawn ever more researchers' interest, since pH adaptation mechanisms hold the key to unravel various questions of enzyme evolution, flexibility, and adaptability. Apart from their ecological functions in biogeochemical cycles and stress responses, acidic and alkaline hydrolases turned out to be invaluable tools in modern biotechnology, which requires highly stable catalysts able to function effectively despite harsh physicochemical stress conditions (Baker-Austin & Dopson, 2007; Dopson & Johnson, 2012; Horikoshi, 1999). Such enzymes are particularly abundant in extremophilic microorganisms like *Acidithiobacillus*,

Sulfolobus, and Bacillus species, which have evolved hydrolases that maintain catalytic activity and structural stability under extreme acidic or alkaline pH through specialized amino acid composition and adaptive folding mechanisms (Tiquia-Arashiro & Rodrigues, 2016; Satyanarayana et al., 2016; Tütüncü & Vardar-Yel, 2022).

2.2.1 pH-dependent Activity and Stability

The catalytic activity of enzymes is markedly pH-dependent, with pH regulating both the ionization state of active-site residues and the structural stability of the protein. Enzymes typically exhibit a bell-shaped pH-activity curve, with optimal catalytic efficiency at an optimum pH that ensures the correct protonation state of key residues (Horikoshi, 1999). Deviation from the optimum results in altered electrostatic interactions, conformational instability, and reduction in turnover numbers. Acidic hydrolases, which have evolved in pH conditions lower than 4, are structurally highly resilient to proton-rich environments. They possess surfaces that are typically enriched with acidic amino acids, i.e., aspartate and glutamate, which enable solvation and prevent denaturation. The enzymes also typically possess reinforced intramolecular hydrogen bonding and compact hydrophobic cores that are resistant to acid-induced unfolding (González et al., 2024). These adaptations are similar to those observed in acidophilic microorganisms, whose enzymes maintain net negative surface charge, ensuring stability and solubility in acidic habitats (Mamo & Mattiasson, 2015). In contrast, alkaline hydrolases adjusted to high-pH environments possess an excess of basic residues (lysine, arginine, histidine) on their surface, which provides charge compensation and electrostatic stability under hydroxide-rich conditions. Moreover, alkaliphilic hydrolases often show modified hydrogen-bonding networks and increased salt-bridge interactions that contribute to enhanced folding stability in low-proton environments (Tiquia-Arashiro & Rodrigues, 2016; Satyanarayana et al., 2016).

Also, the microenvironment around the catalytic site is very crucial for retaining enzymatic function under extreme pH. Acid-stable enzymes often contain a downward shift in the pKa values of catalytic residues like aspartate and glutamate, enabling them to remain ionized and catalytically active even in proton-rich environments (Sharma et al., 2016). On the other hand, alkaline enzymes exhibit increased pKa values of lysine and histidine residues, which maintain proton-accepting capabilities under alkaline stress (Horikoshi, 1999). At the structural level, extremozymes often show increased disulfide bonding, tight tertiary

packing, and hydrophobic shielding of core residues, which prevent denaturation by protonation or deprotonation events (Gupta et al., 2014; Razia et al., 2023). Another characteristic feature that most extremophilic hydrolases have in common, especially those produced by thermoacidophiles and haloalkaliphiles, is that their pH and thermal stability coexist. They usually retain activity in a wide pH and temperature range (Mirete et al., 2017; Kour et al., 2019).

On an applied level, pH-tolerant hydrolases are of immense biotechnological significance. Acid-stable enzymes have been used in various applications, such as the clarification of fruit juice, processing of starch, and treatment of leather, while alkaline-stable hydrolases-essentially lipases and proteases-find a place in the formulation of detergents and in bioremediation processes (Mamo & Mattiasson, 2015; Tiquia-Arashiro & Rodrigues, 2016; Satyanarayana et al., 2016). Such dual stability regarding pH and temperature underlines their industrial relevance as robust biocatalysts for processes that require harsh operational conditions.

2.2.2 Industrial and Environmental Applications

The high stability as well as catalytic flexibility of acidic and alkaline hydrolases has ensured that they play an indispensable role in many industrial processes. Moreover, their catalytic flexibility enables them to act as substitutes for chemical catalysts in various production processes. Esterases and lipases of *Acidithiobacillus ferrooxidans*, which belong to the group of acidophilic hydrolases with particular interest in bioleaching, play an important role in the degradation of the organic matrix that envelops metal sulphides by improving copper, gold, and nickel metal mobilization (Bornscheuer, 2002; Kirk et al., 2002). The enzymes help in the oxidation of sulfur, which ensures that a low pH prevails. Similarly, other acid-stable hydrolases from genera such as *Sulfolobus* and *Picrophilus* have shown potential in polymer degradation and organic waste treatment under acidic industrial conditions (Tütüncü & Vardar-Yel, 2022). However, there has been an increase in the use of alkaline hydrolases in the detergent industry, where proteases and lipases from Bacillus species show efficient catalytic activity in highly basic compositions (Horikoshi, 1999; Maurer, 2004). Their tolerance properties of surfactants, oxidizing agents, and thermal stability make them highly preferable in the removal of proteinaceous or lipids in basic pH (pH 10-11). Alkaline proteases find uses in cheese ripening, while lipases and esterases with heat-stable activity

in an acidic environment help fermentation in the production of cocoa, as well as the clarification of juice by breaking down barriers of lipids or pectin (Gupta et al., 2002). In addition, alkaliphilic esterases and cellulases derived from haloalkaliphilic bacteria have recently been applied in textile and paper processing, where their ability to withstand high salt and high pH improves process sustainability (Mamo & Mattiasson, 2015; Satyanarayana et al., 2016). Environmentally, both acidic and alkaline hydrolases can play pivotal roles in bioremediation. Acid-stable enzymes find application in breaking down contaminants in acidic industrial wastewaters, including those from tanneries and dyestuff manufacturing units (Huang et al., 2021). Alkaline hydrolases, on the other hand, can find application in eliminating organic contaminants from wastewaters with higher pH values from textile mills and paper production units (Kirk et al., 2002). The use of these extremophilic hydrolases in all of these industries not only enhances the efficiency of the process but also reduces the effect on the environment by decreasing the use of corrosive chemicals and energy. The use of hydrolases represents how biotechnology can help modernize traditional industries. These applications clearly demonstrate that extremophilic enzymes are not merely laboratory curiosities but vital components of sustainable biotechnological innovation (Tiquia-Arashiro & Rodrigues, 2016; Tütüncü & Vardar-Yel, 2022).

2.3 ALPHA BETA HYDROLASES

The α/β-hydrolases are a superfamily of enzymes comprising one of the largest functional families of enzymes found in nature. The enzymes from this superfamily all share a common structural framework, known as the α/β-hydrolase fold, which enables them to perform a wide range of catalytic activities, including hydrolysis, isomerization, or transesterification (Nardini & Dijkstra, 1999). α/β-hydrolases from this superfamily can be found in bacteria, archaea, as well as eukaryote. The functional adaptability of this superfamily of enzymes has seen it take center stage in a number of various natural phenomena. Due to their functional diversity as well as their utility in industry, α/β-hydrolases have been the subject of interest in research efforts trying to understand the structural basis of enzyme specificity and stability, not least in adverse environment conditions of pH (Ollis & Carr, 2009).

2.3.1 Structural Features of Alpha/Beta Hydrolases

The common structural framework of the α/β-hydrolases is known as the α/β-hydrolase fold. The eight-stranded β-sheet is flanked on both sides by α-helices. Alpha/beta hydrolases (ABHs) represent one of the most structurally conserved and functionally diverse enzyme superfamilies in biology. Despite low sequence similarity among members, all share a characteristic α/β-fold architecture, which forms a stable catalytic scaffold accommodating a broad range of substrates and catalytic activities (Nardini & Dijkstra, 1999). The canonical fold consists of a central eight-stranded mostly parallel β-sheet (with the second strand antiparallel) flanked by α-helices on both sides. This configuration produces a left-handed superhelical twist, where the first and last strands intersect at approximately 90°, providing structural stability (Nardini & Dijkstra, 1999). A defining feature of ABHs is the nucleophile elbow, a sharp β–α turn that harbors the nucleophilic residue of the catalytic triad, typically composed of a nucleophile (Ser, Cys, or Asp), an acidic residue (Asp or Glu), and an absolutely conserved His residue (Lenfant et al., 2013). This triad is spatially arranged to enable efficient hydrolysis via proton transfer and stabilization of the transition state. The oxyanion hole, formed by backbone amide nitrogen or side chains (often glycine, tyrosine, or aspartate), stabilizes negatively charged intermediates during catalysis (Bauer et al., 2019). Recent large-scale structural analyses have revealed that ABHs can adopt 12 distinct modular architectures, built upon a conserved catalytic core (Bauer et al., 2019). These modules include lids (one or two α-helices that regulate substrate access), caps (static α-helical domains above the active site), and N- or C-terminal auxiliary domains such as Rossmann or β-sandwich folds. Such modular additions diversify substrate specificity and confer functional adaptability while maintaining the same catalytic core geometry. For instance, the lid domain observed in lipases regulates the open or closed conformation of the active site, explaining their “interfacial activation” behavior (Nardini & Dijkstra, 1999). Caps and terminal domains may also influence substrate orientation, stability, or interaction with cofactors. Bauer et al. (2019) further classified ABHs based on their oxyanion hole motifs GX-, GGGX-, and Y-types each correlating with substrate preferences and catalytic environments. GX-type hydrolases, for example, often act on linear esters, whereas GGGX-types accommodate bulkier substrates such as tertiary alcohols.

The high degree of fold plasticity allows insertions, deletions, and domain shuffling without compromising catalytic integrity (Nardini & Dijkstra, 1999). This structural resilience explains how members like lipases, dehalogenases, esterases, and proteases have diverged to catalyze chemically distinct reactions. Structural databases such as ESTHER and the Lipase Engineering Database (LED) have been instrumental in identifying conserved motifs and predicting catalytic residues (Lenfant et al., 2013; Bauer et al., 2019).

In summary, α/β hydrolases combine a conserved catalytic core with remarkable architectural variability. Their modular design underlies both evolutionary robustness and functional versatility, making them key models for understanding enzyme evolution, adaptation, and engineering. The β-sheet serves as a structural framework that encompasses the active site of the proteins (Ollis & Carr, 2009). The nomenclature of this structural framework follows the order β1–α1–β2–α2–β3–α3–β4–α4–β5–α5–β6–α6–β7–α7–β8, whereby the β-sheet serves as the structural framework of the protein. The hallmark of α/β-hydrolases is the catalytic triad. The catalytic triad always comprises the nucleophile serine, histidine, and an acidic residue (aspartic or glutamate). The three amino acids take on a three-dimensional structure that facilitates the protonation of the substrate's carbonyl carbon. The nucleophile serine can also take on the role of nucleophile elbow. The nucleophile elbow always comprises residues like Gly–X–serine–X–Gly. The nucleophile elbow position ensures proper alignment of the serine hydroxyl group (Nardini & Dijkstra, 1999). The oxyanion hole is another feature of this enzyme. The role of the oxyanion hole is the stabilization of the tetrahedral intermediate formed during catalysis by hydrogen bonding, which can be from amides or residues. This particular feature allows α/β-hydrolases to catalyze efficiently while still retaining flexibility in binding substrates (Ollis & Carr, 2009). The structural variation of α/β hydrolases despite the preservation of the catalytic triad and the overall fold allows great functional diversity. This is achieved by variations or insertions of peripheral loops. The flexible regions of the protein show specificity for substrates. They can catalyze various chemical reactions. These include hydrolysis of various substrates like ester, amides, epoxide, dehalogenation, or hydroxynitrile lyases (Ollis et al., 1992). Schematic representation of the classical α/β-hydrolase fold highlighting the β-sheet core, flanking α-helices, as well as their three-dimensional disposition in forming the catalytic triad in the active site pocket is shown in Figure 2.1. The characteristic α/β-hydrolase fold features a central eight-stranded β-sheet surrounded by α-helices, while its catalytic triad

consists of a nucleophilic residue, an acidic residue, and a histidine (Lord et al., 2013; Kind & Kursula, 2019).

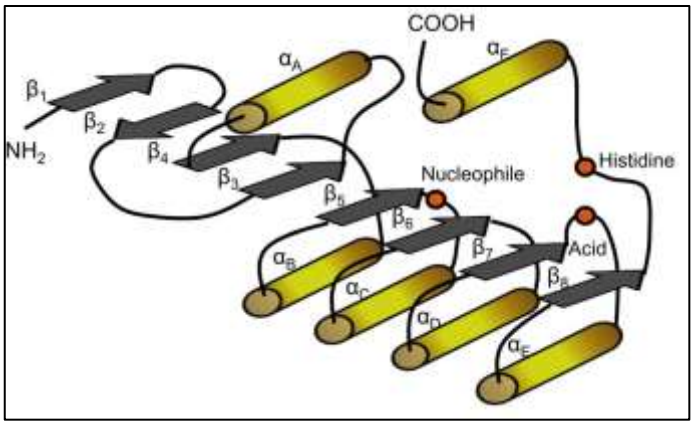

Figure 2.1: Schematic Representation of the Canonical α/β-hydrolase Fold.

2.3.2 Biological Functions and Diversity of Alpha/Beta Hydrolases

The α/β-hydrolases belong to one of the most functionally diverse families of enzymes known from nature. The structural similarity of the members of this superfamily of enzymes belies their immense diversity in substrate specificity, catalytic mechanisms, and functions (Holmquist, 2000; Nardini & Dijkstra, 1999). The biochemistry of the members of this superfamily of enzymes encompasses a whole range of biochemical transformations, including hydrolysis, transesterification, halogenation, and carbon-carbon bond formation (Ollis & Carr, 2009). This diversity is chiefly owing to variations in the peripheral regions of the catalytic domains of α/β-hydrolases, which can affect substrate binding or the specificity of the reaction without deviating from the α/β-hydrolase domain skeleton.

The α/β-hydrolase fold superfamily encompasses an extensive range of enzymes performing hydrolytic, transferase, lyase, and even non-catalytic functions across all domains of life (Lenfant et al., 2013). Initially identified through structural homology among enzymes such as acetylcholinesterase, dienelactone hydrolase, and lipases, the ABH fold has since been recognized in thousands of proteins with widely differing biological roles (Nardini & Dijkstra, 1999). Catalytic diversity arises from variation in substrate-binding loops and associated modular domains. Members include lipases, esterases, epoxide hydrolases, dehalogenases, and proteases, many of which play vital roles in metabolism, detoxification, and signalling (Bauer et al., 2019). Some ABHs, such as the acetylcholinesterases, function

in neurotransmission, while others like epoxide hydrolases are critical in xenobiotic metabolism. Haloalkane dehalogenases detoxify halogenated compounds, contributing to bioremediation. Meanwhile, lipases and esterases have become indispensable biocatalysts in industrial synthesis, owing to their high regio- and stereoselectivity. Beyond classical hydrolytic activity, ABHs also exhibit moonlighting functions non-catalytic roles that include acting as hormone precursors, transporters, or chaperones (Lenfant et al., 2013). The discovery of catalytically inactive homologs, such as neuroligins, which mediate synaptic adhesion, illustrates the functional plasticity of this fold. Evolutionary diversification through gene duplication and domain fusion has allowed ABHs to populate nearly every ecological niche. Environmental adaptations further expand ABH diversity.

Psychrophilic, mesophilic, and thermophilic variants demonstrate how subtle sequence and compositional changes modulate structural dynamics. For instance, psychrophilic α/β hydrolases exhibit increased flexibility in the active site and surface loops, enabling high catalytic efficiency at low temperatures (Vardar-Yel, 2025). Elevated contents of serine and threonine enhance hydrogen bonding, maintaining activity under cold stress. Conversely, thermophilic counterparts stabilize through increased hydrophobic and ionic interactions.

Machine learning-based analyses have recently revealed that amino acid composition patterns, particularly high Ala, Gly, Ser, and Thr, are predictive of psychrophilic adaptation, illustrating how bioinformatics can uncover evolutionary trends (Vardar-Yel, 2025). Database-driven approaches, such as ESTHER and LED, have facilitated the functional classification of ABHs by linking structure to catalysis. Phylogenetic analyses reveal that while ABHs share a conserved catalytic mechanism, their substrate space overlaps, blurring the boundaries between enzyme classes (Bauer et al., 2019). This interconnectedness supports the concept of a continuum of catalytic functions rather than discrete families.

In conclusion, the α/β-hydrolase superfamily exemplifies nature's strategy for functional innovation built upon a robust structural foundation. Their widespread occurrence, mechanistic versatility, and adaptability to extreme conditions make them central to metabolism, biotechnology, and enzyme evolution studies. For example, differences in the substrate-binding site or the oxyanion hole can switch the specificity of hydrolases from an esterase into a lipase or an amidase. The flexibility of this structural unit makes it an ideal candidate for the efficient biological diversification of the α/β-hydrolase structural

framework from one physiological role into another (Lenfant et al., 2013). There was significant functional benefit conferred by α/β-hydrolases on prokaryotes. Their functions included acting as lipases, phospholipases, or dehalogenases. All of this helped them adapt to their environment (Bornscheuer, 2002). In eukaryotes, they play central roles in processes such as neurotransmitter hydrolysis (acetylcholinesterase), hormone regulation (epoxide hydrolases), and protein modification (ubiquitin carboxy-terminal hydrolases) (Lenfant et al., 2013; Ollis et al., 1992). The adaptability of this enzyme group in terms of evolutionary history can also be seen in its existence among extremophiles. The α/β-hydrolases of extremophiles show ideal levels of stability and catalysis in highly adverse conditions of temperature, pH, and salinity. The three-dimensional structure of this enzyme is maintained by strong intra-molecular forces, which help it perform its catalysis function in highly adverse physico-chemical conditions that would denature mesophilic proteins (Littlechild, 2015). Individually, the biological diversity of α/β-hydrolases attests to their success as nature's modular catalysts, capable of conserving the same fold while differentiating into enzymes of survival, adaptation, or biotechnological interest.

However, despite the enormous biochemical and structural diversity of α/β-hydrolase enzymes, their accurate classification remains a persistent challenge in enzymology. The mere growth of genomic and metagenomic databases has resulted in the identification of thousands of potential hydrolases, a large number of which show low sequence similarity even with the occurrence of conserved structural folds and reaction mechanisms (Holmquist, 2000). This gap between sequence and structure has frustrated functional annotation, resulting in systematic misclassification and a lengthening list of enzymes being annotated as "hypothetical proteins" in the public databases (Nardini & Dijkstra, 1999). In addition, the α/β-hydrolase superfamily contains a very wide variety of catalytic activities lipases, esterases, epoxide hydrolases, and amidases that frequently have overlapping substrate specificities. Classical classification procedures relying mainly on sequence homology are thus not adequate to reflect the subtle functional relationships of this enzyme family. Even enzymes with very similar folds can act on completely different substrates, whereas others with low sequence identity can catalyze the same reactions (Ollis & Carr, 2009). These complexities require the development of more sophisticated analytical models that transcend primary sequence comparisons and incorporate structural, biochemical, and environmental parameters. These are especially significant for extremophilic microbial-derived hydrolases,

whose novel physicochemical adaptations to pH, temperature, and salinity increasingly obscure traditional classification boundaries (Gupta et al., 2002).

Classification of the α/β-hydrolase fold enzymes based on sequence similarity alone is intrinsically challenging due to the very low sequence identity among members of the superfamily often below 20% (Nardini & Dijkstra, 1999). Although all of these enzymes possess a conserved α/β-fold and a classical catalytic triad (Ser–His–Asp/Glu), their primary sequences have diverged considerably over time. As a result, alignment-based methods such as BLAST or Clustal frequently fail to detect homologous relations and therefore lead to functional misannotations or incomplete classification in genomic databases (Arpigny & Jaeger, 1999). This is because the α/β-hydrolase fold is structurally plastic: it can accommodate diverse loop insertions, deletions, and secondary structure rearrangements without interfering with the basic catalytic geometry (Ollis & Carr, 2009). The consequence is that enzymes with nearly identical folds catalyze entirely different reactions, while others with low sequence identity have indistinguishable catalytic mechanisms (Holmquist, 2000). This makes sequence homology by itself an inadequate basis for functional classification. The other disadvantage is that clustering based on sequences does not correctly account for convergent evolution a phenomenon where unrelated enzymes evolve similar motifs or active-site residues for performing similar activities. This type of convergence most times leads to false positives in annotation, where enzymes from various evolutionary backgrounds are incorrectly assigned to the same functional class (Bornscheuer, 2002) In addition, auto-annotation pipelines, which rely heavily on homology cut-offs, tend to assign general or vague function descriptions such as "putative esterase" or "lipase-like protein," thereby masking the true diversity of the superfamily. While sequence analysis provides a convenient starting point for enzyme discovery, its limitations emphasize the need for integrative classification systems that fold in structural bioinformatics, biochemical characterization, and machine learning approaches with the capacity to unearth subtle sequence–structure–function relationships missed by conventional alignment tools.

Experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and biochemical assays have been the gold standard for decades to elucidate the structure, function, and catalytic activity of α/β-hydrolase enzymes. These techniques provide atomic-level data on active-site architecture, substrate binding, and conformational

motions vital to the understanding of enzyme behavior (Holmquist, 2000). Nevertheless, while extremely precise, such approaches are inherently limited in terms of price, time investment, and technical complexity particularly when applied to extremophilic enzymes.

X-ray crystallography requires highly pure and crystallizable proteins, yet some extremozymes resist crystallization since they possess flexible surface loops, are of high salt concentration, or are inherently unstable at conditions outside their natural pH and temperature(Littlechild, 2015). Even when crystals are obtained, diffraction quality is generally poor, with labor-intensive optimization and resort to synchrotron radiation necessary in order to establish reasonable resolution. Similarly, NMR spectroscopy, while powerful at characterization of enzyme motion in solution, is restricted by molecular size generally useful only for proteins with molecular weights below 30–40 kDa—and necessitates isotopic labelling and significant amounts of data processing, making it impractical for large-scale enzyme screening (Kay, 2016). In addition, biochemical and kinetic tests to characterize hydrolase activity are substrate-dependent and optimal reaction conditions, and the latter may be distinct from the enzyme's native conditions. Most acidophilic and alkaliphilic hydrolases have activity at extreme pH or ionic environments, which add more challenges for protein stability during purification and quantification Therefore, the yield of such experimental methods is low compared with the rate at which novel hydrolase sequences are discovered using genomic and metagenomic studies.

The interplay of these limitations environmental instability, cost, and time required has represented a serious bottleneck to the comprehensive classification of α/β-hydrolases. Consequently, researchers rely more and more on computational and data-intensive strategies to complement experimental data. Machine learning, homology modelling, and molecular dynamics simulations now present scalable options with the capability to predict enzyme pH adaptability, structure, and function more quickly and with larger scope (Yang et al., 2024).

2.4 MACHINE LEARNING IN ENZYME CLASSIFICATION

The two decades of runaway expansion of genomic and proteomic information have delivered both unparalleled opportunity and unparalleled analytic challenge to enzymology. Classical tools of classification based on sequence similarity, structural homology, and

biochemical tests have been uniquely valuable for shedding light on enzyme mechanisms but still fail when confronted with the enormity and heterogeneity of modern biological data sets. In this regard, machine learning (ML) has emerged as a paradigm breaker that offers the ability to retrieve latent patterns in high-dimensional biological data and deduce functional relationships out of scope of common practices (Libbrecht & Noble, 2015). Machine learning algorithms have proven particularly effective when applied to complex biological systems, as they can handle nonlinear relationships and extract informative features that traditional statistical approaches often overlook (Tao et al., 2020; Pandey et al., 2019). In enzyme studies, algorithms such as Support Vector Machines (SVM), Random Forests, and Artificial Neural Networks (ANN) have demonstrated superior accuracy when amino acid composition, dipeptide frequency, and hydrophobicity profiles are used as input descriptors (Akinsola et al., 2017). Machine learning approaches are particularly well suited to enzyme classification because they have the ability to integrate a variety of sources of information e.g., amino acid sequences, secondary and tertiary structures, and physicochemical properties to build predictive models that can generalize among enzyme families. These models learn discriminative features from data directly, so they are able to distinguish between highly similar proteins, predict catalytic activities, and even predict new enzyme subclasses (Chandra et al., 2023). Recent work has shown that incorporating physicochemical parameters such as hydrophobicity, residue charge, and amino acid distribution enhances the robustness of enzyme classification models, particularly in differentiating thermophilic and mesophilic proteins (Gromiha & Suresh, 2007). Moreover, ensemble-based learning methods that aggregate multiple classifiers such as Gradient Boosting and AdaBoost further improve prediction accuracy by minimizing overfitting and bias (Pandey et al., 2019). The recent advances in deep learning architectures, including convolutional neural networks (CNNs), graph neural networks (GNNs), and transformer-based protein language models, have further revolutionized computational enzymology. These methods can pick up non-linear sequence–structure–function relationships that are overlooked by traditional alignment-based methods, allowing improved accuracy and interpretability of annotations. For example, convolutional and recurrent neural network architectures have achieved over 95% classification accuracy in predicting EC numbers and enzyme subclasses by automatically learning from raw sequence data without explicit feature engineering (The Classification of Enzymes by Deep Learning, 2021). These

advancements indicate that hybrid deep learning systems can capture hierarchical biochemical information, facilitating high-confidence annotation of enzymes with low sequence identity. Consequently, ML-based frameworks today are the key to the prediction of enzyme function, structural prediction, and classification of enzyme superfamilies with low sequence identity like α/β-hydrolases (Yang et al., 2024).

2.4.1 Overview of Machine Learning Techniques

Machine learning (ML) refers to a broad category of computational methods by which systems can learn patterns and predict on the basis of empirical data without explicit rule-based programming. In enzyme classification, ML offers an adaptable and scalable solution compared to manual curation or sequence comparison and is particularly suited to handle large-scale biological data generated by genomic and metagenomic surveys (Lavecchia, 2015). Recent reviews highlight that machine learning models used for enzyme prediction can be broadly grouped into supervised, unsupervised, and deep learning paradigms, each suited for different levels of annotation and data availability (Pandey et al., 2019). Overall, ML techniques can be categorized as supervised, unsupervised, and deep learning techniques, differing in methodologies and bioinformatics utilities.

In supervised learning, algorithms are learned from labelled datasets where enzyme sequences or structures are tagged with known functional classes to learn discriminative patterns that can predict the class of novel, unseen samples. Classical supervised models such as support vector machines (SVMs), random forests (RFs), and k-nearest neighbors (kNN) have been extensively employed to predict enzyme function based on sequence-derived descriptors such as amino acid composition, dipeptide frequency, and physicochemical indices(Chandra et al., 2023). Studies comparing these algorithms have shown that SVM and RF consistently outperform kNN and Naïve Bayes models for enzyme activity prediction, achieving classification accuracies between 85% and 92% (Akinsola et al., 2017; Pandey et al., 2019). On the other hand, unsupervised learning methods like clustering algorithms k-means, hierarchical clustering, and self-organizing maps aim to find hidden patterns or groupings within unlabelled sets of enzymes. These types of methods are informative in finding hidden relationships or evolutionary subgroups among enzyme families, for instance, to find new subtypes of hydrolases from metagenomic data or clustering enzymes by reaction type by structural similarity(Lavecchia, 2015; Rives et al.,

2021). The most revolutionary developments in recent years have been the result of deep learning, which utilizes multi-layered neural networks with the ability to learn automatically hierarchical features from unstructured biological information. Protein sequence classification has also been tackled effectively with architectures such as convolutional neural networks (CNNs), while structure prediction and functional annotation have been revolutionized by recurrent neural networks (RNNs) and transformer-based protein language models (such as AlphaFold, ESM, ProtT5), which can learn long-range dependencies in amino acid sequences (Rives et al., 2021). Moreover, graph neural networks (GNNs) represent protein structures in the form of node–edge graphs, which allow direct learning from spatial and topological features and thereby improve catalytic site prediction and substrate specificity modelling (Gligorijević et al., 2021). According to Pandey et al. (2019), combining deep neural architectures with ensemble learning yields the best generalization performance for enzyme-related tasks, as hybrid models can leverage both global and local sequence–structure representations. Together, these approaches collectively enable a paradigm shift away from rule-based annotation toward data-driven discovery, in which enzymes can be classified, clustered, and functionally annotated based on the composite interpretation of large-scale biochemical, structural, and environmental data sets. This multi-faceted perspective provides a solid foundation for the classification of α/β-hydrolases, which are generally found to have low sequence homology but structurally and catalytically conserved motifs.

2.4.2 Previous Applications of ML in Enzymology

Machine learning enzymology has grown remarkably over the past decade, with the paradigm shifting from sequence alignment–based conventional approaches to data-driven predictive modelling. The early computational approaches were limited to enzyme function prediction, where enzymes were categorized according to their Enzyme Commission (EC) numbers based on sequence-derived features by algorithms. One of the initial tools in this group, EFICAz, used a combination of sequence similarity, conserved motifs, and machine learning classifiers to predict enzyme functions more accurately than BLAST-based annotation (Arakaki et al., 2009). With the increasing availability of genomic data, new architectures were suggested to scale up and improve accuracy. The DeepEC model, for instance, employed a deep neural network framework trained on millions of protein

sequences to successfully predict full four-level EC numbers with high accuracy, outperforming traditional sequence-alignment techniques in both accuracy and computational efficiency (Lavecchia, 2015). Similarly, DEEPre fused convolutional neural networks (CNNs) and multi-task learning for the prediction of both EC numbers and enzyme subfamily membership with good generalization even for low sequence identity proteins. In addition to EC classification, machine learning approaches have been used in enzyme–substrate specificity prediction, catalytic residue prediction, and reaction mechanism prediction. For example, EnzymeMiner, a webserver integrating random forest classifiers with structural and functional descriptors, was capable of discovering novel hydrolases with potential industrial applications in metagenomic databases (Hon et al., 2020). Similarly, Gligorijević et al., (2021) demonstrated th.at graph convolutional networks (GNNs) can learn from three-dimensional protein contact maps, enabling direct structure-based function prediction without the need for explicit alignment or hand-crafted feature extraction. Other studies confirm that CNN–LSTM hybrid architectures trained on large enzyme datasets achieved accuracy levels above 95% for EC number prediction and functional annotation (The Classification of Enzymes by Deep Learning, 2021). Moreover, models that integrate amino acid composition, structural topology, and physicochemical descriptors have successfully classified α/β-hydrolases, esterases, and dehalogenases, providing evidence that ML can generalize across enzyme superfamilies (Classification of Enzymes Using Machine Learning, 2020). Feature selection methods such as PCA and RFE were also reported to improve interpretability and performance by reducing noise in biochemical data (Pandey et al., 2019). More recent advances in transformer-based protein language models such as ESM-1b and ProtT5 have pushed computational enzymology to a new level with unsupervised learning from billions of amino acid sequences. These models capture subtle contextual relationships in protein sequences that enable function prediction with high accuracy even for enzymes with limited current annotation (Rives et al., 2021). These models have been especially helpful for α/β-hydrolases, where low sequence similarity and high structural diversity frustrate traditional classification schemes. Collectively, these endeavors underscore the transformative potential of ML and DL in enzymology. By coupling large-scale biological datasets with cutting-edge learning architectures, researchers have created computational frameworks capable of expediting enzyme discovery, improving

annotation accuracy, and elucidating new catalytic mechanisms successfully bridging the divide between experimental biochemistry and predictive bioinformatics.

2.5 RATIONALE AND OBJECTIVES OF THE STUDY

2.5.1 Gaps in Current Classification Methods

Despite considerable progress in the classification of α/β hydrolases, several critical challenges remain unresolved. Sequence-based classification, though widely applied, suffers from inherent limitations due to the very low sequence identity across members of this superfamily. Many hydrolases that share the conserved α/β fold display less than 20% sequence similarity, making reliable annotation based solely on homology-based approaches highly problematic (Nardini & Dijkstra, 1999). Consequently, a substantial number of enzymes are incorrectly assigned to broad categories such as esterases or lipases, resulting in widespread functional misannotations in genomic repositories (Solanki & Lee, 2010). Experimental approaches including biochemical assays, X-ray crystallography, and nuclear magnetic resonance (NMR) spectroscopy provide high-resolution insights into enzyme structure and function. However, they lack scalability in the face of the thousands of novel sequences generated by high-throughput genomic projects. These methods are also resource- and time-intensive, requiring purified protein samples and specialized laboratory facilities, which constrains their applicability for large-scale classification efforts. Furthermore, extremophilic enzymes such as acidic and alkaline hydrolases pose additional difficulties, as they often lose stability outside their native ecological conditions(Littlechild, 2015). Machine learning (ML) methods have emerged as promising alternatives, yet they are not without shortcomings. Early ML models primarily utilized simple sequence-derived features (e.g., amino acid composition, k-mer frequencies) without considering structural or environmental information, thereby limiting their predictive capacity. Although deep learning and graph neural networks (GNNs) have significantly improved predictive accuracy, persistent challenges remain in terms of interpretability and in dealing with imbalanced datasets, particularly where certain classes of hydrolases (e.g., alkaline hydrolases) are underrepresented (Yang et al., 2024). Moreover, the lack of integrated frameworks that unify sequence, structural, and ecological features continues to restrict the precision with which acidic and alkaline hydrolases can be distinguished.

In summary, these unresolved gaps highlight the urgent need for a comprehensive, data-driven framework that integrates heterogeneous biological features, incorporates robust validation strategies, and balances scalability with accuracy. Addressing these issues forms the central motivation for the present study, which seeks to advance enzyme classification methodologies beyond the constraints of current experimental and computational approaches.

2.5.2 Study Goals and Expected Contributions

Building on the identified gaps in current classification methods, this study is designed to develop and evaluate a data-driven machine learning framework for the discrimination of acidic and alkaline Alpha/Beta hydrolases. The primary goal is to integrate sequence-derived, structural, and physicochemical features into predictive models capable of achieving both high accuracy and scalability. By leveraging modern supervised and deep learning techniques, the study aims to provide a more reliable approach to enzyme classification than existing sequence-based or experimental strategies.

Specific Goals:

Feature Integration: To construct a comprehensive dataset that combines amino acid composition, evolutionary profiles, predicted structural motifs, and environmental pH-related adaptations.

Model Development: To design and implement supervised and deep learning models (e.g., random forests, convolutional neural networks, and transformers) optimized for enzyme classification.

Performance Evaluation: To rigorously assess the models using multiple performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, with special attention to class imbalance between acidic and alkaline hydrolases.

Comparative Analysis: To compare the proposed machine learning framework against traditional sequence-based and experimental benchmarks, highlighting strengths and limitations.

Application Scenarios: To explore potential applications of the developed models in industrial and environmental contexts, such as predicting candidate enzymes for detergents, food biotechnology, and bioremediation.

Scientific Contribution: The study will advance the field of computational enzymology by providing a novel machine learning framework that integrates diverse biological features, offering improved accuracy in distinguishing between acidic and alkaline Alpha/Beta hydrolases.

Methodological Contribution: It will demonstrate the value of combining multiple feature types—sequence, structure, and environmental adaptations in predictive models, serving as a methodological template for future enzyme classification studies.

Practical Contribution: The framework is expected to facilitate the rapid discovery and annotation of extremophilic hydrolases, accelerating their application in industrial and environmental biotechnology where robust enzymes are required under extreme pH conditions.

Broader Impact: Beyond Alpha/Beta hydrolases, the approach developed here can be extended to other enzyme families, contributing to the global effort of functional annotation in the post-genomic era.

3. METHODOLOGY

3.1 THE DATASET

The data for the present analysis were obtained from the Universal Protein Knowledgebase (UniProt), which served as the main reference for identifying α/β hydrolase enzymes from microorganisms adapted to both acidic and alkaline environments. A total of 200 alkaline and 203 acidic or acid-tolerant bacterial hydrolases were gathered from the database. Each sequence underwent homology modeling, and the resulting proteins were typically between 201 and 400 amino acids long. This size range aligns with what is generally reported for α/β hydrolases, enzymes known for maintaining a balance between structural stability and catalytic flexibility. To minimize redundancy and remove inaccurate entries, the CD-HIT algorithm was used, as it is a standard tool for cleaning and clustering protein datasets in bioinformatics.

The refined collection included several types of α/β hydrolase enzymes such as esterase, carboxylesterase, thioesterase, S-formylglutathione hydrolase, proline iminopeptidase, dienelactone hydrolase, pimeloyl-[acyl-carrier protein] methyl ester esterase, palmitoyl-protein thioesterase, carboxymethylenebutenolidase, peptidase S9 prolyl oligopeptidase, 3-oxoadipate enol-lactonase, biotin synthesis protein BioH, acetylxylan esterase, prolyl oligopeptidase, epoxide hydrolase, and PHB depolymerase. These enzymes cover a wide functional spectrum, ranging from ester hydrolysis and lactone ring opening to protein modification. The variation in catalytic behavior highlights how the α/β hydrolase fold has evolved to perform different reactions while maintaining a conserved structural framework.

Taxonomic profiling showed that the identified sequences originated from several acidophilic and halophilic bacterial species. Representative organisms included *Acidiferrobacter sp.*, *Acidiferrobacter thiooxydans*, *Acidihalobacter prosperus*, *Acidiphilium rubrum*, *Acidisoma cellulosilyticum*, *Acidisoma sylvae*, *Acidithiobacillus caldus*, *Acidithiobacillus ferrianus*, *Acidithiobacillus marinus*, and *Acidithiobacillus sulfuriphilus*. The presence of multiple *Acidithiobacillus* and *Acidiferrobacter* species agrees with their established role in acidic, metal-rich ecological niches. Detailed enzyme lists, including UniProt accession numbers and source organisms, are presented in Tables 3.1-3.2.

To further describe these enzymes, several physicochemical and structural parameters were calculated. The analysis included amino acid composition, predicted secondary structure, theoretical isoelectric point (pI), instability index, GRAVY (Grand Average of Hydropathicity) value, and aliphatic index. These properties provide insight into enzyme stability and solubility under different environmental conditions. Data were obtained using reliable online tools such as Expasy-ProtParam, COPid (Calculate Composition of Whole Protein), and the CAMPR3 feature calculator (Kumar et al., 2008; Ku et al., 2009). Overall, this integrated computational approach enabled a detailed overview of the structural diversity within the α/β hydrolase family, serving as a foundation for subsequent comparative and functional analyses.

Table 3.1: Acidic α/β Hydrolases Screened from Different Bacterial Sources.

Organism	Enzyme name	Uniprot ID
Acidiferrobacter sp.	Alpha/beta hydrolase	A0A2Z3QX72 A0A2Z3QYI8 A0A2Z3R0V3 A0A2D7VJ80 A0A2D7VKQ3 A0A2D7VKW2 A0A2D7VLY1 A0A2D7VM02 A0A2D7VMW0 A0A2D7VND7 A0A2D7VNF3 A0A2E8QNH4 A0A2E8QP06 A0A2E8QP75 A0A2E8QPD7 A0A2E8QQT6 A0A2E8QR17 A0A2E8QSF6
Acidiferrobacter sp.	Hydrolase	A0A2Z3QXP7
Acidiferrobacter sp.	Carboxylesterase	A0A2Z3QZR1 A0A2E8QSQ0
Acidiferrobacter sp.	Pimeloyl-[acyl-carrier protein] methyl ester esterase	A0A2Z3R4A6
Acidiferrobacter sp	Carboxymethylenebutenolidase	A0A2Z3R4K1
Acidiferrobacter sp.	S-formylglutathione hydrolase	A0A2Z3R9A1
Acidiferrobacter sp.	Dienelactone hydrolase	A0A2Z3QUX3
Acidiferrobacter thiooxydans	Hydrolase	A0A1C2FXT7
Acidiferrobacter thiooxydans	Alpha/beta hydrolase	A0A1C2G0X1 A0A1C2G1D9
Acidiferrobacter thiooxydans	Carboxylesterase	A0A1C2G2E2
Acidiferrobacter thiooxydans	Proline iminopeptidase	A0A1C2G3R6

Table 3.1: Acidic α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Acidiferrobacter thiooxydans	S-formylglutathione hydrolase	A0A1C2G4L6
Acidiferrobacter thiooxydans	Dienelactone hydrolase family protein	A0A368HBP7 A0A368HJP8
Acidiferrobacter thiooxydans	Thioesterase	A0A368HGM3
Acidiferrobacter thiooxydans	Pimeloyl-[acyl-carrier protein] methyl ester esterase	A0A368HGP1
Acidihalobacter prosperus	Proline iminopeptidase	A0A1A6C121
Acidihalobacter prosperus	Carboxymethylenebutenolidase	A0A1A6C1J4
Acidihalobacter prosperus	S-formylglutathione hydrolase	A0A1A6C1N8
Acidihalobacter prosperus	Lipase	A0A1A6C380
Acidihalobacter prosperus	Peptidase S9 prolyl oligopeptidase catalytic domain-containing protein	A0A1A6C445
Acidihalobacter prosperus	Carboxymethylenebutenolidase	A0A1A6C7I0
Acidihalobacter prosperus	Alpha/beta hydrolase	A0A1A6C8W5
Acidisoma cellulosilyticum	Acetylxylan esterase	A0A963YWY9
Acidisoma cellulosilyticum	Alpha/beta hydrolase	A0A963YY11 A0A963YY14 A0A963YYF1 A0A963YYL1 A0A963YZ93 A0A963YZP8 A0A963Z134 A0A963Z1M9 A0A963Z1P4 A0A963Z269 A0A963Z362 A0A963Z3S1 A0A963Z3Y6 A0A963Z433 A0A963Z4G1 A0A963Z4I4 A0A963Z4U1 A0A963Z502 A0A963Z522 A0A963Z611 A0A963Z7M5 A0A964E235 A0A964E297 A0A964E576 A0A964E5N2 A0A964E5Y4 A0A964E685 A0A964E690 A0A964E6F4 A0A963Z6P1 A0A964E7B0 A0A963YXJ5 A0A963YYG2 A0A963YZ56 A0A963Z0P0 A0A963Z240

Table 3.1: Acidic α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Acidisoma cellulosilyticum	Alpha/beta hydrolase	A0A963Z2R6 A0A963Z3S3 A0A963Z591 A0A963Z5U8 A0A963Z748 A0A963YZ75
Acidisoma cellulosilyticum	Esterase family protein	A0A963YYY0
Acidisoma cellulosilyticum	Prolyl oligopeptidase family serine peptidase	A0A963Z057
Acidisoma cellulosilyticum	Dienelactone hydrolase family protein	A0A963Z0N4 A0A963Z1A3 A0A964E6C2
Acidisoma cellulosilyticum	S-formylglutathione hydrolase	A0A963Z280
Acidisoma cellulosilyticum	Carboxymethylenebutenolidase	A0A963Z4X2 A0A964E1W3
Acidisoma cellulosilyticum	Proline iminopeptidase	A0A964E635 A0A964E2F7
Acidisoma silvae	Alpha/beta hydrolase	A0A963YN31 A0A963YNE6 A0A963YP29 A0A963YQS0 A0A963YSA1 A0A963YT45 A0A963YT72 A0A963YTC1 A0A963YTD2 A0A963YV36 A0A963YVK6 A0A963YVU1 A0A963YWG6 A0A963YWC6 A0A963YWV3 A0A963YWZ8 A0A964DYQ7 A0A964DZD5 A0A964E0N6 A0A964E0W0 A0A964E0W9 A0A964E178 A0A964E1C8 A0A963YNX7 A0A963YP67 A0A963YPZ9 A0A963YQT0 A0A963YVE1 A0A963YVH3 A0A964DZP1 A0A963YUJ3 A0A963YSN9
Acidisoma silvae	Proline iminopeptidase	A0A963YNT5 A0A963YNF7
Acidisoma silvae	Esterase family protein	A0A963YP63
Acidisoma silvae	Carboxymethylenebutenolidase	A0A963YPA5
Acidisoma silvae	Phosphodiesterase	A0A963YPJ9 A0A963YRJ4

Table 3.1: Acidic α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Acidisoma silvae	S-formylglutathione hydrolase	A0A963YPQ7
Acidisoma silvae	Acetylxylan esterase	A0A963YPU2 A0A963YR69
Acidisoma silvae	Dienelactone hydrolase family protein	A0A963YQ22 A0A964DX54
Acidisoma silvae	PHB depolymerase family esterase	A0A963YTV5
Acidisoma silvae	Dienelactone hydrolase family protein	A0A963YW29 A0A964DZ05 A0A964E076
Acidisoma silvae	Prolyl oligopeptidase family serine peptidase	A0A964DX33
Acidithiobacillus caldus	Alpha/beta hydrolase	A0A059ZLN3 A0A059ZPU3
Acidithiobacillus caldus	Biotin synthesis protein BioH	A0A060A2J5
Acidithiobacillus ferrianus	Palmitoyl-protein thioesterase	A0A845U2C6
Acidithiobacillus ferrianus	Alpha/beta fold hydrolase	A0A845U2M1 A0A845U4U5 A0A845U917 A0A845UL72 A0A845U7I1 A0A845UFN4
Acidithiobacillus ferrianus	Phospholipase	A0A845U8X6
Acidithiobacillus ferrianus	Dienelactone hydrolase family protein	A0A845UF23 A0A845ULX4
Acidithiobacillus ferrianus	S-formylglutathione hydrolase	A0A845UMC1
Acidithiobacillus marinus	Carboxymethylenebutenolidase	A0A2I1DK85
Acidithiobacillus marinus	Palmitoyl-protein thioesterase	A0A2I1DKD1
Acidithiobacillus marinus	Alpha/beta hydrolase	A0A2I1DMD4
Acidithiobacillus marinus	Phospholipase	A0A2I1DMN0 A0A2I1DNI6
Acidithiobacillus marinus	Esterase	A0A2I1DP66
Acidithiobacillus marinus	Pimeloyl-[acyl-carrier protein] methyl ester esterase	A0A2I1DPZ6
Acidithiobacillus marinus	S-formylglutathione hydrolase	A0A2I1DQW4
Acidithiobacillus sulfuriphilus	Alpha/beta fold hydrolase	A0A3M8QQ08
Acidithiobacillus sulfuriphilus	Dienelactone hydrolase family protein	A0A3M8QSY7
Acidithiobacillus sulfuriphilus	S-formylglutathione hydrolase	A0A3M8QXY3
Alicyclobacillaceae bacterium	Alpha/beta hydrolase	A0A418MA17 A0A418MDG8 A0A418MG93 A0A418MGE5 A0A418MH38
Alicyclobacillaceae bacterium	Proline iminopeptidase	A0A418MA48 A0A418MLG5
Alicyclobacillaceae bacterium	Esterase family protein	A0A418MCW1
Alicyclobacillaceae bacterium	Histidine biosynthesis bifunctional protein HisIE	A0A418MH28
Alicyclobacillus mengziensis	Serine aminopeptidase S33 domain-containing protein	A0A9X7VUU4

Table 3.1: Acidic α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Alicyclobacillus mengziensis	Proline iminopeptidase	A0A9X7VWQ3 A0A9X7Z9M1
Alicyclobacillus mengziensis	Alpha/beta hydrolase	A0A9X7VX17 A0A9X7W0T5 A0A9X7W230 A0A9X7W2T A0A9X7W341 A0A9X7W4C4 A0A9X7Z5N3 A0A9X7Z6M2 A0A9X7Z6M3 A0A9X7Z722 A0A9X7Z7C5 A0A9X7Z7I2 A0A0D8FY94
Alicyclobacillus mengziensis	Dienelactone hydrolase	A0A9X7VZX3
Alicyclobacillus mengziensis	Lipase family protein	A0A9X7W2B3
Alicyclobacillus mengziensis	Esterase family protein	A0A9X7W2U7
Ferrimicrobium acidiphilum	Carboxymethylenebutenolidase	A0A0D8FRK6 A0A0D8FWP9 A0A0D8FWP9
Ferrimicrobium acidiphilum	Alpha/beta hydrolase	A0A0D8FS68 A0A0D8FY94 A0A0D8FTG0 A0A0D8FU54
Ferrimicrobium acidiphilum	Proline iminopeptidase	A0A0D8FUQ1
Ferrimicrobium acidiphilum	Feruloyl esterase	A0A0D8FSW6
Ferrimicrobium acidiphilum	Carboxylic ester hydrolase	A0A0D8FW75

Table 3.2: Alkaline α/β-hydrolases Screened From Different Bacterial Sources.

Organism	Enzyme name	Uniprot ID
Alkalicoccus luteus	Alpha/beta hydrolase	A0A969PQN5 A0A969PRA3 A0A969PRU0 A0A969PRY3 A0A969PSA1 A0A969PTJ9 A0A969PV57 A0A969TUF8 A0A969TUM9 A0A969TV94 A0A969TV99 A0A969TVF7 A0A969TW32 A0A969TWJ7 A0A969TWP9 A0A969PRK9 A0A969PS19 A0A969TUA9 A0A969TWB4 A0A969TWG9 A0A969TYC8
Alkalicoccus luteus	Hydrolase-1 domain-containing protein	A0A969PTL1
Alkalicoccus luteus	Acetylxylan esterase	A0A969TUL3
Alkalihalobacillus alcalophilus	Histidine biosynthesis bifunctional protein HisIE	A0A094XFI9
Alkalihalobacillus alcalophilus	Alpha/beta hydrolase	A0A094WEE3 A0A094WSD7 A0A094XB56 A0A094YUE3 A0A094YYX0
Alkalihalobacillus alcalophilus	Beta-ketoadipate enol-lactone hydrolase	A0A094WCN2
Alkalihalobacillus alcalophilus	Hydrolase/Esterase/Carboxylesterase/Lipase	A0A094WHK9 A0A094WN42 A0A094WIG6 A0A094XFL6 A0A094WMT5 A0A094XF26
Alkalihalobacillus alcalophilus	Enterochelin esterase	A0A094WLL5
Alkalihalobacillus alcalophilus	Dienelactone hydrolase domain-containing protein	A0A094WN90
Alkalihalobacillus alcalophilus	Phospholipase	A0A094WRY2
Alkalihalobacillus alcalophilus	Acetyl esterase	A0A094XCL7 A0A094YWB7
Alkalihalobacillus alcalophilus	Proline iminopeptidase	A0A094XD17
Pseudomonas toyotomiensis	Alpha/beta hydrolase	A0A1I5M9L1 A0A1I5MVG3 A0A1I5N2C9

Table 3.2: Alkaline α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Pseudomonas toyotomiensis	Alpha/beta hydrolase	A0A1I5PWL7 A0A1I5R5G3 A0A1I5RDV3 A0A1I5RGN3 A0A1I5S2U7 A0A1I5S430 A0A1I5SJ53 A0A1I5SN36 A0A1I5T9I6 A0A1I5TEX9 A0A1I5TST1 A0A1I5UML1 A0A1I5V6E4 A0A1I5VL79 A0A1I5VX50 A0A1I5W614 A0A1I5WBL3 A0A1I5WDG7 A0A1I5WQ50 A0A1I5XLK8 A0A1I5YA42 A0A1I5YST3 A0A1I5YXU0 A0A2G5FFB2 A0A2G5FKN3 A0A2G5FQP5 A0A2G5FYX2 A0A1I5ML26 A0A1I5NEU6 A0A1I5S348 A0A1I5RIV9
Pseudomonas toyotomiensis	3-oxoadipate enol-lactonase	A0A1I5MRF4 A0A1I5YBS8
Pseudomonas toyotomiensis	Esterase/Carboxylesterase/Hydrolase/Acetylhydrolase	A0A1I5PQS9 A0A2G5FGC0 A0A1I5X7G1 A0A1I5WWC8 A0A2G5FXE2
Pseudomonas toyotomiensis	Acetyl esterase/lipase	A0A1I5RJW3 A0A1I5RLW2 A0A1I5UE95 A0A1I5NXZ1 A0A1I5SQB6
Pseudomonas toyotomiensis	Dienelactone hydrolase	A0A1I5S1B5
Pseudomonas toyotomiensis	Medium-chain acyl-[acyl-carrier-protein] hydrolase	A0A1I5UM98
Pseudomonas toyotomiensis	Proline iminopeptidase	A0A1I5PH39
Pseudomonas toyotomiensis	Pimeloyl-[acyl-carrier protein] methyl ester esterase	A0A1I5WS52

Table 3.2: Alkaline α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Pseudomonas toyotomiensis	S-formylglutathione hydrolase	A0A1I5XNF9
Pseudomonas mandelii	LipS	H2ETP9
Proteus sp	Alkaline lipase	F2Y9R8
Alkalihalophilus marmarensis	Hydrolase-1 domain-containing protein	U6SK75 U6SPU2
Alkalihalophilus marmarensis	Serine aminopeptidase S33 domain-containing protein	U6SI45 U6SJA4 U6SQN4
Alkalihalophilus marmarensis	Phospholipase	U6SLC8
Alkalihalophilus marmarensis	Carboxylesterase/Esterase/Hydrolase	U6SN19 U6SNU6 U6SPV4 U6SM81 U6SM54 U6SPW5
Alkalihalophilus marmarensis	Lipase	U6SP33 U6SRJ6
Alkalihalophilus marmarensis	Alpha/beta hydrolase	U6SQ64 U6SR41 U6SS49 U6SSQ3 U6SU10 U6SU37 U6SQH1
Alkalihalophilus marmarensis	Peptidase S9 prolyl oligopeptidase catalytic domain-containing protein	U6SV57
Alkalihalophilus marmarensis	Enterochelin esterase	U6SW26
Alkalihalophilus marmarensis	Histidine biosynthesis bifunctional protein HisIE	U6SK43
Dietzia timorensis	AB hydrolase-1 domain-containing protein	A0A173LH07 A0A173LMI5 A0A173LMS6 A0A173LRQ4
Lysobacter alkalisoli	Alpha/beta hydrolase	A0A514BMU0 A0A514BN25 A0A514BN47 A0A514BND8 A0A514BPS9 A0A514BQR4 A0A514BQU2 A0A514BSE5 A0A514BVL6 A0A514BWP5 A0A514BWQ3 A0A514BX10
Lysobacter alkalisoli	Lysophospholipase	A0A514BN74

Table 3.2: Alkaline α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Lysobacter alkalisoli	Lysophospholipase	A0A514BRU2 A0A514BUR1
Lysobacter alkalisoli	S-formylglutathione hydrolase	A0A514BNM1
Lysobacter alkalisoli	Dienelactone hydrolase	A0A514BPT3 A0A514BSB3
Lysobacter alkalisoli	Carboxymethylenebutenolidase	A0A514BQY9
Lysobacter alkalisoli	Proline iminopeptidase	A0A514BTL8
Lysobacter alkalisoli	Carboxylesterase/Esterase	A0A514BV49 A0A514BWX7
Planococcus salinus	Alpha/beta hydrolase	A0A3M8P2X0 A0A3M8P306 A0A3M8P410 A0A3M8P4S4 A0A3M8P5P1 A0A3M8P5R7 A0A3M8P6M2 A0A3M8P7C2 A0A3M8P835 A0A3M8PAH5 A0A3M8PBE4 A0A3M8PBJ4 A0A3M8PC50 A0A3M8PET2 A0A926Q1C7 A0A926Q2F9 A0A926Q4U0 A0A926Q5F4 A0A926Q5J5
Planococcus salinus	Esterase	A0A3M8P331 A0A3M8P4E0 A0A926Q431
Planococcus salinus	Proline iminopeptidase	A0A3M8P6C7
Sinomicrobium weinanense	Esterase	A0A926JNJ1 A0A926Q0H8 0A926Q3P7
Sinomicrobium weinanense	Alpha/beta hydrolase	A0A926JNP1 A0A926JP64 A0A926JQ49 A0A926JQY3 A0A926JR67 A0A926JR69 A0A926JRL8 A0A926JRN0 A0A926JRZ4 A0A926JSI5 A0A926JUU6 A0A926JV33 A0A926JV66 A0A926JVB2 A0A926Q3B2 A0A926Q3G4

Table 3.2: Alkaline α/β-hydrolases Screened From Different Bacterial Sources "Table Continued".

Organism	Enzyme name	Uniprot ID
Sinomicrobium weinanense	Dienelactone hydrolase family protein	A0A926JR19 A0A926JTD3 A0A926JW00 A0A926Q2P0
Sinomicrobium weinanense	Prolyl oligopeptidase family serine peptidase	A0A926JTT2 A0A926Q3D3
Sinomicrobium weinanense	Thioesterase	A0A926Q2S8

3.2 MACHINE LEARNING ANALYSIS

3.2.1 Cross-Validation

The overall reliability of any predictive model depends not only on how well it fits the data but also on how consistently it performs with unseen samples. One of the most practical ways to test this is by using *k*-fold cross-validation, a method that has become almost standard in model evaluation. In this technique, the dataset is divided into *k* roughly equal parts. The model is trained on *k–1* of these parts and tested on the remaining one. This process is repeated several times so that every subset is used once as a test set (Bates et al., 2023). The model's true stability or dependability can be reasonably estimated from the average of the obtained error values.

For this particular study, a 5-fold cross-validation was selected. This number was chosen because the dataset was of medium size, and $k = 5$ usually provides a good balance between computational cost and accuracy. Using this setup also reduced the random bias that can occur when the data are split only once. For comparison, a simple train–test split was also carried out, with 70% of the data assigned for training and the remaining 30% for testing. Comparing these two validation approaches made it easier to see whether cross-validation provided a more dependable measure of performance.

3.2.2 Comparison of Different Algorithms

In selecting the machine learning algorithms for this study, we focused on practical considerations related to the dataset and the research objectives. The dataset was relatively small, so we prioritized methods that have shown reliable performance with limited data and

that do not require extensive parameter tuning. Class imbalance and the presence of occasional outliers were also important factors in our decision-making process, as these issues could affect model stability. All analyses and implementations were conducted in Python. Data management and preprocessing were performed using NumPy and Pandas, visualizations were created with Seaborn, and model training and evaluation were carried out with Scikit-learn. This approach allowed us to maintain a reproducible workflow while keeping the process manageable and aligned with the study goals.

3.2.2.1 Random forest (RF)

Random Forest (RF) was employed as a primary classification method in this study due to its well-documented ability to handle complex datasets. The algorithm constructs a large ensemble of decision trees, with each tree independently evaluating the input features and casting a vote for a predicted class. The final classification is then determined by aggregating these individual votes, typically through a majority rule, which helps smooth out the influence of any single noisy or atypical tree. RF has been widely recognized for providing high predictive accuracy, demonstrating robustness in the presence of noisy data or outliers, and reducing the likelihood of overfitting, making it particularly suitable for datasets where both variability and potential anomalies may exist (Prinzie & Van den Poel, 2008; Liu et al., 2012). Moreover, its flexibility in handling both categorical and continuous variables allows it to adapt effectively to a variety of feature types, further enhancing its utility in real-world applications.

3.2.2.2 Decision trees (DTs)

Decision Trees (DTs) were also included in this study as a core method for both classification and regression tasks, primarily due to their intuitive structure and interpretability. These models allow the underlying decision-making process to remain transparent, which is particularly valuable when explaining predictions or integrating the trees with other computational systems. The construction of a DT begins with the dataset in its most general form, and through successive splitting based on feature values and specific criteria, the model gradually specializes to capture patterns and relationships present in the data. This stepwise process, influenced by initial parameter settings and the chosen methodology, enables DTs to adapt effectively to diverse datasets while maintaining clarity in their

decision pathways (Kotsiantis, 2013; Costa & Pedreira, 2023). Additionally, their modular nature facilitates integration with ensemble methods, further extending their practical utility in more complex predictive tasks.

3.2.2.3 Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) was selected for this study as a powerful and highly scalable method based on gradient-boosted decision trees (GBDT). This algorithm has demonstrated exceptional performance across a wide range of predictive tasks, including classification, regression, and ranking, making it a versatile tool for handling complex datasets (Deka & Weiner, 2024). One of the key advantages of XGBoost is its ability to perform parallel computations, which allows for faster model training and optimization, particularly when dealing with large or performance-intensive datasets. Additionally, the algorithm incorporates regularization techniques and flexible objective functions, which contribute to its robustness against overfitting and enhance its generalization capabilities, further increasing its suitability for real-world applications where both accuracy and efficiency are critical.

3.2.2.4 Support vector machines (SVMs)

Support Vector Machines (SVMs), first developed by Vapnik and colleagues, were employed in this study as a core method for tackling classification tasks due to their solid theoretical foundation in statistical learning. These algorithms are particularly well-suited for scenarios involving relatively small datasets, as they are capable of identifying optimal decision boundaries while minimizing classification errors (Vapnik, 2013; Pisner & Schnyer, 2020). In addition to their accuracy, SVMs offer flexibility through the use of kernel functions, which allow them to capture complex, non-linear relationships within the data. This combination of precision, adaptability, and strong theoretical grounding makes SVMs a valuable tool in both experimental and applied machine learning contexts, particularly when interpretability and performance are critical.

3.2.3 Hyperparameter Tuning

Hyperparameter tuning for each machine learning model was carried out using the GridSearchCV approach in conjunction with 5-fold stratified cross-validation, allowing for

a systematic exploration of the parameter space while maintaining balanced class representation across folds. Key model-specific parameters, including *n_estimators*, *max_depth*, *num_leaves*, *min_child_samples*, and *min_split_gain*, were carefully adjusted to identify the configuration that maximized predictive performance. The effectiveness of each tuned model was then evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curves, and the corresponding area under the curve (AUC). This multi-metric evaluation provided a thorough assessment of both overall performance and class-specific predictive capabilities, ensuring that the selected models were not only accurate but also reliable and robust for practical applications.

3.2.4 Performance Evaluation of Machine Learning Models

A confusion matrix is a structured, two-dimensional table that provides a detailed framework for evaluating the performance of a classification model by directly comparing the true labels of the dataset with the labels predicted by the model. Also commonly referred to as an error matrix, this tool offers a clear and interpretable visualization of how well the model distinguishes between different classes, allowing researchers to identify not only the correctly classified instances but also the patterns of misclassification. By examining the distribution of true positives, true negatives, false positives, and false negatives within the matrix, it becomes possible to gain nuanced insights into the strengths and weaknesses of the predictive algorithm. Consequently, a series of quantitative metrics, calculated using Equations 1–4, were derived from the confusion matrix to provide a more comprehensive assessment of model performance, encompassing measures such as accuracy, precision, recall, and F1 score, and thereby enabling a thorough evaluation of both overall and class-specific predictive capabilities.

$$Accuracy = \frac{TN+TP}{TP+TN+FP+FN} \quad (3.1)$$

$$P = \frac{TP}{TP+FP} \quad (3.2)$$

$$R = \frac{TP}{TP+FN} \quad (3.3)$$

$$F\ score = \frac{2(R*P)}{R+P} \quad (3.4)$$

4. RESULTS

4.1 DATA COLLECTION AND CROSS-VALIDATION PROCEDURES

A comprehensive search of the UniProt database identified a total of 403 bacterial α/β hydrolase enzymes, including 200 classified as alkaline and 203 as acidic or acid-resistant, with homology-modeled sequences ranging from 201 to 400 amino acids in length. Further analysis revealed that many of the alkaline enzymes also exhibited resistance to high salinity, whereas a majority of the acidic enzymes demonstrated notable thermostability. The distinction between acidic and alkaline α/β hydrolases was based on differences in their amino acid sequences as well as specific physicochemical properties. In this study, machine learning methods were applied to classify these enzymes using primary sequence information and associated physicochemical features, without relying on secondary structural data. The dataset was structured with features serving as independent variables and enzyme classes as dependent labels, and it was used to train and evaluate models including Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT).

Table 4.1: Accuracy of Models Using 5-Fold Cross-Validation with a 70:30 Train-Test Split.

	Accuracy	
	Train-Test (7:3)	Cross Validation (mean)
DT	0,73	0,70 ± 0,08
RF	0,77	0,76 ± 0,08
XGB	0,75	0,75 ± 0,07
SVM	0,76	0,75± 0,09

To assess the performance of these algorithms, multiple evaluation metrics were considered, including accuracy, precision, recall, F1 score, area under the curve (AUC), and corresponding ROC curves. Table 4.1 summarizes the accuracy of the four models as determined using five-fold cross-validation combined with a 70:30 train-test split. Overall, the results indicated that most machine learning approaches were capable of distinguishing between acidic and alkaline bacterial α/β hydrolases with an accuracy ranging from 70 to

75%. While no substantial differences were observed in the overall effectiveness of the algorithms, the Random Forest model demonstrated slightly superior predictive performance compared to the other methods, as detailed in Table 4.1.

Following the initial analyses, further evaluation was conducted using the Random Forest (RF) algorithm. Hyperparameter optimization for the RF model indicated that the best-performing configuration consisted of 200 estimators, a maximum tree depth of 10, and a minimum sample split of 2. These parameter settings yielded the highest accuracy during cross-validation and were consequently adopted for the final model assessment. At the beginning of the study, accuracy values obtained from both the train-test split and cross-validation approaches were compared. A t-test conducted on these results indicated that there was no statistically significant difference between the two methods. Across all four classification algorithms examined, the train-test approach consistently produced slightly higher accuracy values. Accordingly, the train-test split was used for model training under the assumption that the dataset was reasonably balanced and that cross-validation would not provide additional performance benefits. Figure 4.1 presents the confusion matrices and ROC curves generated using the Random Forest method, providing a detailed illustration of model performance and predictive reliability.

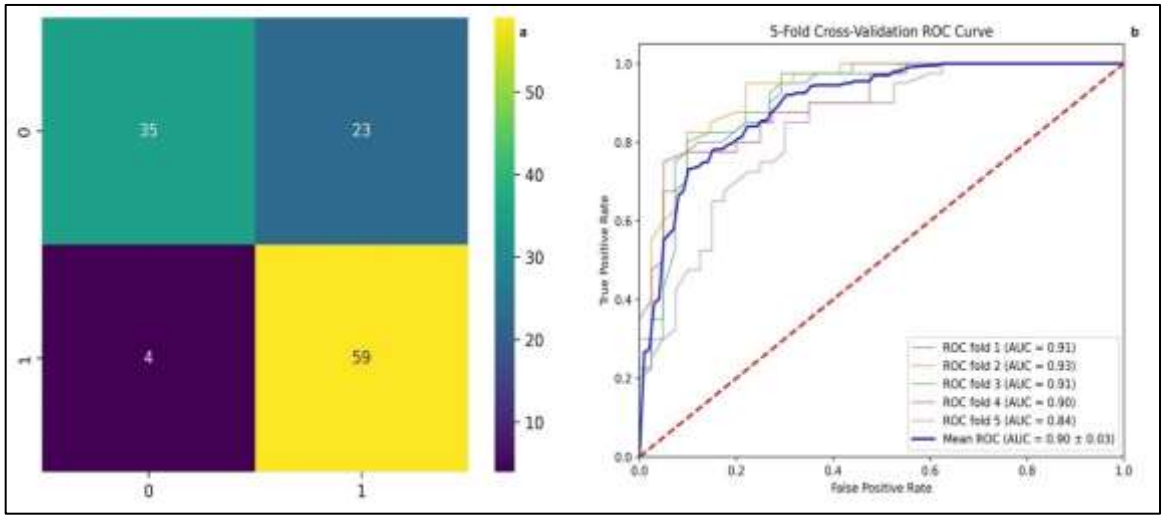

Figure 4.1: a, Confusion Matrix for RF Used in the Prediction of Acidic Bacterial α/β hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; **b**, ROC Curve with 5-fold Cross-Validation with the RF Algorithm.

The first and fourth cells of the confusion matrix (TN:35, TP:59) show the correctly expected values, while the second and third cells (FP:23, FN:4) show the mistakenly predicted values (Figure 4.1). The area under the ROC curve, or AUC value, measures how effective the separation capability is. An AUC value of 0.90 ± 0.03 was obtained using the random forest (RF) approach (Figure 4.1).

Confusion matrices were also created using other models and compared with each other. Figures 4.2-4.4 show the confusion matrices and ROC curves generated using the DTs, SVM and XGboost models, respectively.

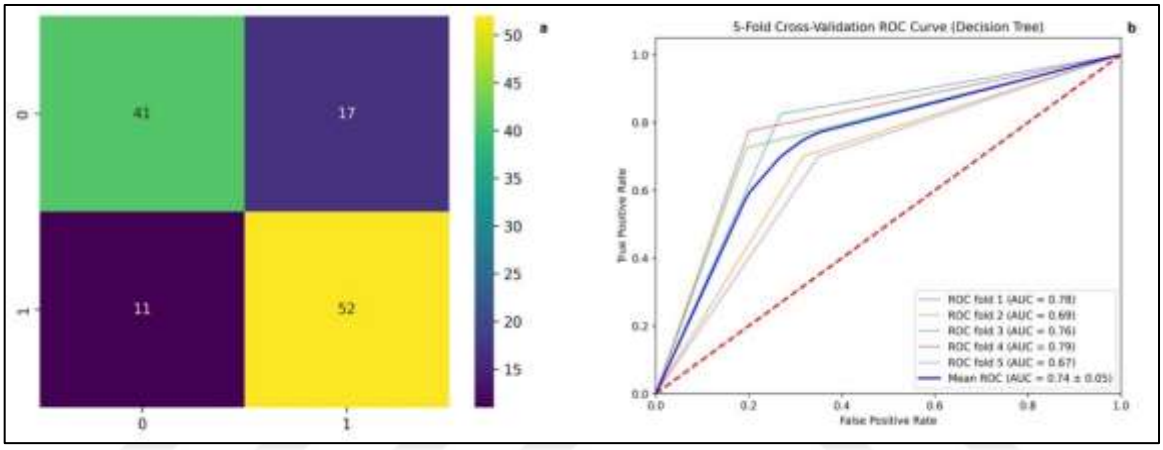

Figure 4.2: a, Confusion Matrix for DT Used in the Prediction of Acidic Bacterial α/β Hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; **b**, ROC Curve with 5-Fold Cross-validation Using the DT Algorithm.

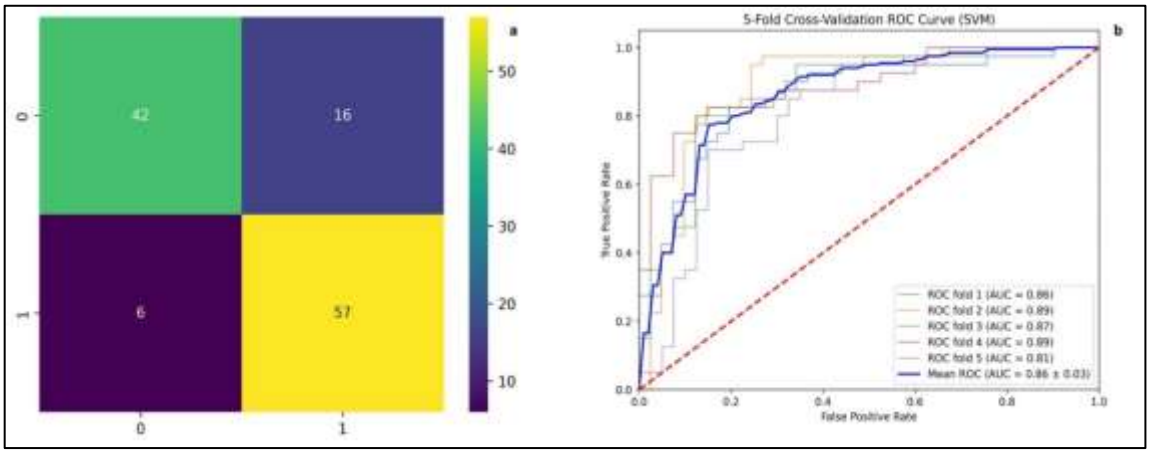

Figure 4.3: a, Confusion Matrix for SVM Used in the Prediction of Acidic Bacterial α/β Hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; **b**, ROC Curve With 5-fold Cross-validation Using the SVM Algorithm.

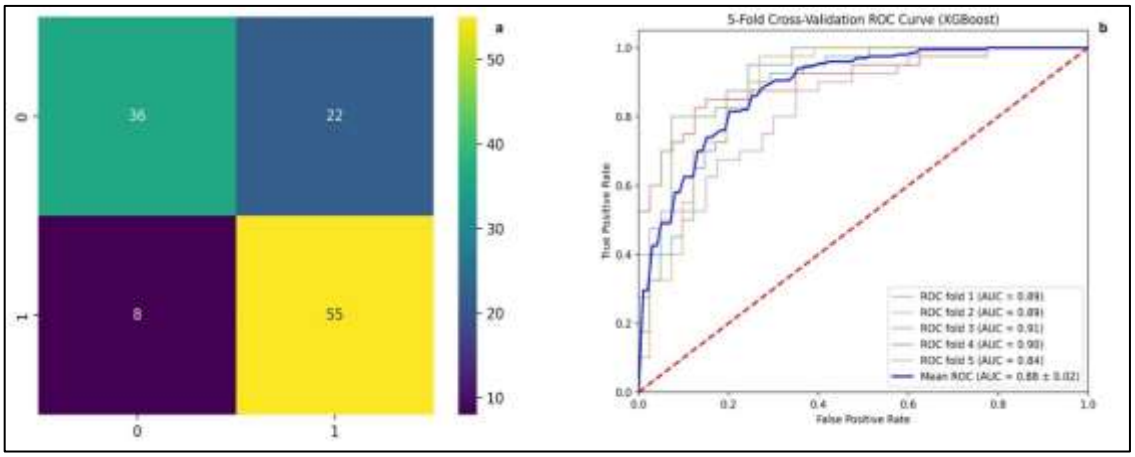

Figure 4.4: a, Confusion Matrix for XGBoost Used in the Prediction of Acidic Bacterial α/β Hydrolase Enzymes. TN: True Negative; Cell2=FP: False Positive; Cell3=FN: False Negative; Cell4=TP: True Positive; **b**, ROC Curve with 5-fold Cross-validation Using the XGBoost Algorithm.

4.2 FEATURE IMPORTANCE ANALYSIS

To further explore the factors influencing bacterial enzyme classification, the Random Forest (RF) algorithm was selected, given the consistency of performance across the various evaluation metrics. This approach was applied to investigate how the intrinsic properties of the dataset relate to bacterial categories defined by their optimal growth pH. By examining the feature importance values generated by the final RF model, as depicted in Figure 4.5, it becomes possible to identify which variables contribute most significantly to accurate classification. These insights not only highlight the primary determinants of model performance but also provide a clearer understanding of the underlying biological characteristics that distinguish acidic from alkaline bacterial enzymes.

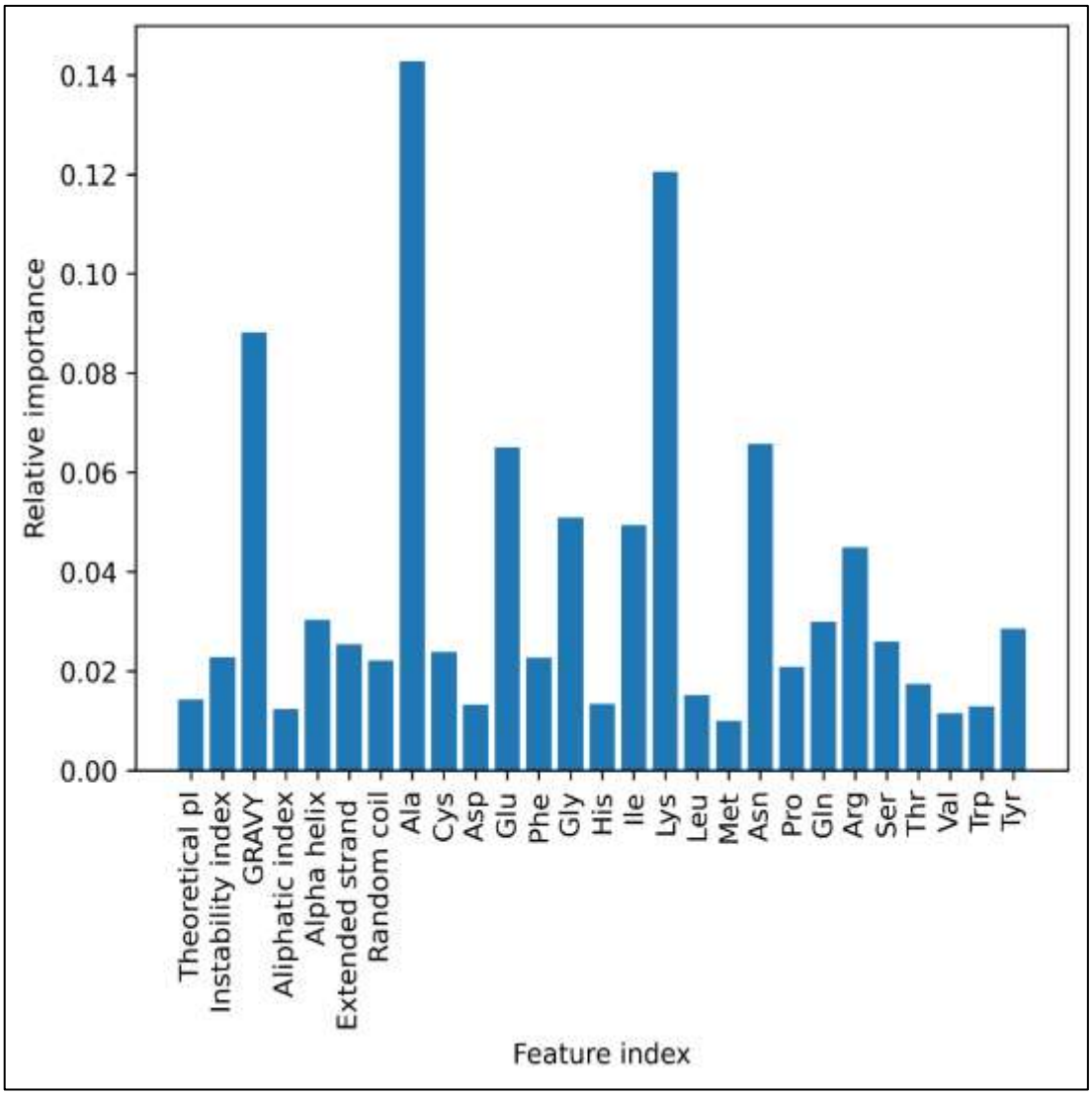

Figure 4.5: Feature Importance Results for Random Forest Algorithm.

The results indicated that among the amino acids, alanine and lysine played the most significant roles in distinguishing between acidic and alkaline enzyme groups, with relative importance values of 14.2% and 12.1%, respectively. Other properties, including the GRAVY score as well as the amino acids glutamic acid and asparagine, also contributed meaningfully to the model, with relative importance ratios of 8.9%, 6.5%, and 6.5%, respectively. The remaining features exhibited minimal impact, each accounting for less than 5% of the model's predictive power. Interestingly, the contributions of secondary structural elements—namely alpha helices, extended strands, and random coils—were found to be negligible. This lack of influence is likely due to the uniformity of secondary structures

across all enzymes in the dataset, as they all belong to the α/β hydrolase family, resulting in limited variation for the model to exploit.

Feature importance analysis indicates the relative contribution of each variable to the model's predictions, but it does not provide information about the direction of the effect (i.e., whether a feature positively or negatively influences the outcome). Feature importance, as computed by models like Random Forest, reflects the contribution of a variable to the predictive power of the model, regardless of whether its effect is positive or negative. SHAP analysis was performed to validate feature importance and examine the direction of effects and to validate the feature importance results. To address this, we compared the outcomes derived from SHAP analysis (directional and sample-specific feature contributions) with those obtained from Random Forest feature importance (overall relative contribution to the model). Both analyses consistently highlight Ala as the most influential feature, confirming its dominant role in model prediction. In addition, GRAVY and Lys are ranked among the top contributors in both approaches, further validating their importance. Similarly, structural indices (e.g., alpha helix, extended strand) display moderate effects in both analyses. A comparison table is provided as Table 4.2.

Table 4.2: Comparative Results of SHAP and Feature Importance.

Feature	SHAP summary (directional effect)	Feature importance (relative importance)	Consistency
Ala	Strong positive contribution (highest SHAP values)	**Highest importance**	Consistent
GRAVY	Strong positive contribution	High importance	Consistent
Lys	Negative contribution when high	High importance	Consistent
Ile	Moderate contribution	Moderate importance	Consistent
Glu	Negative contribution when high	Moderate importance	Consistent
Gly	Moderate positive contribution	Moderate importance	Consistent
Alpha helix / Extended strand	Moderate contributions	Moderate importance	Consistent

Additionally, the effects and directions of the variables identified by SHAP analysis are shown in Figure 4.6. In the SHAP graph below, Lys points mostly shift to the red side (high values) with negative SHAP values. A high Lys frequency has a decreasing effect on the model outcome (pushing towards the negative class). Conversely, Ala and GRAVY points are mostly located on the red side in the positive region, indicating that higher content increases the predicted outcome (pushing towards the positive class).

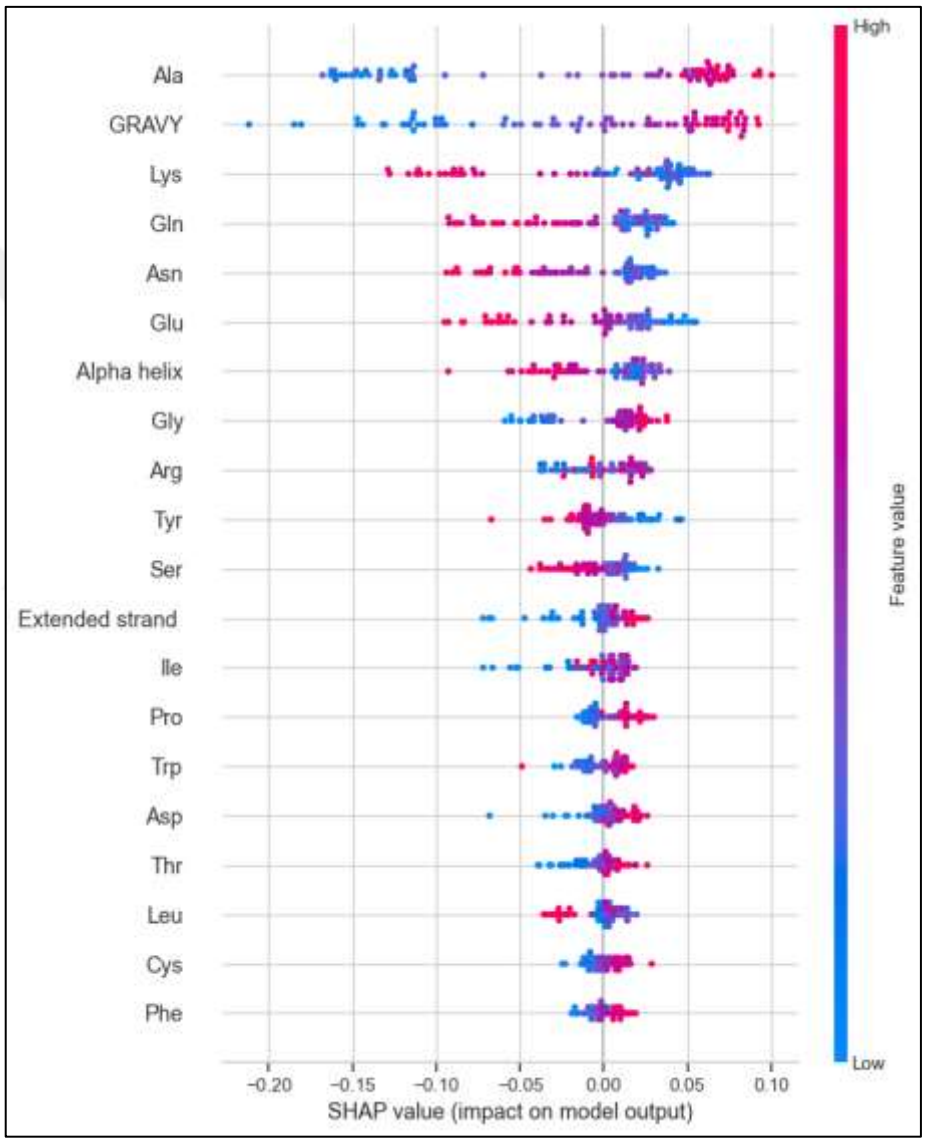

Figure 4.6: SHapley Additive Explanations (SHAP) Analysis.

When considered together, the consistency between the two methods supports the robustness of our findings. To further corroborate the reliability of the identified feature importances, a permutation test was conducted over 100 iterations using a Random Forest

classifier. The baseline model achieved an accuracy of 0.7903, indicating that the observed feature contributions are statistically meaningful and not attributable to random chance.

4.3 CORRELATION MATRIX

The correlation matrix, a straightforward analysis that does not rely on any machine learning algorithm, provides a graphical representation of the data used to identify the features most strongly correlated with one another and with the result as shown in Figure 4.7.

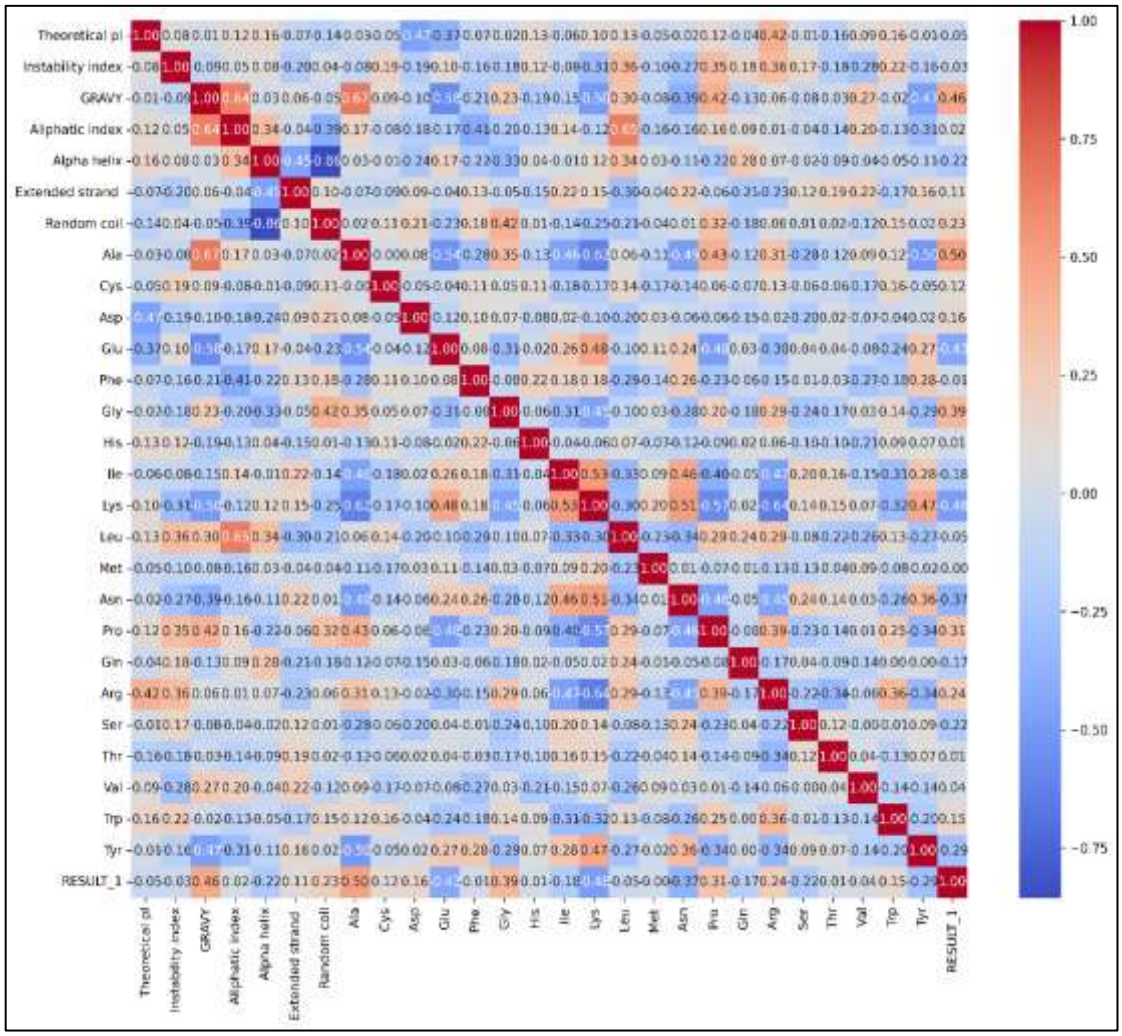

Figure 4.7: Feature Correlation Matrix Depicting Variable Interrelationships.

The acidic characteristic is positively correlated with alanine, the GRAVY score, and glycine, with correlation values of 0.5, 0.46, and 0.39, respectively, according to the correlation matrix's heatmap. The amino acids Arg, Asp, Trp, and Cys also had correlation values of 0.24, 0.16, 0.15, and 0.12, respectively. These correlation coefficients showed a weak positive connection, ranging from 0 to 0.25. Furthermore, the heatmap's dark blue indicates that the negatively linked characteristics Lys, Glu, and Asn have correlation coefficients of -0.48, -0.43, and -0.37, respectively. Additionally, the amino acids Ile, Leu, Gln, Ser, and Tyr show a negative link (correlation coefficient ranging from 0 to - 0.4) with the acidic characteristic.

5. DISCUSSION AND CONCLUSION

Acidic enzymes play a vital role in various biological and industrial applications, as they remain active and stable under low-pH conditions. In recent years, machine learning approaches have proven highly effective for classifying enzymes by evaluating their amino acid composition (AAC), secondary-structure tendencies, and physicochemical parameters (Amidi et al., 2016; Huang et al., 2023; Wang et al., 2020; Zhao et al., 2011; Ai et al., 2018; Albayrak & Sezerman, 2012; Zhang et al., 2009). Through computational algorithms that identify correlations and recurring patterns, enzyme functions can be inferred and optimized for specific biotechnological or medical purposes. Such analyses deepen our understanding of the molecular factors shaping enzyme activity and stability.

Although most previous studies have concentrated on distinguishing thermophilic enzymes from mesophilic or psychrophilic counterparts, relatively few have investigated the separation of acidic enzymes (Huang et al., 2023; Wang et al., 2020; Zhao et al., 2011; Ai et al., 2018; Albayrak & Sezerman, 2012; Taylor & Vaisman, 2010; Zhang & Fang, 2006a, 2006b, 2006c, 2007). In our earlier work (Vardar-Yel, 2025), machine-learning models were used to differentiate psychrophilic and non-psychrophilic α/β-hydrolases with respect to temperature adaptation. The present study extends this approach to pH-based classification, focusing on the distinction between acidic and alkaline members of the α/β-hydrolase family. Its novelty lies in providing a systematic, machine-learning-driven framework for identifying sequence- and property-based signatures associated with adaptation to extreme pH conditions. The models rely solely on the primary sequence and physicochemical descriptors, independent of secondary-structure variation.

Heatmap analysis of the correlation matrix revealed notable associations between amino acid characteristics and acidic behavior, with meaningful biochemical implications. Alanine showed the strongest positive correlation ($r = 0.5$) with acidity, implying that its small, non-polar side chain may aid the stabilization and functionality of acidic enzymes. It also received the highest importance score within the classification model, confirming its key role in differentiating acidic and alkaline α/β-hydrolases. Glycine ($r = 0.46$) and GRAVY value ($r = 0.39$) displayed moderate positive relationships, emphasizing the contribution of small hydrophobic residues to structural integrity under acidic conditions. Conversely, Lys,

Glu, and Asn were negatively correlated (r = −0.48, −0.41, and −0.37, respectively), suggesting that their presence may reduce enzyme stability in acidic environments. Even so, lysine remained an informative predictor, since feature-importance metrics indicate predictive contribution rather than correlation direction.

Zhang et al. (2009) similarly reported that acidic and alkaline enzymes differ in both sequence and secondary-structure composition when analyzed with the Random Forest model. They found Ser, Thr, and Asn enriched in acidic enzymes, while Glu, Arg, Leu, and Ala were characteristic of alkaline forms. Their study also noted structural preferences: α-helices in alkaline enzymes tend to contain Lys, Glu, Ile, Gln, Asp, and Arg, whereas acidic ones favor Leu, Thr, Ala, and Gly; β-sheets of alkaline enzymes include Ile, Gly, Cys, and Leu, while acidic enzymes more often have Thr, Trp, Tyr, Gln, and Phe. In coil regions, alkaline enzymes are rich in Lys, Leu, Glu, and Pro, whereas acidic enzymes contain more Ser, Thr, and Tyr.

Our analysis, limited to α/β-hydrolases, excluded secondary-structure effects and revealed positive correlations between alanine and glycine and acidic adaptation—consistent with their structural roles in maintaining enzyme stability. Alanine's small, non-polar side chain promotes α-helix formation and strengthens protein rigidity (Marqusee et al., 1989; Serrano et al., 1992; Fágáin, 1995). Under acidic conditions, this property may prevent denaturation and preserve catalytic-site geometry, as also observed by Zhang et al. (2009). Glycine's minimal steric hindrance allows conformational flexibility, enabling enzymes to withstand low-pH stress (Neurath, 1943; Yan & Sun, 1997).

A mild positive correlation for arginine (r = 0.24) aligns with Lin et al. (2013), who identified Arg as a determinant in acidic enzymes, though their dataset linked Ile and Leu more strongly to acidic types possibly reflecting differences in enzyme groups analyzed. The GRAVY index further highlighted the stabilizing role of hydrophobic residues that shield catalytic sites from protonation, maintaining both structure and activity (Morrison, 2020; Kellis et al., 1989; Pakula & Sauer, 1989; Creighton, 1990). Collectively, these observations agree with the general view that compact hydrophobic amino acids enhance protein resilience under challenging conditions (Qing et al., 2022; Pace et al., 2011; Khrustalev & Barkovsky, 2012; White, 1992).

Overall, this study demonstrates the effective use of machine learning for classifying acidic and alkaline α/β-hydrolases. The identified amino acids and physicochemical traits shed light on molecular adaptations that enable enzymatic activity at extreme pH, offering valuable insights for protein engineering and industrial biotechnology.

REFERENCES

Ai, H., Zhang, L., Zhang, J., Cui, T., Chang, A. K., & Liu, H. (2018). Discrimination of thermophilic and mesophilic proteins using support vector machine and decision tree. Current Proteomics, 15(4), 374–383.https://doi.org/10.2174/1570164615666180718143606

Akinsola, J. E. T., Osisanwo, F. Y., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. International Journal of Computer Trends and Technology, 48(3), 128–131.

Albayrak, A., & Sezerman, O. U. (2012). Discrimination of thermophilic and mesophilic proteins using reduced amino acid alphabets with n-grams. Current Bioinformatics, 7(2), 152–158. https://doi.org/10.2174/157489312800604435

Amidi, A., Amidi, S., Vlachakis, D., Paragios, N., & Zacharaki, E. I. (2016). A machine learning methodology for enzyme functional classification combining structural and protein sequence descriptors. In I. Rojas & F. Ortuño (Eds.), Bioinformatics and biomedical engineering IWBBIO 2016 (pp. 728–738). Granada: Springer. https://doi.org/10.1007/978-3-319-31744-1_63

Arakaki, A. K., Huang, Y., & Skolnick, J. (2009). EFICAz2: Enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinformatics, 10(1), 107. https://doi.org/10.1186/1471-2105-10-107.

Arpigny, J. L., & Jaeger, K. E. (1999). Bacterial lipolytic enzymes: Classification and properties. Biochemical Journal, 343(1), 177–183. https://doi.org/10.1042/0264-6021:3430177

Baker-Austin, C., & Dopson, M. (2007). Life in acid: pH homeostasis in acidophiles. Trends in Microbiology, 15(4), 165–171. https://doi.org/10.1016/j.tim.2007.02.005.

Bates, S., Hastie, T., & Tibshirani, R. (2023). Cross-validation: What does it estimate and how well does it do it? Journal of the American Statistical Association, 119(548), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Bauer, T. L., Buchholz, P. C., & Pleiss, J. (2019). The modular structure of α/β-hydrolases. The FEBS Journal, 287(5), 1035–1053. https://doi.org/10.1111/febs.15071.

Booth, I. R. (1985). Regulation of cytoplasmic pH in bacteria. Microbiological Reviews, 49(4), 359–378.

Bornscheuer, U. T. (2002). Microbial carboxyl esterases: Classification, properties and application in biocatalysis. FEMS Microbiology Reviews, 26(1), 73–81. https://doi.org/10.1111/j.1574-6976.2002.tb00599.x.

Chandra, A., Tünnermann, L., Löfstedt, T., & Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. eLife, 12, e82819. https://doi.org/10.7554/eLife.82819.

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. Artificial Intelligence Review, 56(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

Creighton, T. E. (1990). Protein folding. Biochemical Journal, 270(1), 1–16. https://doi.org/10.1042/bj2700001

Dave, S. R., & Tipre, D. R. (2012). Coal mine drainage pollution and its remediation. In T. Satyanarayana & B. Johri (Eds.), Microorganisms in environmental management (pp. 719–743). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-2229-3_32

Deka, P. P., & Weiner, J. (2024). Xgboost for regression predictive modeling and time series analysis: Learn how to build, evaluate, and deploy predictive models with expert guidance. Birmingham: Packt Publishing.

Dopson, M., & Johnson, D. B. (2012). Biodiversity, metabolism and applications of acidophilic sulfur-metabolizing microorganisms. Environmental Microbiology, 14(10), 2620–2631. https://doi.org/10.1111/j.1462-2920.2012.02749.x.

Fágáin, C. O. (1995). Understanding and increasing protein stability. Biochimica et Biophysica Acta, 1252(1), 1–14. https://doi.org/10.1016/0167-4838(95)00133-F

Ghosh, S., & Banerjee, S. (2022). The extremophiles indigenous to acid mines. In M. P. Shah & S. Dey (Eds.), Extremophiles: A paradox of nature with biotechnological implications (pp. 23–45). Berlin: De Gruyter.

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., Xavier, R. J., Knight, R., Cho, K., & Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. Nature Communications, 12(1), 3168. https://doi.org/10.1038/s41467-021-23303-9.

González, E., Vera, F., Scott, F., Guerrero, C., Bolívar, J. M., Aroca, G., Muñoz, J. Á., Ladero, M., & Santos, V. E. (2024). Acidophilic heterotrophs: Basic aspects and technological applications. Frontiers in Microbiology, 15, 1374800. https://doi.org/10.3389/fmicb.2024.1374800.

Gromiha, M. M., & Suresh, M. X. (2007). Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins: Structure, Function, and Bioinformatics, 70(4), 1274–1279. https://doi.org/10.1002/prot.21616

Gupta, G. N., Srivastava, S., Khare, S. K., & Prakash, V. (2014). Extremophiles: An overview of microorganism from extreme environment. International Journal of Agriculture, Environment and Biotechnology, 7(2), 371–380. https://doi.org/10.5958/2230-732X.2014.00258.7

Gupta, R., Biyani, Q., Pophaly, B., & Ladhani, L. (2002). Bacterial alkaline proteases: Molecular approaches and industrial applications. Applied Microbiology and Biotechnology, 59(1), 15–32. https://doi.org/10.1007/s00253-002-0975-y.

Hallberg, K. B., & Johnson, D. B. (2003). Biodiversity of acidophilic prokaryotes. Advances in Applied Microbiology, 49(3), 37–84. https://doi.org/10.1016/S0065-2164(01)49009-5

Holmquist, M. (2000). Alpha beta-hydrolase fold enzymes structures, functions and mechanisms. Current Protein and Peptide Science, 1(2), 209–235. https://doi.org/10.2174/1389203003381405.

Hon, J., Borko, S., Stourac, J., Prokop, Z., Zendulka, J., Bednar, D., Martinek, T., & Damborsky, J. (2020). EnzymeMiner: Automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Research, 48(W1), W104–W109.

Horikoshi, K. (1999). Alkaliphiles: Some applications of their products for biotechnology. Microbiology and Molecular Biology Reviews, 63(4), 735–750. https://doi.org/10.1128/MMBR.63.4.735-750.1999.

Huang, A., Lu, F., & Liu, F. (2023). Discrimination of psychrophilic enzymes using machine learning algorithms with amino acid composition descriptor. Frontiers in Microbiology, 14, 1130594. https://doi.org/10.3389/fmicb.2023.1130594

Huang, Y., Li, X.T., Jiang, Z., Liang, Z.L., Wang, P., Liu, Z.H., Li, L.Z, Yin, H.Q., Jia, Y., Huang, Z.S., Liu, S.J., & Jiang, C.Y. (2021). Key factors governing microbial community in extremely acidic mine drainage (pH< 3). Frontiers in Microbiology, 12, 761579. https://doi.org/10.3389/fmicb.2021

Johnson, D. B., & Aguilera, A. (2016). The microbiology of extremely acidic environments. In: M. V. Yates, C. H. Nakatsu, R. V. Miller, & S. D. Pillai (Eds.), Manual of environmental microbiology (4th ed.) (pp 1–24). Washington: Wiley. https://doi.org/10.1128/9781555818821.ch4.3.1

Johnson, D.B. (1995). Acidophilic microbial communities: candidates for bioremediation of acidic mine effluents. Int Biodeterior Biodegrad 35:41-58. https://doi.org/10.1016/0964-8305(95)00065-D

Johnson, D.B. (1998). Biodiversity and ecology of acidophilic microorganisms. FEMS Microbiol Ecol 27: 307-17. https://doi.org/10.1111/j.1574-6941.1998.tb00547.x

Johnson, D.B. (2007). Physiology and ecology of acidophilic microorganisms. In: Gerday C, Glansdorff N (Eds.) Physiology and biochemistry of extremophiles, Wiley, Washington, pp 255-270. https://doi.org/10.1128/9781555815813.ch20

Johnson, D.B., & Hallberg, K.B. (2008). Carbon, iron, and sulfur metabolism in acidophilic micro-organisms. Adv Microb Physiol 54: 201-55. https://doi.org/10.1016/S0065-2911(08)00003-9

Johnson, D.B., & Quatrini, R. (2020). Acidophile microbiology in space and time. Curr Issues Mol Biol 39: 63-76. https://doi.org/10.21775/cimb.039.063

Kanekar, P. P., & Kanekar, S. P. (2022). Acidophilic microorganisms. In N. K. Arora (Ed.), Diversity and biotechnology of extremophilic microorganisms from India (pp. 155–185). Singapore: Springer. https://doi.org/10.1007/978-981-19-1573-4_5

Kay, L.E. (2016). New Views of Functionally Dynamic Proteins by Solution NMR Spectroscopy. Journal of Molecular Biology, 428(2), 323–331. https://doi.org/10.1016/j.jmb.2015.11.028

Kellis, J.R., Nyberg, K., & Fersht, A.R. (1989). Energetics of complementary side chain packing in a protein hydrophobic core. Biochemistry 28:4914–22. https://doi.org/10.1021/bi00437a058

Khrustalev, V.V., & Barkovsky, E.V. (2012). Stabilization of secondary structure elements by specific combinations of hydrophilic and hydrophobic amino acid residues is more important for proteins encoded by gc-poor genes. Biochimie 94: 2706–15. https://doi.org/10.1016/j.biochi.2012.08.008

Kind, L., & Kursula, P. (2019). Structural properties and role of the endocannabinoid lipases ABHD6 and ABHD12 in lipid signalling and disease. Amino Acids, 51(2), 151–174.

Kirk, O., Borchert, T. V., & Fuglsang, C. C. (2002). Industrial enzyme applications. Current Opinion in Biotechnology, 13(4), 345–351. https://doi.org/10.1016/S0958-1669(02)00328-2.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. Artificial Intelligence Review, 39(3), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kour, D., Rana, K. L., & Kaur, T. (2019). Extremophiles for hydrolytic enzymes productions: Biodiversity and potential biotechnological applications. In G. Molina, V. Gupta, B. Singh, & N. Gathergood (Eds.), Bioprocessing for biomolecules production (pp. 321–372). Hoboken: Wiley. https://doi.org/10.1002/9781119434436.ch16

Ku, T., Lu, P., Chan, C., Wang, T., Lai, S., Lyu, P., & Hsiao, N. (2009). Predicting melting temperature directly from protein sequences. Computational Biology and Chemistry, 33(6), 445–450. https://doi.org/10.1016/j.compbiolchem.2009.10.002

Kumar, M., Thakur, V., & Raghava, G.P. (2008). Copid: composition-based protein identification. In Silico Biol 8:121-8. https://doi.org/10.3233/ISB-00350

Kumar, R., Merugu, R., Mohapatra, S., & Sharma, S. (2022). Extremophiles life of microorganisms in extreme environments. In P. Dheeran & S. Kumar (Eds.), Extremophiles general and plant biomass based biorefinery (pp. 43–66). Boca Raton: CRC Press. https://doi.org/10.1201/9781003014188

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. Drug Discovery Today, 20(3), 318–331. https://doi.org/10.1016/j.drudis.2014.10.012.

Lenfant, N., Hotelier, T., Bourne, Y., Marchot, P., & Chatonnet, A. (2013). Proteins with an alpha/beta hydrolase fold: Relationships between subfamilies in an ever-growing superfamily. Chemico-Biological Interactions, 203(2), 266–268. https://doi.org/10.1016/j.cbi.2012.09.003

Lenfant, N., Hotelier, T., Velluet, E., Bourne, Y., Marchot, P., & Chatonnet, A. (2012). ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: tools to explore diversity of functions. Nucleic Acids Research, 41(D1), D423–D429. https://doi.org/10.1093/nar/gks1154

Libbrecht, M.W., & Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920

Liese, A., Seelbach, K., & Wandrey, C. (2006). Industrial biotransformations. Wiley. https://doi.org/10.1002/3527608184

Lin, H., Chen, W., & Ding, H. (2013). Acalpred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PLoS One 8: e75726. https://doi.org/10.1371/journal.pone.0075726

Littlechild, J.A. (2015). Enzymes from Extreme Environments and Their Industrial Applications. Frontiers in Bioengineering and Biotechnology, 3. https://doi.org/10.3389/fbioe.2015.00161

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: random forest. Third International Conference, ICICA 2012. https://doi.org/10.1007/978-3-642-34062-8_32 . Accessed 25 April 2025

Lord, C.C., Thomas, G., & Brown, J.M. (2013). Mammalian alpha beta hydrolase domain (abhd) proteins: lipid metabolizing enzymes at the interface of cell signaling and energy metabolism. Biochim Biophys Acta 1831(4):792–802. https://doi.org/10.1016/j.bbalip.2013.01.002

Mamo, G., & Mattiasson, B. (2015). Alkaliphilic microorganisms in biotechnology. In S. Das et al. (Eds.), Microbial diversity and biotechnology in food security (pp. 243–270). Springer.

Marqusee, S., Robbins, V.H., & Baldwin, R.L. (1989). Unusually stable helix formation in short alanine-based peptides. Pnas 86(14): 5286–5290. https://doi.org/10.1073/pnas.86.14.528

Matin, A. (2009). Ion transport in acidophiles. In C. Gerday & N. Glansdorff (Eds.), Extremophiles (pp. 119–143). Oxford: Eolss.

Maurer, K.H. (2004). Detergent proteases. Current Opinion in Biotechnology, 15(4), 330–334. https://doi.org/10.1016/j.copbio.2004.06.005

Menzel, U., & Gottschalk, G. (1985). The internal ph of Acetobacterium wieringae and Acetobacter aceti during growth and production of acetic acid. Arch Microbiol 143: 47-51. https://doi.org/10.1007/BF00414767

Mirete, S., Morgante, V., & González-pastor, J. E. (2017). Acidophiles: Diversity and mechanisms of adaptation to acidic environments. In H. Stan-Lotter & S. Fendrihan (Eds.), Adaption of microbial life to environmental extremes: Novel research results and application (2nd ed.) (pp. 227–251). Springer. https://doi.org/10.1007/978-3-319-48327-6_9

Morrison, H. (2020). Enzyme active sites and their reaction mechanisms. Academic Press, London.

Nardini, M., & Dijkstra, B.W. (1999). α/β Hydrolase fold enzymes: The family keeps growing. Current Opinion in Structural Biology, 9(6), 732–737. https://doi.org/10.1016/S0959-440X(99)00037-8

Nardini, M., & Dijkstra, B.W. (1999). α/β Hydrolase fold enzymes: the family keeps growing. Current Opinion in Structural Biology, 9(6), 732–737. https://doi.org/10.1016/S0959-440X(99)00037-8

Neurath, H. (1943). The role of glycine in protein structure. JACS 65:2039-41. https://doi.org/10.1021/ja01250a504

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., Sussman, J. L., Verschueren, K. H., & Goldman, A. (1992). The α / β hydrolase fold. Protein Engineering, Design and Selection, 5(3), 197–211. https://doi.org/10.1093/protein/5.3.197

Ollis, D.L., & Carr, P.D. (2009). Hydrolase Fold: An Update. Protein & Peptide Letters, 16(10), 1137–1148. https://doi.org/10.2174/092986609789071298

Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Shirley, B. A., & Et, A. L. (2011). Contribution of hydrophobic interactions to protein stability. Journal of Molecular Biology, 408(3), 514–528. https://doi.org/10.1016/j.jmb.2011.02.053

Pakula, A.A., & Sauer, R.T. (1989). Genetic analysis of protein stability and function. Annu Rev Genet 23:289–310. https://doi.org/10.1146/annurev.ge.23.120189.001445

Pandey, D., Niwaria, K., & Chourasia, B. (2019). Machine learning algorithms: A review. IRJET, 6(4), 1–6.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In A. Mechelli & S. Vieira (Eds.), Machine learning methods and applications to brain disorders (pp. 101–121). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Prinzie, A., & Van den Poel, D. (2008). Random forests for multiclass classification: Random multinomial logit. Expert Systems with Applications, 34(3), 1721–1732. https://doi.org/10.1016/j.eswa.2007.01.029

Qing, R., Hao, S., Smorodina, E., Jin, D., Zalevsky, A., & Zhang, S. (2022). Protein design: From the aspect of water solubility and stability. Chemical Reviews, 122(16), 14085–14179. https://doi.org/10.1021/acs.chemrev.1c00757

Qiu, X., Zhang, Y., & Hong, H. (2021). Classification of acetic acid bacteria and their acid resistant mechanism. AMB Express, 11(1), 29. https://doi.org/10.1186/s13568-021-01189-6

Razia, S., Hadibarata, T., & Lau, S. Y. (2023). Acidophilic microorganisms in remediation of contaminants present in extremely acidic conditions. Bioprocess and Biosystems Engineering, 46(2), 341–358. https://doi.org/10.1007/s00449-022-02844-3

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., Fergus, & R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118. https://doi.org/10.1073/pnas.2016239118.

Robinson, P. K. (2015). Enzymes: Principles and biotechnological applications. Essays in Biochemistry, 59, 1–41. https://doi.org/10.1042/bse0590001.

Satyanarayana, T., Johri, B. N., & Prakash, A. (Eds.). (2016). Microorganisms in sustainable agriculture and biotechnology. Springer.

Seckbach, J. (2000). Acidophilic microorganisms. In J. Seckbach (Ed.), Journey to diverse microbial worlds: Adaptation to exotic environments (pp. 107–116). Springer.

Serrano, L., Neira, J. L., Sancho, J., & Fersht, A. R. (1992). Effect of alanine versus glycine in α-helices on protein stability. Nature, 356(6367), 453–455. https://doi.org/10.1038/356453a0

Sharma, A., Kawarabayasi, Y., & Satyanarayana, T. (2012). Acidophilic bacteria and archaea: Acid stable biocatalysts and their potential applications. Extremophiles, 16(1), 1–19. https://doi.org/10.1007/s00792-011-0402-3

Sharma, A., Parashar, D., & Satyanarayana, T. (2016). Acidophilic microbes: Biology and applications. In P. H. Rampelotto (Ed.), Biotechnology of extremophiles: Advances and challenges (pp. 215–241). Springer. https://doi.org/10.1007/978-3-319-13521-2_7

Slonczewski, J. L., Fujisawa, M., Dopson, M., & Krulwich, T. A. (2009). Cytoplasmic pH measurement and homeostasis in bacteria and archaea. Advances in Microbial Physiology, 55, 1–317. https://doi.org/10.1016/S0065-2911(09)05501-5

Solanki, A., & Lee, K. B. (2010). A step closer to complete chemical reprogramming for generating iPS cells. ChemBioChem, 11(6), 755–757. https://doi.org/10.1002/cbic.201000032.

Tao, Z., Dong, B., Teng, Z., & Zhao, Y. (2020). The classification of enzymes by deep learning. IEEE Access, 8, 89802–89811.

Taylor, T. J., & Vaisman, I. I. (2010). Discrimination of thermophilic and mesophilic proteins. BMC Structural Biology, 10(S1), 1–10. https://doi.org/10.1186/1472-6807-10-S1-S5

Tiquia-Arashiro, S., & Rodrigues, D. (Eds.). (2016). Extremophiles: Applications in nanotechnology. Springer.

Trček, J., Mira, N. P., & Jarboe, L. R. (2015). Adaptation and tolerance of bacteria against acetic acid. Applied Microbiology and Biotechnology, 99(15), 6215–6223. https://doi.org/10.1007/s00253-015-6762-3

Tütüncü, S., & Vardar-Yel, N. (2022). Extremophiles: General characteristics and industrial potential. Serüven Yayınevi.

Vapnik, V. (2013). The nature of statistical learning theory. New York: Springer.

Vardar-Yel, N. (2025). Machine learning techniques for differentiating psychrophilic and non-psychrophilic bacterial α/β hydrolase enzymes. Biologia, 80(7), 1557–1564. https://doi.org/10.1007/s11756-025-01920-9.

Wang, X. F., Gao, P., Liu, Y. F., Li, H. F., & Lu, F. (2020). Predicting thermophilic proteins by machine learning. Current Bioinformatics, 15(6), 493–502. https://doi.org/10.2174/1574893615666200207094357

White, S. H. (1992). Amino acid preferences of small proteins: Implications for protein stability and evolution. Journal of Molecular Biology, 227(4), 991–995. https://doi.org/10.1016/0022-2836(92)90515-L

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of Analytical and Testing, 2(4), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Yan, B. X., & Sun, Y. Q. (1997). Glycine residues provide flexibility for enzyme active sites. Journal of Biological Chemistry, 272(6), 3190–3194. https://doi.org/10.1074/jbc.272.6.3190

Yang, J., Li, F.-Z., & Arnold, F. H. (2024). Opportunities and challenges for machine learning-assisted enzyme engineering. ACS Central Science, 10(2), 226–241. https://doi.org/10.1021/acscentsci.3c01275.

Zhang, G., & Fang, B. (2006a). Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. Process Biochemistry, 41(5), 552–556. https://doi.org/10.1016/j.procbio.2005.09.003

Zhang, G., & Fang, B. (2006b). Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process Biochemistry, 41(8), 1792–1798. https://doi.org/10.1016/j.procbio.2006.03.026

Zhang, G., & Fang, B. (2006c). Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. Protein and Peptide Letters, 13(10), 965–970. https://doi.org/10.2174/092986606778777560

Zhang, G., & Fang, B. (2007). Logitboost classifier for discriminating thermophilic and mesophilic proteins. Journal of Biotechnology, 127(3), 417–424. https://doi.org/10.1016/j.jbiotec.2006.07.020

Zhang, G., Li, H., & Fang, B. (2009). Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition. Process Biochemistry, 44(6), 654–660. https://doi.org/10.1016/j.procbio.2009.02.007

Zhao, W., Wang, X., Deng, R., Wang, J., & Zhou, H. (2011). Discrimination of thermostable and thermophilic lipases using support vector machines. Protein and Peptide Letters, 18(7), 707–717. https://doi.org/10.2174/092986611795446094