Machine Learning Meta-analysis of Proteolytic Cleavage Specificity

by

Suhyeon Kim

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 3, 2024

Keywords: Proteolytic Cleavage Specificity, Machine Learning, Feature Selection

Approved by

Christopher Kieslich, Chair, Assistant Professor of Chemical Engineering
Selen Cremaschi, B. Redd & Susan W. Redd Eminent Scholar Chair Professor of Chemical Engineering
Peter He, George E. & Dorothy Stafford Uthlaut Endowed Professor of Chemical Engineering

Abstract

Proteolytic enzymes, such as cathepsins and matrix metalloproteinases (MMPs), play crucial roles in various physiological processes, including metabolism, cell signaling, and apoptosis. Identifying their cleavage sites is a complex challenge due to the diverse substrate specificities and regulatory mechanisms of these enzymes. This thesis investigates the use of machine learning models, particularly Support Vector Machines (SVMs) and One-Class Support Vector Machines (OCSVMs), to predict proteolytic cleavage specificity. The study introduces a novel approach utilizing Fourier Transform-based encoding of peptide sequences to capture essential biochemical properties and structural characteristics, which are used as inputs into SVM algorithms.

The research encompasses a comprehensive meta-analysis using SVM-based feature selection techniques to compare and contrast the substrate specificity of different proteases. This analysis aims to uncover distinct patterns in substrate interaction, offering valuable insights for therapeutic strategies and biomarker discovery. The datasets used in this study were sourced from the MEROPS database and included both positive data points (cleaved sequences) and synthetic negative data points (non-cleaved sequences) to ensure robustness and diversity.

Through rigorous cross-validation and hyper-parameter optimization, the SVM models demonstrated high predictive accuracy, achieving Area Under the Receiver Operating Characteristic (AUC-ROC) scores close to 1.00 for several proteases. The study also explores the performance of OCSVM models, both with and without negative class data, revealing that tailored feature selection and weighting strategies significantly enhance model performance.

The findings of this research underscore the potential of machine learning techniques in advancing bioinformatics and protease research. The developed models not only improve the precision of proteomic analyses but also support the broader field of precision medicine by providing deeper insights into protease functions in health and disease.

Acknowledgments

I would like to extend my heartfelt gratitude to several individuals and institutions that made this research possible. First and foremost, I am deeply indebted to my advisor, Dr. Christopher Kieslich, for his unwavering support, insightful guidance, and invaluable feedback throughout the course of this study. His expertise and dedication were instrumental in shaping the direction and success of this research.

I am also profoundly grateful to Dr. Selen Cremaschi, the chair of department, whose profound knowledge and rigorous approach have greatly enriched this work. Her encouragement and constructive critiques were essential in refining my research objectives and methodologies. Special thanks to Dr. Peter He for his critical feedback and for being an inspiring mentor. His thorough reviews and insightful comments significantly improved the quality of this thesis.

I would like to acknowledge my colleagues and friends in the Department of Chemical Engineering at Auburn University for their camaraderie, support, and intellectual discussions that have greatly contributed to my academic and personal growth. A special note of thanks to the administrative staff for their assistance and for ensuring a smooth research process.

I am immensely thankful to my family for their unconditional love, support, and patience. Their belief in my abilities has been a constant source of motivation and strength throughout my academic journey. Lastly, I would like to thank Auburn University for providing the resources and a conducive environment for research, and the funding agencies for their financial support, which made this study possible.

This thesis is dedicated to everyone who supported me along the way. Thank you all for your contributions and encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of Vaccines

Vaccines represent one of the most significant achievements in public health, preventing millions of deaths annually by protecting individuals from infectious diseases. [22] Vaccines work by mimicking a natural infection, stimulating the body's immune system to recognize and combat pathogens without causing the disease itself. When a vaccine introduces an antigen into the body, it triggers an immune response. [36] This involves the activation of B-cells and T-cells, which produce antibodies and kill infected cells, respectively. The immune system then "remembers" the pathogen, enabling it to respond more quickly and effectively if exposed to the actual disease in the future.

The development of vaccines is a meticulous scientific endeavor, requiring rigorous testing in clinical trials to ensure efficacy and safety before they can be widely administered. [18] Over the years, the evolution of vaccine technology has seen a shift from traditional live-attenuated and inactivated vaccines to more sophisticated approaches, such as recombinant vector vaccines and nucleic acid vaccines. These advanced methods offer the promise of quicker development times and enhanced immune responses. [10]

The COVID-19 pandemic exemplified the rapid development and deployment of vaccines. [7] Traditionally, vaccine development spans 10-15 years, but COVID-19 vaccines were developed and made available within 12-24 months. [29], [2] This unprecedented acceleration was possible due to global collaboration, prior research on related coronaviruses (SARS and MERS) and overlapping clinical trial phases. [9]Initiatives like

Operation Warp Speed in the United States further expedited this process, demonstrating the potential for swift responses to future pandemics. [24]

Vaccines can be categorized into several types based on their development methods and the technologies they use. There are inactivated vaccines,live attenuated vaccines, sub-unit vaccines, mRNA vaccines, and Peptide-based vaccines, etc. [15] Each type has its unique characteristics and advantages, making them suitable for different applications and target populations. The development and application of these diverse types of vaccines showcase the advancements in vaccine technology, offering tailored solutions for preventing a wide range of infectious diseases.

1.2 Advancements in Peptide-based Vaccines

Peptide-based vaccines are a rapidly advancing area in the field of immunotherapy, offering targeted and specific responses against various diseases, including infectious diseases, cancer, and allergies. [20]These vaccines use short sequences of proteins or peptides derived from pathogens or tumor cells to elicit an immune response without the risk of introducing whole pathogens into the body. Recent advancements in peptide vaccine technology have focused on improving the stability, delivery, and immunogenicity of peptide sequences. [34], [31]

Coupled with computational methods for epitope mapping and prediction, these approaches are streamlining the development of highly specific vaccines that can be tailored to individual immunological profiles, potentially revolutionizing personalized medicine. Computational tools like NetMHCpan and NetMHCIIpan are frequently used to predict which peptides will bind to MHC molecules, aiding in the design of vaccines that elicit strong and specific immune responses. [1]

1.3 Mechanism of Immune Response

Immune Response, specifically MHC class II molecule with T-cells, can be broken down into three main stages: antigen processing, peptide presentation, and activation of CD4+ T-cells. [37] In antigen processing stage, antigen-presenting cells capture and internalize pathogens. Inside these cells, proteolytic enzymes degrade the protein antigens into smaller peptide fragments. Next, the peptide fragments are bound to MHC class II molecules within the antigen-presenting cell. These peptide-MHC complexes are transported to the cell surface, where they are displayed for recognition by T-cells. Finally, when a T-cell receptor on a CD4+ T-cell recognizes and binds to the peptide-MHC complex on the antigen-presenting cell, it triggers the activation of the T-cell. This interaction is further stabilized by the CD4 molecule, ensuring a strong and specific immune response.

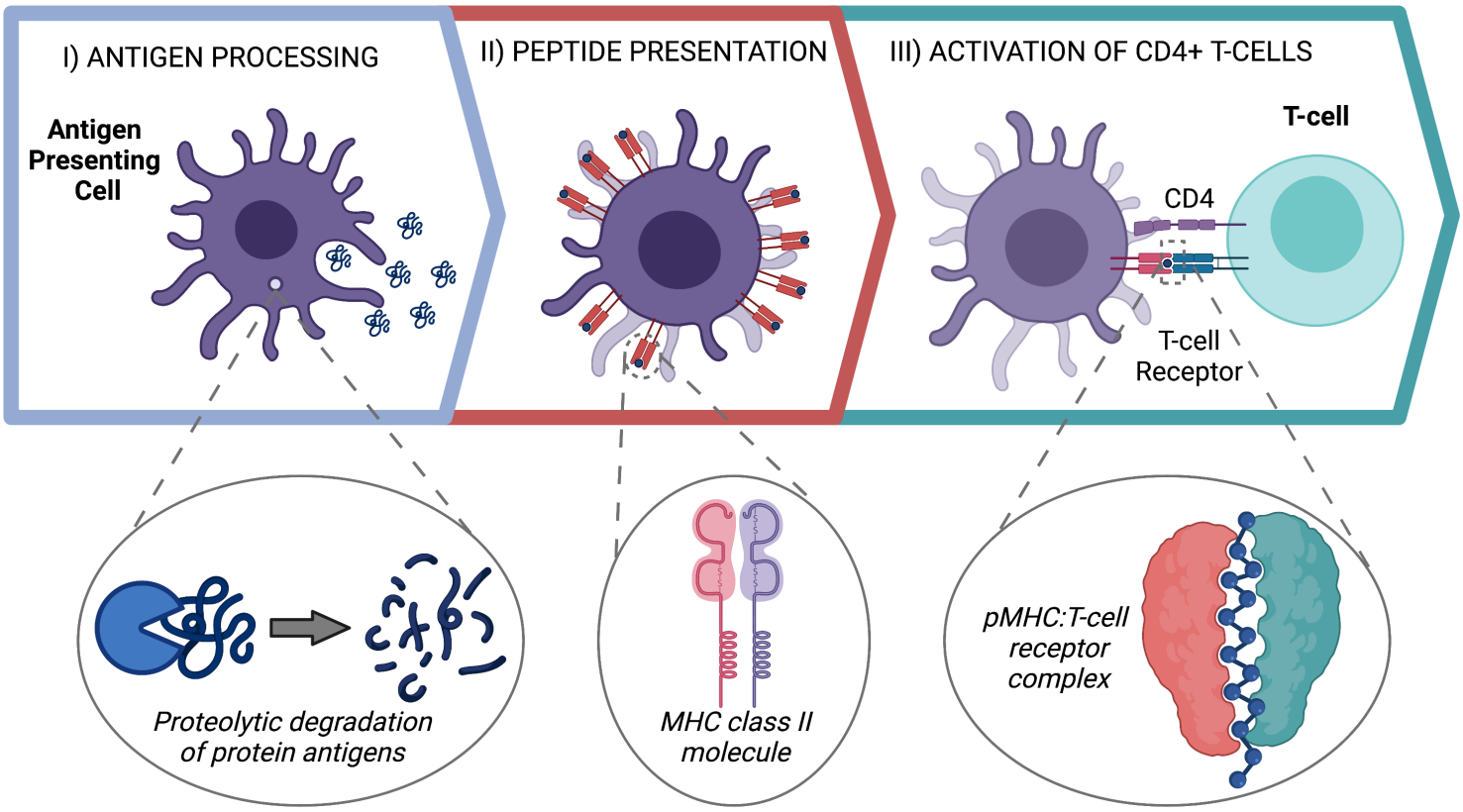

Figure 1.1: Mechanism of Immune Response

Moreover, ongoing research into understanding the mechanisms of peptide recognition by T-cells and B-cells continues to refine our ability to design more effective and long-lasting peptide vaccines. By studying how peptides interact with the immune system, scientists are able to create vaccines that not only target specific diseases but also provide durable protection. [11], [41] This knowledge is crucial for developing vaccines against cancers and chronic infections, where the immune system's ability to recognize and remember pathogens is essential for long-term efficacy. [3], [25]

These advancements highlight the potential of peptide-based vaccines to transform the landscape of vaccine development, offering new avenues for treatment and prevention in a variety of medical fields.

1.4 Role of Proteases in the Immune System

Proteases, a diverse group of enzymes, are critical components of the immune system, orchestrating a myriad of processes that maintain homeostasis and defend against pathogens. These enzymes achieve this by cleaving peptide bonds in protein substrates, a mechanism crucial for the activation, regulation, and eventual recycling of immune molecules. Proteases such as serine proteases, cysteine proteases, aspartate proteases, and metalloproteases influence various immune functions, including inflammation, cell signaling, and apoptosis. [35] By processing key molecules like cytokines, chemokines, and receptors, proteases fine-tune the immune response to ensure it is effective yet does not lead to autoimmunity or chronic inflammation, illustrating a sophisticated balance within the immune system.

Proteases are integral to both innate and adaptive immune responses, serving critical roles in antigen processing, immune surveillance, and cell signaling pathways. [38] Their activity affects the maturation of immune cells and the activation states of enzymes and receptors involved in immune signaling. For instance, proteases activate cytokines and chemokines that mediate inflammation and immune cell recruitment to sites of infection or injury. The strategic cleavage of these molecules by proteases like convertases enhances or terminates signaling pathways, thereby controlling the scope and duration of the immune response. This regulation is essential for preventing the immune system from attacking host tissues, thus avoiding autoimmune diseases. Moreover, proteases are involved in the activation of the complement system, a crucial part of the innate immune response that helps clear pathogens and damaged cells from an organism, promotes inflammation, and attacks the pathogen's cell membrane., [28]

In adaptive immunity, proteases play a vital role in antigen presentation. They degrade intracellular proteins into peptides, which are then presented on the cell surface by major histocompatibility complex (MHC) molecules. [37] This process is essential for the recognition of infected or malignant cells by T cells.

Proteases also modulate the function of immune checkpoints, which are critical for maintaining self-tolerance and modulating the duration and amplitude of physiological immune responses. [23]

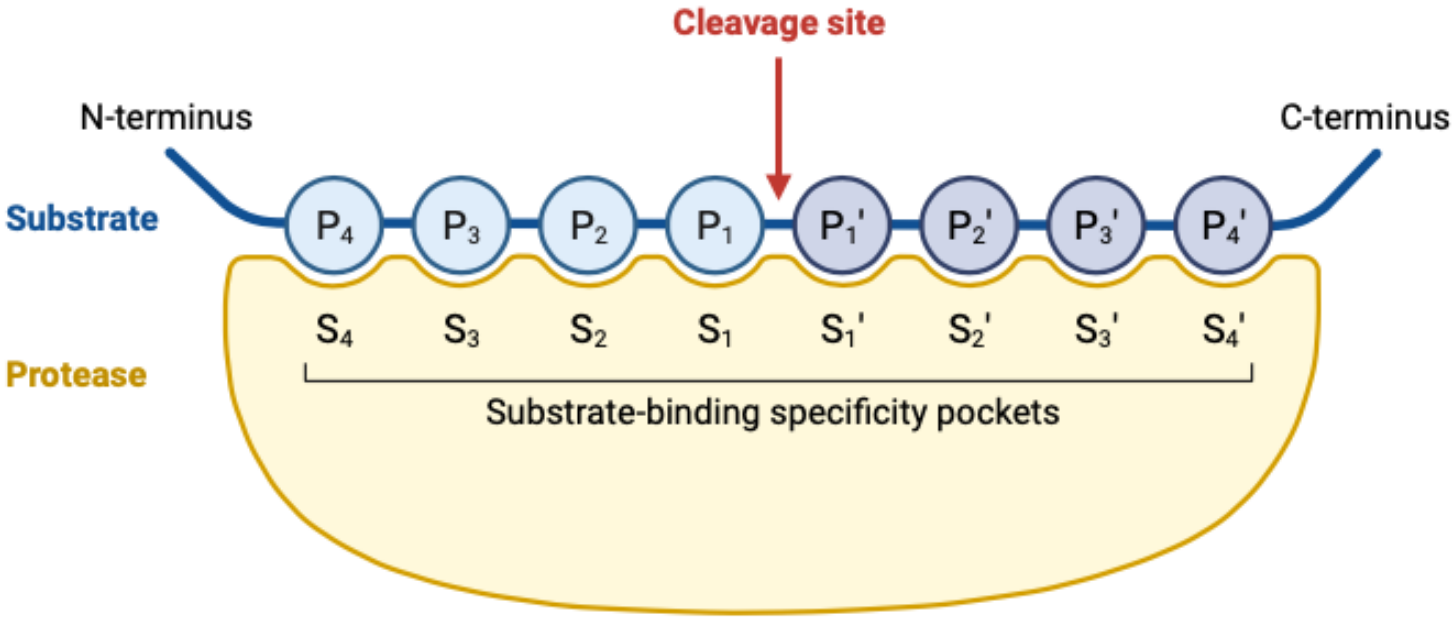

Figure 1.2: Proteases Specificity Diagram

Above figure is the diagram that shows how proteolytic degradation occurs in detail. This diagram illustrates the interaction between a protease and its substrate, focusing on the binding specificity and the cleavage process. [17] The substrate, represented by the blue chain, is a peptide or protein sequence that is recognized and cleaved by the protease, represented by the yellow region. The cleavage site, where is marked by the red arrow, is the specific location where the protease cuts the substrate. In this diagram, the cleavage occurs between the P1 and P1' positions. The substrate has an N-terminus on the left and a C-terminus on the right, indicating the direction of the peptide chain from the amino end to the carboxyl end. The substrate positions are labeled as like this, P4 to P4'. These positions represent the specific amino acid residues in the

substrate around the cleavage site. Correspondingly, the protease has substrate-binding specificity pockets labeled S4 to S4'. These pockets are where the substrate residues bind to the protease.

1.5 Research Challenges

Identifying proteolytic cleavage sites remains a formidable challenge in the field of biochemistry due to the intrinsic complexity and diversity of proteases. These enzymes exhibit a wide range of substrate specificities, catalytic mechanisms, and regulatory controls that can vary significantly across different types and families. [33] The functionality of proteases is heavily influenced by their biological context; factors such as cellular location, activity levels, and dynamic interactions with other proteins are crucial in determining their biological outcomes. This complexity is compounded by the proteases' involvement in numerous physiological processes, including metabolism, cell signaling, and apoptosis, necessitating precise predictive models to understand and manipulate their actions effectively.

One of the primary difficulties is developing computational tools that can accurately model and predict the highly specific yet variable interactions between proteases and their substrates. [5] Current predictive models often struggle to account for the full range of biological contexts in which proteases operate. For instance, the same protease might exhibit different substrate specificities in different cellular environments or under varying physiological conditions. This variability necessitates advanced algorithms that can integrate diverse data sources, including sequence motifs, structural information, and dynamic protein interactions, to improve prediction accuracy.

Moreover, the regulatory mechanisms controlling protease activity add another layer of complexity. [5] Proteases are often regulated by inhibitors, activators, and post-translational modifications, which can alter their activity and substrate specificity. Understanding these regulatory networks is essential for accurately

modeling protease function and predicting cleavage sites. Experimental validation remains a critical component of this research, as computational predictions must be corroborated with empirical data to ensure their reliability.

In summary, the challenge of identifying proteolytic cleavage sites lies in the intricate and variable nature of proteases and their interactions. Advances in computational biology, coupled with comprehensive experimental validation, are essential to overcome these hurdles and develop accurate predictive models for protease activity. This research is vital for advancing our understanding of protease functions in health and disease and for developing therapeutic strategies that target these critical enzymes effectively.

1.6 Objectives of the Study

This study aims to tackle the challenges of predicting peptide substrates for key proteolytic enzymes, specifically Cathepsin and Matrix Metalloproteinases (MMPs), by developing a robust, machine learning-based framework. Utilizing Support Vector Machines (SVMs), which are known for their effectiveness in managing complex, high-dimensional data, the study seeks to improve the accuracy of protease activity predictions. To achieve this, a novel approach involving Fourier Transform-based encoding of peptide sequences will be implemented. This method is designed to capture crucial biochemical properties and structural characteristics of peptides, thereby enhancing the predictive capabilities of the SVM algorithm.

Additionally, the study will undertake a comprehensive meta-analysis using SVM-based feature selection techniques to compare and contrast the substrate specificity of different proteases. This aspect of the research aims to elucidate distinct patterns in substrate interactions across various proteases, providing valuable insights that could inform therapeutic strategies and biomarker discovery in diseases related to protease dysregulation.

These objectives highlight a commitment to advancing the predictive capabilities of bioinformatics tools in protease research. The ultimate goal is to contribute to targeted therapeutic interventions and a deeper understanding of protease functions in health and disease. This work promises to enhance the precision of proteomic analyses and support the broader field of precision medicine.

Chapter 2

Background Information

2.1 Specific Types of Proteases

2.1.1 Cathepsin Proteases: Characteristics and Functions

Cathepsins are a family of lysosomal proteases that play pivotal roles in protein degradation within immune cells, antigen presentation, and apoptosis. These proteases are particularly important in the processing of antigens within the endosomal/lysosomal compartments of antigen-presenting cells such as dendritic cells and macrophages. By cleaving proteins into peptide fragments that can be loaded onto MHC-II molecules, cathepsins facilitate the crucial mechanism of antigen presentation to helper T cells. This is vital for initiating an adaptive immune response. For example, Cathepsin S is critical in the invariant chain processing, which is a necessary step for peptide loading onto MHC-II molecules.

Furthermore, cathepsins contribute to apoptosis, aiding in the removal of cells that could potentially become cancerous or autoimmunogenic. This process involves the activation of caspases, a family of proteases central to the execution of apoptosis. Cathepsins can initiate the caspase cascade or act downstream to amplify apoptotic signals. [42] Additionally, cathepsins regulate various other immune processes, including cytokine production and the modulation of the immune response during infections and inflammatory conditions. For example, Cathepsin G can modulate the activity of several cytokines, enhancing or inhibiting their functions, thus influencing the inflammatory response.

2.1.2 Matrix Metalloproteinases (MMPs): Characteristics and Functions

Matrix metalloproteinases (MMPs) are zinc-dependent endopeptidases involved in the degradation of extracellular matrix components, which is a key process in cellular migration, wound healing, and angiogenesis. [39] In the immune system, MMPs are crucial for the trafficking of cells to and from sites of inflammation. They modify the extracellular matrix and cleave cell surface receptors, thereby regulating the accessibility of tissues to immune cells and influencing the inflammatory response. MMPs can process a variety of bioactive molecules, including cytokines, chemokines, and growth factors, thereby modulating their activity and availability. This regulatory role is essential for coordinating the immune response and tissue remodeling.

However, the dysregulation of MMP activity can contribute to a range of diseases, including cancer, where they are implicated in tumor invasion and metastasis. MMPs can degrade the extracellular matrix barriers, facilitating cancer cell migration and invasion into surrounding tissues and the bloodstream. They also promote angiogenesis by releasing sequestered growth factors, thereby supporting tumor growth and metastasis. Consequently, MMPs are viewed as potential therapeutic targets, with inhibitors being developed to block their activity in pathological conditions. For instance, several synthetic MMP inhibitors have been investigated in clinical trials for cancer treatment, although achieving selective inhibition without adverse side effects remains a significant challenge.

In summary, proteases are indispensable for the regulation and execution of various immune processes. Their ability to cleave specific substrates allows them to modulate immune responses precisely, maintaining a balance between effective pathogen defense and prevention of autoimmunity. Understanding the intricate roles of different protease families, such as cathepsins and MMPs, in the immune system can provide valuable insights into their functions in health and disease, and guide the development of therapeutic interventions targeting these critical enzymes.

2.2 Overview of Peptidase Databases: Focus on MEROPS

MEROPS is a comprehensive database focused on peptidases (also known as proteases, proteinases, and proteolytic enzymes), their substrates, and inhibitors. Hosted by the EMBL-European Bioinformatics Institute (EMBL-EBI), the database offers extensive information on the classification and nomenclature of peptidases, alongside details on their structures, sequences, and biochemical functions. [26]

The database employs a hierarchical, structure-based classification system where peptidases are grouped into families based on statistically significant similarities in amino acid sequences. These families are further clustered into clans if they are believed to share a common evolutionary origin. Each entry in MEROPS is equipped with a summary page that provides comprehensive information, including sequence identifiers, structural data, literature references, and links to supplementary pages for deeper insights.

MEROPS also includes tools for searching and analyzing peptidase-related data, such as batch BLAST services, evolutionary trees, and substrate specificity data. [27] Researchers can access a variety of search functions, allowing them to explore peptidase and inhibitor genes, structures, and their known cleavage sites in proteins. The database also supports comparative genomics studies by facilitating searches for common peptidases across different organisms and strains.

The MEROPS database is distinguished by several unique features that set it apart from other protease databases. One of its standout characteristics is its hierarchical classification system. This system categorizes peptidases into families and clans based on statistically significant similarities in their amino acid sequences. This approach not only facilitates a clear understanding of the evolutionary relationships among enzymes but also allows for more precise categorization and comparison of peptidases.

In addition to this, MEROPS provides comprehensive data on peptidases, substrates, and inhibitors. By encompassing detailed information on not just the peptidases themselves but also their substrates and

inhibitors, MEROPS offers a holistic view of proteolytic activity. This makes it an invaluable resource for researchers looking to understand the full spectrum of interactions and functions involving proteases.

Another key feature is the extensive cross-references and links included in each entry. Each peptidase entry in MEROPS is supplemented with links to additional pages that offer data such as sequence identifiers, structural information, and literature references. [40] This thorough cross-referencing significantly enhances the usability and depth of the database, allowing researchers to access a wide array of related information easily.

The database also boasts a variety of tools for searching and analyzing data. These include batch BLAST services, evolutionary trees, and substrate specificity data. These tools facilitate the navigation and utilization of the extensive information available in MEROPS, making it easier for researchers to conduct detailed analyses and draw meaningful conclusions from the data.

Moreover, MEROPS supports comparative genomics studies. The database allows for searches that identify common peptidases across different organisms and strains. This feature is particularly useful for studies focused on the evolutionary and functional aspects of peptidases.

Finally, MEROPS is regularly updated and encourages community contributions. This ensures that the database remains current and relevant, fostering a collaborative environment where researchers can share their findings and insights. This commitment to regular updates and community involvement helps maintain the database as a cutting-edge resource for protease research.

2.3 Foundations of Machine Learning in Biomedical Research

The integration of machine learning techniques in biomedical research represents a transformative shift in how biological data are analyzed and interpreted. These techniques are particularly adept at handling the large volumes and complex nature of biomedical data, facilitating discoveries in genomics, proteomics, and

disease pathology. [16] Machine learning's ability to learn from data and make predictions can dramatically accelerate the pace of research and lead to more personalized approaches in medicine, such as in precision oncology and individualized treatment plans. [30] Furthermore, machine learning models can uncover patterns and insights that are not apparent through traditional bio-statistical methods, providing a powerful tool for predictive analytics and decision support in clinical settings. [6]

Machine learning encompasses a broad range of algorithms and techniques that enable computers to learn from and make decisions based on data. At its core, machine learning uses algorithms to parse data, learn from that data, and make informed predictions or decisions regarding new data. [13] The two main types of machine learning are supervised and unsupervised learning. Supervised learning involves training a model on a labeled dataset, which means the data includes an answer key that the model can learn to predict. Unsupervised learning, on the other hand, involves training a model on a dataset without explicit instructions on what to predict. The model tries to identify patterns and relationships within the data. These core principles are fundamental to developing algorithms that can perform a variety of tasks in biomedical research, such as diagnosing diseases from medical images or predicting patient outcomes from clinical data.

Machine learning algorithms have diverse applications in the field of biomedicine, ranging from diagnostic imaging to genetic analysis. Common algorithms include decision trees, which are simple yet powerful tools for classification and regression; random forests, which improve on the performance of decision trees by combining multiple trees to produce a more accurate and stable prediction; and neural networks, which are particularly good at processing patterns in complex datasets, such as those found in large-scale genomic studies or in sophisticated imaging techniques. Each algorithm brings strengths to specific types of data and problems, making the field of machine learning rich and varied in its approach to tackling biomedical challenges.

2.3.1 Support Vector Machines: Principles and Application in Bioinfomatics

Support Vector Machines (SVMs) are a class of supervised learning models primarily used for classification and regression tasks in the field of machine learning. SVMs have gained prominence due to their ability to handle high-dimensional data and provide robust predictive accuracy. The core principle of SVM is to find the hyperplane that best separates the data into different classes. [21] This is achieved by maximizing the margin, defined as the distance between the hyperplane and the nearest data points from each class, known as support vectors. By focusing on these critical points, SVMs can create a decision boundary that generalizes well to unseen data.

SVMs also excel in managing high-dimensional datasets, which are prevalent in fields such as genomics and text classification. Their ability to handle many features without suffering from the curse of dimensionality stems from the model's reliance on support vectors rather than the entire dataset. Additionally, SVMs are robust to overfitting, especially when the appropriate regularization parameter (C) is selected. This regularization parameter controls the trade-off between achieving a low training error and minimizing the model complexity, thus enhancing the generalization capability.

In the context of bioinformatics, SVMs have been widely adopted for tasks such as gene expression analysis, protein structure prediction, and proteolytic cleavage site prediction. Their ability to manage large, complex datasets with numerous features makes them an invaluable tool in extracting meaningful biological insights. Furthermore, SVMs' robustness to overfitting is particularly advantageous when dealing with noisy biological data. The adaptability and precision of SVMs, combined with their theoretical foundations and empirical success, underscore their critical role in advancing computational biology and bioinformatics research.

2.4 Kernel Methods

One of the distinctive features of SVMs is their use of kernel functions, which enable the model to perform non-linear classification by transforming the input data into a higher-dimensional space where a linear separation is possible. Commonly used kernels include the linear, polynomial, radial basis function (RBF), and sigmoid kernels. The choice of kernel and its parameters significantly impacts the model's performance and must be carefully tuned through cross-validation. This flexibility allows SVMs to effectively capture complex patterns and relationships within the data, making them suitable for various applications, from image recognition to bioinformatics .

The core of SVM's functionality lies in its use of kernel functions. These functions transform the original data into a higher-dimensional space without the need to compute the coordinates of the data in that space. A common kernel function is the Radial Basis Function (RBF), defined as:

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2) \tag{2.1}$$

where x_i and x_j are two feature vectors in the input space, γ is a parameter that sets the 'spread' of the kernel, and $||x_i - x_j||^2$ denotes the squared Euclidean distance between the two vectors. This kernel is particularly effective in scenarios involving complex datasets common in bioinformatics, as it can handle non-linear relationships between features.

2.5 Case Studies in Predictive Modeling of Proteolytic Cleavage Sites

2.5.1 DeepCleave: A Deep Learning Predictor for Protease Cleavage Sites

DeepCleave is a cutting-edge deep learning framework that utilizes Convolutional Neural Networks (CNNs) to predict protease cleavage sites. By leveraging sequence information, DeepCleave significantly

outperforms traditional machine learning methods in terms of accuracy and robustness. [19] The model specifically targets caspase and matrix metalloprotease substrates and cleavage sites, showcasing substantial improvements in prediction performance. One of the primary advantages of DeepCleave is its high accuracy, attributed to the deep learning techniques that can identify complex sequence patterns often overlooked by traditional methods. However, this model requires large datasets and substantial computational resources for training, which can be a limitation. Additionally, the complexity of deep learning models may reduce interpretability compared to simpler models.

2.5.2 PROSPERous: Integrating Machine Learning with Structural Information for Predicting Protease Substrate Cleavage Sites

PROSPERous integrates machine learning algorithms with structural information to accurately predict protease substrate cleavage sites. By combining features such as sequence motifs and structural data, PROSPERous enhances prediction accuracy and reliability. [32] This tool is particularly advantageous due to its comprehensive feature set, making it useful for predicting cleavage sites for a wide range of proteases. However, it relies heavily on detailed structural information, which may not always be available, potentially limiting its applicability. Additionally, the quality of the input data significantly impacts its performance, necessitating high-quality and well-curated datasets for optimal results.

2.6 Past Work of Research

In previous research, comprehensive computational analyses were conducted across various protease datasets, including Cathepsin B. The focus was on feature selection and binary classification using Support Vector Machines (SVMs). To identify the most relevant features for prediction, importance criteria were employed, allowing for ranking and selection based on their contribution to model performance. This

method ensured only the most significant variables were considered, leading to enhanced model accuracy and interpretability.

The feature selection process involved evaluating the importance of each feature through techniques such as recursive feature elimination and the use of feature importance scores derived from tree-based models. This approach was crucial in reducing the dimensionality of the datasets, thus improving the efficiency and effectiveness of the SVM models.

Following the identification of key features, this work employed Support Vector Machines (SVM) for binary classification to predict protease activity. SVM's inherent robustness and effectiveness in handling high-dimensional data made it a suitable choice. Through optimization of hyperparameters and utilization of appropriate kernel functions, the model achieved high accuracy in distinguishing between active and inactive protease states. This computational approach not only streamlined the analysis process but also yielded valuable insights into the behavior of different proteases. This deeper understanding contributes to elucidating their roles in various physiological and pathological conditions.

The methodologies and results from this past work have laid a solid foundation for further research and development in the field of protease activity prediction, offering potential pathways for targeted therapeutic interventions.

Chapter 3

Methodological Framework

3.1 Data Acquisition Strategies

To develop a predictive model for identifying proteolytic cleavage sites, the comprehensive online MEROPS database was leveraged. This resource proved crucial for obtaining datasets of substrate sequences specifically cleaved by cathepsin and matrix metalloproteinase (MMP) proteases in Homo sapiens. These known cleaved sequences formed the positive data points for model training. To enhance dataset robustness and diversity, synthetic sequences of random amino acids were generated, assumed to be non-cleaved and used as negative data points. The comprehensive dataset was then meticulously partitioned into five distinct sets to facilitate a rigorous evaluation and validation process. Each dataset functioned sequentially as both training and testing data in a systematic cross-validation scheme, aimed at reinforcing the model's reliability and accuracy.

The following table summarizes the dataset sizes for each protease:

The cathepsin datasets range from smaller samples such as Cathepsin X with 32 data points to larger samples like Cathepsin K with 10508 data points. Similarly, the MMP datasets vary in size, with MMP 1 having 95 data points and MMP 2 comprising 3418 data points.

These datasets were curated from various protease databases and include a wide range of activity measurements to ensure comprehensive analysis. The diversity in dataset sizes reflects the varying availability of protease activity data, which was considered during the feature selection and model training processes.

The selection of these datasets was guided by the need to cover a broad spectrum of protease activities, ensuring that the models developed in this research are robust and generalizable. The varying dataset sizes

Table 3.1: Dataset Sizes for Cathepsins and MMPs

Protease	**Dataset Size**
Cathepsin B	4623
Cathepsin D	915
Cathepsin E	1586
Cathepsin G	448
Cathepsin H	42
Cathepsin K	10508
Cathepsin L	6658
Cathepsin S	3115
Cathepsin V	5764
Cathepsin X	32
MMP 1	83
MMP 10	95
MMP 14	136
MMP 2	3418
MMP 25	167
MMP 3	2462
MMP 7	191
MMP 8	117
MMP 9	400

also present a valuable opportunity to evaluate the performance of the models across both abundant and sparse data conditions.

These datasets provide a solid foundation for evaluating the performance of SVM and one-class SVM models in predicting proteolytic cleavage sites, which is the core focus of this research. The detailed analysis of these datasets, including their preparation and preprocessing, is further discussed in the following sections.

3.1.1 Fourier-based Encoding of Protein Sequences

Effective dataset encoding is a cornerstone for the success of machine learning processes, particularly in biological sequence analysis. In this research, the Blocks Substitution Matrix (BLOSUM 100) utilized the

Blocks Substitution Matrix (BLOSUM 100) for the initial encoding to capture the evolutionary relationships between amino acids in the peptides. [14] BLOSUM 100 was specifically chosen because it focuses on the substitutions within blocks of conserved sequences, which are crucial for maintaining structural and functional integrity of proteins. This matrix is designed to score alignments based on observed substitutions in highly conserved regions of proteins, making it particularly suitable for analyzing evolutionary conserved sequences.

To further refine the data representation and improve the model's ability to interpret these sequences, the Fast Fourier Transform (FFT) is applied. FFT was instrumental in normalizing the encoded sequences, thereby transforming the peptide data into a format that enhances the model's learning efficiency. [8] The FFT process converts the sequences into the frequency domain, emphasizing the periodic patterns within the data rather than their positional sequence. This transformation allows the model to focus on recognizing underlying patterns, which is more effective than simple sequence alignment for many machine learning applications. [4]

By integrating BLOSUM 100 with FFT, the encoding process leverages both evolutionary information and pattern recognition capabilities. This dual approach ensures that the model can capture the essential biochemical properties and structural characteristics of the peptides, thus improving predictive performance and providing deeper insights into the biological phenomena being studied.

3.2 Optimization of Hyper-parameters

Optimizing hyper-parameters is crucial for enhancing a machine learning model's predictive accuracy and efficiency. This study employed a systematic approach to tune these parameters. Grid search and cross-validation methods were utilized to iteratively adjust the hyper-parameters and identify the combination yielding the best performance on the validation data. This process was essential for fine-tuning the model to avoid overfitting and ensure generalizability to unseen data.

3.3 Strategies for Effective Cross-Validation

A structured cross-validation strategy was implemented to validate the effectiveness and robustness of the SVM-based predictive model. Five distinct train/test sets were utilized, enabling comprehensive training and validation across various data subsets. This approach yielded a thorough assessment of model performance while also aiding in the identification of potential biases or variances within the data. By computing and averaging performance metrics across all cross-validation folds, a more precise and reliable estimation of the model's predictive capabilities on unseen data was achieved.

Cross-validation is essential in hyper-parameter optimization to ensure the model's performance is robust and generalizable. It involves partitioning the data into training and validation sets multiple times and averaging the performance metrics across these splits. This approach helps in estimating the generalization error and selecting hyper-parameters that perform well across different subsets of data.

Evaluating the model's performance involves using metrics like accuracy, precision, recall, F1 score, and AUC-ROC, depending on the specific task. These metrics provide a quantitative basis for comparing different hyper-parameter settings and selecting the optimal model configuration.

3.4 Approaches to Feature Selection

Feature selection played a critical role in model development, particularly in optimizing SVM performance. Initially, all possible features derived from peptide sequence encoding were included. To identify the most informative features, a feature-ranking mechanism based on their impact on model performance was employed. Non-contributing features were systematically removed using a predetermined cutoff threshold. This iterative refinement process continued until no significant improvement in model performance was observed. In parallel, the exploration of a One-Class SVM framework as an alternative approach was conducted.

This framework is well-suited for unbalanced datasets with underrepresented negative samples (non-cleaved substrates) due to its ability to define a decision boundary around positive samples. This approach enhances the model's ability to generalize from limited negative data. The dual application of both SVM frameworks ensured the retention of the most relevant features for final model training and validation, solidifying the foundation for a robust predictive model. [12]

In the development of the Support Vector Machine (SVM) model for predicting proteolytic cleavage sites, a crucial aspect of enhancing model performance involved rigorous feature selection. This process was guided by mathematical formulations designed to evaluate the importance and contribution of each feature in the classification task. The objective was to systematically identify and retain the most predictive features, thereby refining the model's efficiency and accuracy. The following equations play a pivotal role in this methodology:

$$crit_k = -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \left(\frac{\partial K(x_i, z, x_j, z)}{\partial x_k}\right)_{z=z^*} \tag{3.1}$$

$$k_{worst} = \arg\max_k crit_k$$

This equation defines the criterion crit_k for assessing the importance of the k-th feature. It is calculated by taking the partial derivative of the kernel function K with respect to the k-th feature at an optimal point z^*. The double summation runs over all pairs of training data points (i, j), where α_i and α_j are their respective labels. This formulation is integral to understanding how changes in a specific feature dimension x_k influence the decision boundary shaped by the SVM. The negative sign indicates a minimization problem, emphasizing features that lead to a greater impact on the model's discriminative capability. This approach helps in optimizing the feature set, ensuring only the most significant features are retained.

$$f(x) = sgn\left(\sum_i \alpha_i k(x_i, x) - \rho\right) \tag{3.2}$$

The decision function $f(x)$ is fundamental to the SVM's operation. It computes the classification of a new data point x by evaluating the sign of the weighted sum of the kernel evaluations between x and the support vectors x_i, offset by a threshold ρ. The kernel function k encapsulates the similarity measure between data points, and α_i represents the learned weight of the i-th support vector. The decision function effectively partitions the feature space into classes based on the sign of the result, classifying each point accordingly.

The development of the one-class Support Vector Machine (SVM) or Support Vector Data Description (SVDD) mode involved a refined approach to feature selection. This approach was crucial for effectively discriminating between the classes in high-dimensional feature spaces, particularly when dealing with imbalanced datasets typical of proteolytic cleavage site prediction. The following equations form the mathematical foundation for the models used and the corresponding feature selection mechanism implemented:

$$crit_k = \sum_{i=1}^{l} \alpha_i^* \left(\frac{\partial K(x_i, z, x_i, z)}{\partial x_k}\right)_{z=z^*} - \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \left(\frac{\partial K(x_i, z, x_j, z)}{\partial x_k}\right)_{z=z^*} \tag{3.3}$$

$$k_{worst} = \arg\max_k crit_k$$

This equation serves as the objective function for assessing the importance of the k-th feature in the dataset. The first term calculates the contribution of each feature to the kernel's diagonal, which relates to the self-similarity of each point and reflects the stability of each point within its own feature space. The second term involves a pairwise evaluation across all data points, weighted by their respective labels and Lagrange multipliers, α_i and α_j. This structure of the equation allows for capturing the influence of individual features on the decision boundary's shape, thus providing a basis for feature importance that informs the selection

process. To identify the least important feature, k_{worst}, this equation is used. It determines the feature with the maximum criterion value, implying that its removal would theoretically have the least impact on the model's performance. This method of feature reduction is targeted to simplify the model while attempting to retain its predictive accuracy.

$$f(x) = sgn\left(R^2 - \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) + 2\sum_{i} \alpha_i k(x_i, x) - k(x, x)\right) \tag{3.4}$$

The decision function defines how new data points are classified based on the model trained with one-class SVM or SVDD. It evaluates whether a new point x lies within a defined boundary R^2 in the feature space. The function computes the difference between this boundary and the kernel-induced distance measures: the summed pairwise distances between support vectors, the distance of the point x from support vectors, and the point's distance from itself. The sign of this evaluation determines if x is similar to the training data (inside the boundary) or dissimilar (outside the boundary), thus classifying it accordingly.

3.5 Enhancing Feature Selection with Weight Optimization

Incorporating weight optimization into the feature selection process for SVM models forms a pivotal aspect of my methodology to balance feature importance and contribution effectively. This approach was designed to refine the SVM and one-class SVM models by evaluating different weighting schemes and determining their impact on model performance.

3.5.1 Weight Optimization in Feature Selection

During the feature selection phase, weights ranging from 0 to 1 were assigned to features at increments of 0.25. These weights adjusted the influence of each feature on the SVM's decision function according to their calculated importance and contribution. The weighting process involved:

- Weight = 0: The weight reflects exclusive reliance on feature importance. The feature's own characteristics dictate its influence on the model, irrespective of its interaction with other features.
- Weight = 0.25: Here, 25% of the weight is attributed to the feature's contribution and 75% to its importance. This setting allows the feature to moderately influence the model based on its relational dynamics with other features, while still maintaining a strong anchor in its intrinsic properties.
- Weight = 0.5: This equal weighting (50% contribution and 50% importance) offers a balanced approach, integrating both the inherent qualities of the feature and its synergistic effects with others.
- Weight = 0.75: At this level, 75% of the weight is based on contribution and 25% on importance, emphasizing the feature's interactive effects over its standalone characteristics.
- Weight = 1: The weight is solely based on feature contribution. This setting maximizes the relational impact of the feature, focusing entirely on how it functions in conjunction with other features, which is crucial in complex models where interactions are significant.

Through systematic adjustment of these weights, the SVM models were fine-tuned to optimize their performance, resulting in a robust and accurate predictive model for protease activity. This approach allowed for a nuanced understanding of how each feature contributes to the overall model, leading to better generalization and improved accuracy on validation data.

3.6 Model Performance Evaluation Strategy

The selection of the best model configuration is driven by rigorous performance evaluations using metrics tailored to the specifics of each SVM variant:

- SVM: For this model, the primary evaluation metrics are the Area Under the Receiver Operating Characteristic (AUC-ROC) and the F1 score. The AUC-ROC is essential for gauging the model's ability to distinguish between classes, providing a single scalar value that represents the model's discriminatory power across various threshold settings. This is particularly vital in bioinformatics, where distinguishing between different biological states is crucial. The F1 score, a harmonic mean of precision and recall, is used to balance the trade-off between the precision (the accuracy of positive predictions) and recall (the ability to find all positive instances), which is important for tasks involving balanced datasets.[45]

- One-Class SVM: The evaluation of this model focuses on the F1 score and recall, given its application in anomaly detection. The F1 score provides a balanced measure that is particularly useful in datasets where anomalies (true positive cases) are rare. High recall is prioritized to ensure comprehensive detection of true anomalies, which is essential for reliable outlier identification in biological data. This ensures that the model captures as many positive instances as possible, crucial in scenarios like identifying rare disease markers or detecting fraudulent activities in biomedical datasets.[46]

The optimized feature weights demonstrating superior performance metrics were integrated into the SVM training process. This involved modifying the kernel function of the SVMs to include these weights, allowing for nuanced interpretation and handling of data features. This methodology not only improved accuracy and computational efficiency but also provided profound insights into the dynamics of each feature within the dataset, enhancing our understanding of complex biological phenomena.

This structured approach to feature selection and model evaluation ensures that the SVM models developed are robust and precise, finely adapted to meet the intricate demands of bioinformatics research. The methodologies outlined in this thesis significantly advance the application of machine learning in the analysis of high-dimensional biological data, providing a solid foundation for future research and practical applications in the field.

Chapter 4

Empirical Results

This chapter presents the key findings from the experiments and analyses conducted in this research. The goal is to provide a clear and structured presentation of the results, focusing on the performance metrics of various proteases using SVM and one-class SVM models.

4.1 Application of Support Vector Machines Compare to Specificity Matrix

4.1.1 Performance of SVM Models

In this section, the performance metrics of the Support Vector Machine (SVM) models are presented across various proteases. The evaluation focuses on identifying the highest scores achieved and determining the best-performing models for each protease. Comprehensive tables and figures are provided to illustrate these results in detail.

The following table summarizes the highest scores and the best models for each protease:

Table 4.1: Model Performance Metrics for Various Proteases

Protease	**Highest Score**	**Best Model**
Cathepsin B	0.85	0.75
Cathepsin D	0.88	0.75
Cathepsin E	0.89	0.50
Cathepsin G	0.88	0.50
Cathepsin H	0.90	1.00
Cathepsin K	0.83	1.00
Cathepsin L	0.84	0.75
Cathepsin S	0.81	1.00
Cathepsin V	0.84	1.00
Cathepsin X	1.00	1.00
MMP 1	0.93	0.75
MMP 2	0.88	0.75
MMP 3	0.85	0.50
MMP 7	0.88	0.50
MMP 8	0.93	0.75
MMP 9	0.91	0.75
MMP 10	0.84	1.00
MMP 14	0.92	0.50
MMP 25	0.92	0.50

The table illustrates the highest AUC scores achieved by the SVM models for each protease and identifies the corresponding best model based on the weight optimization during feature selection. Cathepsin B achieved a highest score of 0.85, with its best model scoring 0.75. Cathepsin D reached a highest score of 0.88, also with a best model score of 0.75. Despite high scores of 0.89 and 0.88 for Cathepsin E and G, respectively, their best model scores are lower at 0.50 each. In contrast, Cathepsin H and K exhibit highest scores of 0.90

and 0.83, with their best models scoring a perfect 1.00. Cathepsin S and V both have highest scores of 0.81 and 0.84, with their best models also scoring 1.00. Notably, Cathepsin X stands out with a perfect score of 1.00 for both metrics. Among the MMP family, MMP 1 and MMP 8 have highest scores of 0.93 and best model scores of 0.75. MMP 2 and MMP 9 show highest scores of 0.88 and 0.91, respectively, with best model scores of 0.75. MMP 3 and MMP 7 have lower highest scores of 0.85 and 0.88, with best model scores of 0.50. Finally, MMP 10 has a highest score of 0.84 with a perfect best model score of 1.00, while MMP 14 and MMP 25 both have highest scores of 0.92 but lower best model scores of 0.50.

Especially, for Cathepsin H and X, and MMP 10, the highest scores were achieved with a weight of 1.00, indicating that feature contribution alone provided the best results.

The performance plot for various proteases illustrates the model scores as a function of the number of features included in the SVM models. The x-axis represents the number of features (nFeats), while the y-axis shows the performance score. Each line corresponds to a different protease, as identified in the legend.

A comparative analysis of model performances reveals that Cathepsin X, with a score of 1.00, represents an ideal model performance, indicating highly relevant feature selection and model robustness. However, the lower scores for Cathepsins like Cathepsin B (0.85) and Cathepsin L (0.84) suggest that additional feature selection or model optimization is required to enhance predictive accuracy. Similarly, for the MMPs, the high scores of MMP 1 and MMP 8 (0.93) reflect strong model performance, whereas the variation in scores among other MMPs (ranging from 0.84 to 0.92) underscores the necessity for customized feature selection strategies. This tailored approach is essential for maximizing model performance and ensuring the accurate prediction of proteolytic activities across diverse protease types. The analysis confirms the efficacy of SVM models in predicting proteolytic cleavage sites, providing crucial insights for therapeutic target identification and other bioinformatics applications.

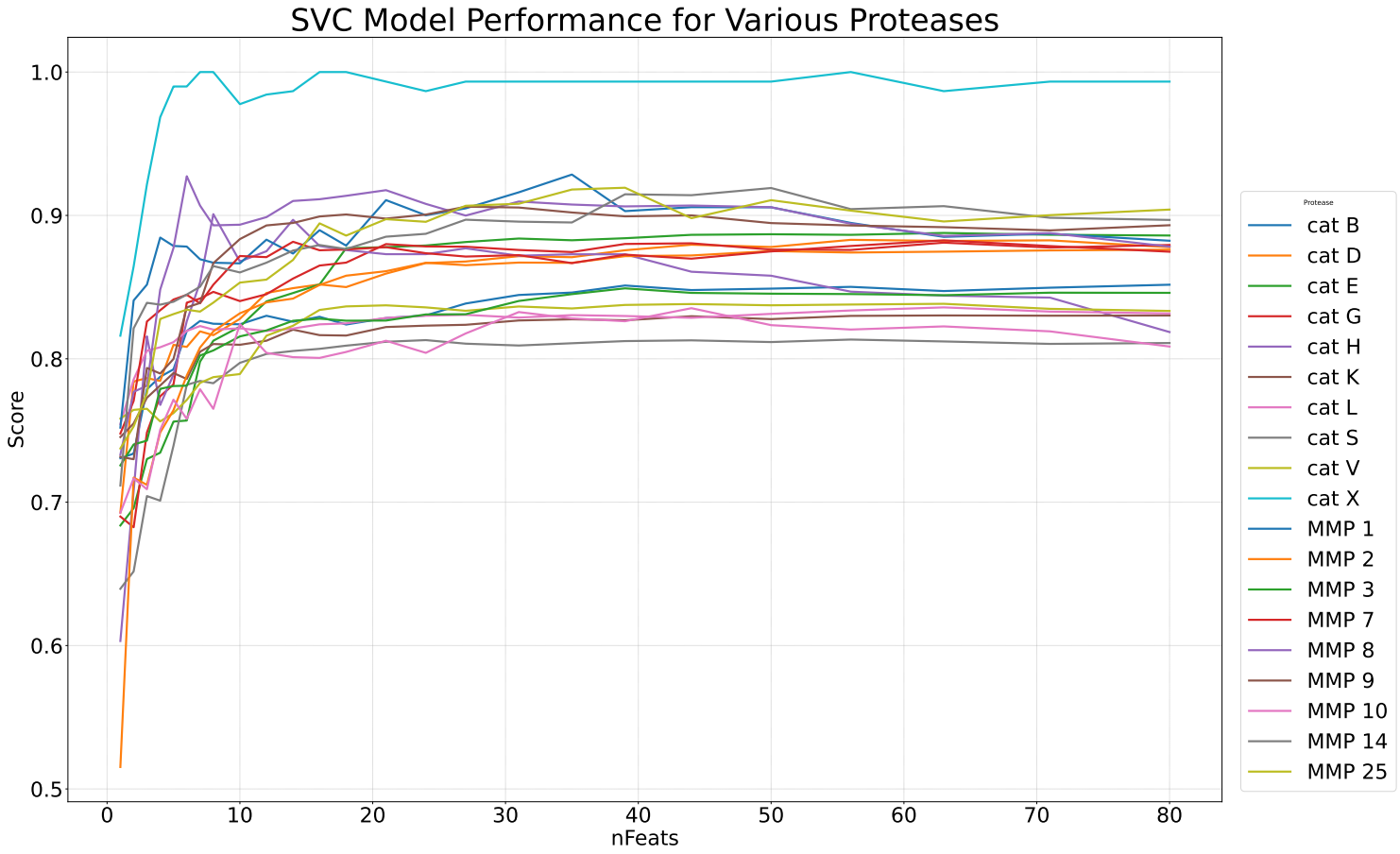

Figure 4.1: Score Plots for Various Proteases

The x-axis represents the number of features (nFeats), while the y-axis represents the performance score.

For most proteases, Initially, all protease models display lower performance scores, which improve rapidly as more features are included. Beyond 20-30 features, the performance scores for many proteases begin to stabilize, indicating that additional features beyond this point contribute less significantly to improving model performance. This suggests that a small number of features are highly informative and contribute significantly to the model's performance. The scores plateau, showing diminishing returns on model improvement with an increasing number of features.

Some proteases exhibit fluctuations in their performance scores. Notably, Cathepsin X achieves a perfect score of 1.00 early on and maintains it, reflecting highly effective feature selection. This could be due to the inclusion of less relevant features that introduce noise into the model. In contrast, other proteases, such as Cathepsin B and Cathepsin L, exhibit moderate performance scores.

The analysis of the performance plot reveals that the optimal number of features for most proteases lies between 20 and 30. This range strikes an effective balance between model complexity and predictive accuracy. The consistently high scores for proteases like Cathepsin X, H, K, S, and V demonstrate robust model performance and effective feature selection. The variability in plateau scores across different proteases highlights the importance of customized feature selection strategies to achieve optimal performance. The diminishing returns observed beyond 20-30 features indicate that further feature addition may not be beneficial, emphasizing the need for efficient model design. For proteases with lower scores, further refinement in feature selection and model parameters is necessary to enhance predictive accuracy.

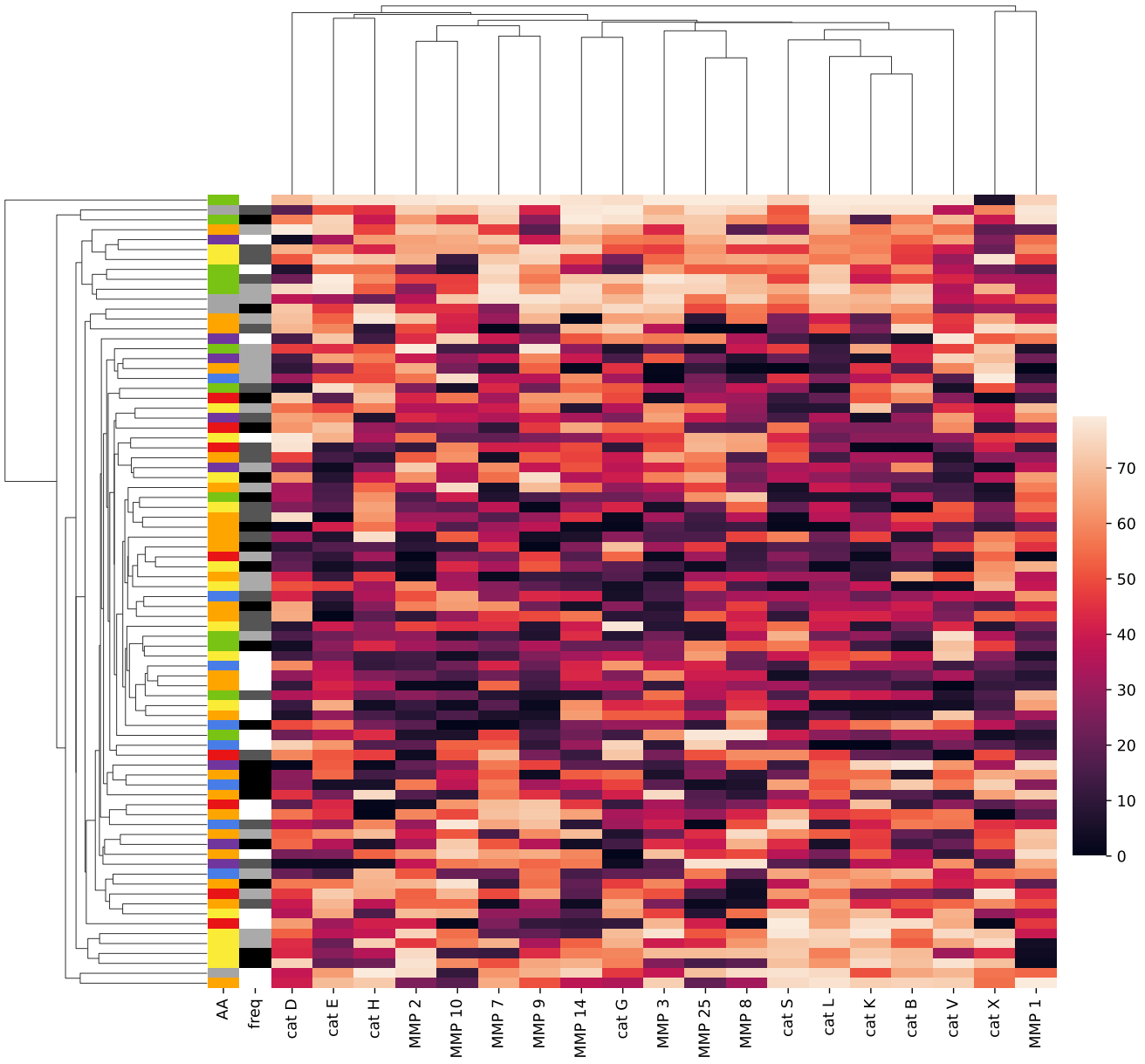

Figure 4.2: Cluster Map for Various Proteases Based on the Feature Rank from SVC

The clustermap illustrates the performance of SVM models across different feature sets for various proteases. The x-axis represents different feature sets, while the y-axis lists the proteases. The color bars on

the cluster map represent the top 80 features for predicting peptides. This visual representation allows us to quickly identify which features are most significant for each protease. Darker colors represent lower values and lighter colors represent higher values. A color code is assigned to visualize the feature. Different colors represent various amino acids. Black and white coding represents the frequency of features. This indicates unique fingerprinting of features. This cluster map provides a powerful visual tool for analyzing the relationships between proteases and their predictive features. By examining these patterns, deeper insights into protease specificity and functionality can be gain.

Analysis of the clustermap reveals clusters of proteases with similar performance characteristics, such as Cathepsin X and MMP 1. This clustering indicates similar performance trends and potentially similar feature relevance. The variability in performance scores across different feature sets underscores the importance of effective feature selection. Feature sets resulting in higher scores are more effective in capturing the relevant biochemical properties and structural characteristics needed for accurate predictions. Proteases with consistently high scores benefit from the chosen features, while those with lower scores may require additional or alternative features to improve model performance. This analysis emphasizes the need for a strategic focus on tailored feature selection and model optimization for individual proteases to enhance predictive accuracy and robustness.

4.1.2 Comparison to Specificity Matrix Method Performance

The plot in Figure X compares the Area Under the Curve (AUC) values for Specificity and Support Vector Machine (SVM) models across different proteases. The x-axis represents the Specificity AUC, while the y-axis represents the SVM AUC. Data points represent individual proteases, with a dashed line (y=x) included for reference.

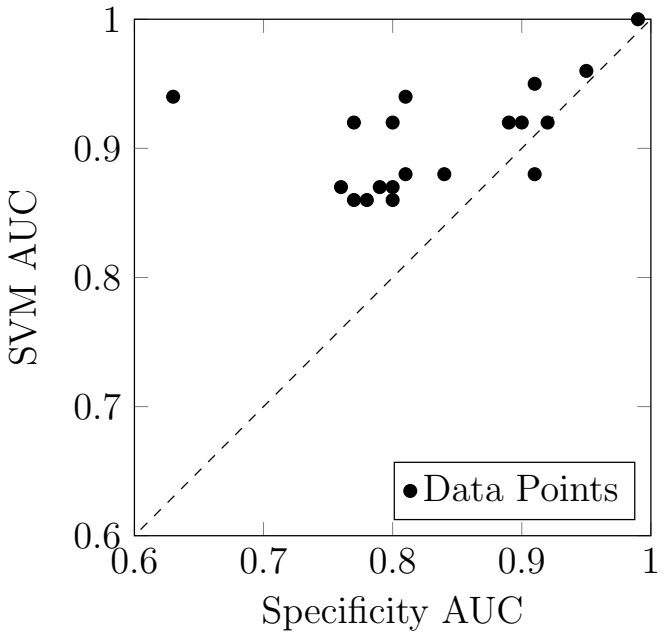

Figure 4.3: Comparison of SVM AUC and Specificity AUC

The majority of the points lie above the diagonal line $y = x$, indicating that the SVM models generally achieve higher AUC scores compared to the Specificity Matrix method.

The analysis reveals that proteases with high AUC scores on both axes demonstrate highly effective feature selection and robust model performance. The clustering of data points above the y=x line indicates that the SVM models generally perform better than the Specificity models, capturing relevant features and biochemical properties more effectively. The superior performance of the SVM model, particularly for proteases with points such as (0.63, 0.94), underscores its greater predictive power and advanced optimization techniques. This analysis highlights the importance of utilizing SVM models for more accurate predictions of proteolytic activities.

For many proteases, the SVM models outperform the Specificity Matrix method, as evidenced by the higher AUC scores. This demonstrates the effectiveness of the SVM approach in capturing the complex

patterns in the data that the Specificity Matrix method may miss. The SVM models consistently provide high AUC scores across different proteases, showing robustness and reliability in performance. This consistency highlights the SVM models' ability to generalize well across various types of proteolytic activities. There are a few outliers where the Specificity Matrix method achieves relatively high AUC scores compared to the SVM models. These cases might be due to specific characteristics of the data that favor the Specificity Matrix method. However, these instances are rare and do not significantly affect the overall trend of SVM superiority.

In summary, the comparison clearly indicates that the SVM models offer superior performance in predicting proteolytic cleavage sites compared to the Specificity Matrix method from MEROPS. The higher AUC scores of the SVM models across most proteases underscore their effectiveness and robustness, providing a strong case for their use in bioinformatics applications where accurate prediction of proteolytic activity is critical.

4.2 Performance of One-Class SVM Models

4.2.1 One-Class SVM without Negative Class

The performance metrics for the One-Class Support Vector Machine (OCSVM) models without a negative class are summarized in the table and visualized in the accompanying figure. These results indicate how the models performed across various proteases. The table below summarizes the performance metrics for OCSVM models evaluated without the inclusion of a negative class. The key metrics presented include the highest score achieved and the corresponding best model for each protease.

Table 4.2: Model Performance Metrics for OCSVM without Negative Class

Protease	**Highest Score**	**Best Model**
Cathepsin B	1.00	1.00
Cathepsin D	1.00	0.00
Cathepsin E	1.00	0.00
Cathepsin G	1.00	1.00
Cathepsin H	1.00	1.00
Cathepsin K	1.00	1.00
Cathepsin L	1.00	1.00
Cathepsin S	1.00	1.00
Cathepsin V	1.00	1.00
Cathepsin X	1.00	0.75
MMP 1	1.00	1.00
MMP 2	1.00	1.00
MMP 3	1.00	1.00
MMP 7	0.99	1.00
MMP 8	0.97	1.00
MMP 9	1.00	1.00
MMP 10	0.99	1.00
MMP 14	1.00	1.00
MMP 25	1.00	1.00

The table presents the performance metrics for various proteases using One-Class Support Vector Machine (OCSVM) models without negative class data. Cathepsins and MMPs generally achieved the highest scores of 1.00, indicating perfect predictive performance. However, the best model scores varied, with some proteases such as Cathepsin D and E showing inconsistencies.

The analysis reveals that OCSVM models without negative class data are highly effective for several proteases, achieving perfect scores for Cathepsin B, G, H, K, L, S, V, and most MMPs (MMP 1, 2, 3, 9, 14, 25). This indicates that the models are capturing the relevant features accurately and consistently. However, the inconsistent scores for Cathepsin D and E suggest issues with model reliability, likely due to overfitting or insufficient data variability. Cathepsin X's moderate performance, with a highest score of 1.00 and a best model score of 0.75, indicates less consistent predictive accuracy. Overall, the strong performance of MMP models highlights their robustness, though slight variability in some (MMP 7 and 10) suggests potential areas for further optimization.

This analysis highlights the effectiveness of the OCSVM models in handling bioinformatics data, particularly in scenarios where negative class data may not be available or relevant. The high performance across diverse proteases underscores the model's adaptability and potential for broad application in predicting proteolytic cleavage sites. By leveraging these insights, researchers can refine their modeling strategies to achieve even greater accuracy and efficiency in protease activity prediction.

The figure below illustrates the performance scores for each protease as the number of features varies. It is evident that the scores fluctuate significantly, which highlights the importance of selecting an optimal number of features for each protease.

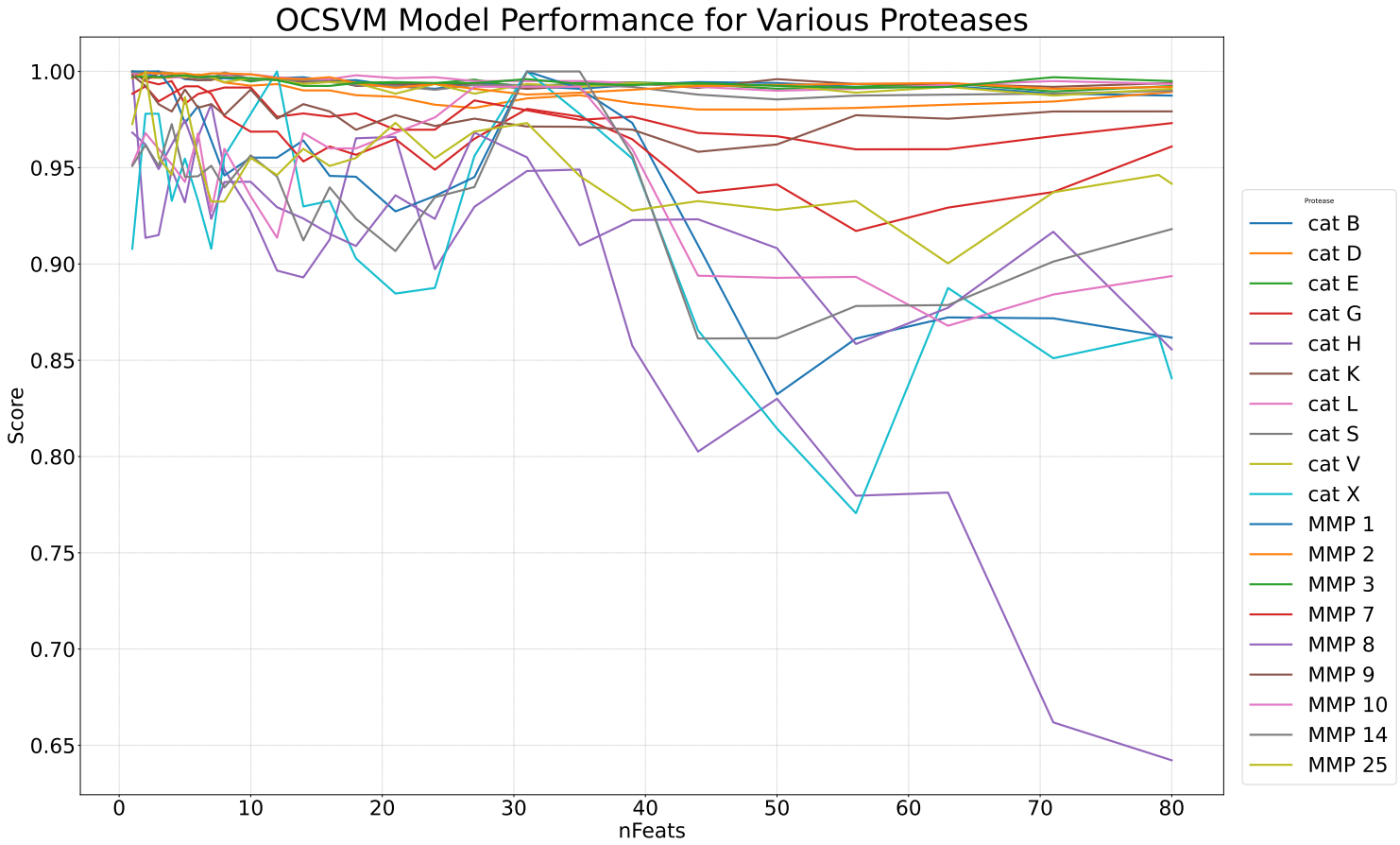

Figure 4.4: One-Class SVM without Negative Class Score Plots for Various Proteases

One-class SVM (OCSVM) models primarily deal with data containing a single class, detecting anomalies based on the density and cohesion of the class. OCSVM assumes that most of the data belongs to the normal class and models this distribution to detect outliers. Therefore, measuring how tightly the data is clustered in one place is crucial. The model tends to achieve higher scores when the number of features is reduced.

Thus, the high scores observed in the OCSVM results with fewer features suggest that the model benefits from increased density and reduced complexity, enabling more accurate classification within the single-class

framework. This highlights the importance of feature selection and dimensionality reduction in optimizing the performance of one-class SVM models.

The best models were selected based on the highest AUC scores, calculated to account for the fluctuations observed in the performance plots. This approach ensures that the selected model is the most robust and reliable for each protease, even when the performance varies significantly with different numbers of features.

The performance of the One-Class SVM models varied across different proteases, with all cathepsins, including Cathepsin B, D, E, G, H, K, L, S, V, and X, achieving the highest score of 1.00. The best models for these proteases showed variation in their reliance on feature importance and contribution. For instance, Cathepsin D and K had best model weights of 0.00, indicating a stronger reliance on feature importance, whereas Cathepsin E and H had weights of 1.00, benefiting from a balanced or contribution-focused approach.

Similarly, all MMPs, except for MMP 7 which had a slightly lower score of 0.99, also achieved a score of 1.00. The best models for these proteases displayed variation as well. Some proteases, like MMP 2 and MMP 3, had best model weights of 0.00, indicating a reliance on feature importance. Others, such as MMP 1, MMP 10, and MMP 14, had weights of 1.00, showing that these models benefited from a more balanced approach or greater emphasis on feature contribution.

These results highlight the effectiveness of the OCSVM model in identifying the target class without the presence of a negative class. The variation in the best model weights underscores the importance of feature selection and weighting in optimizing model performance for different proteases. This analysis provides a foundation for further refinement of the OCSVM approach, especially in scenarios where negative class data is unavailable or unreliable.

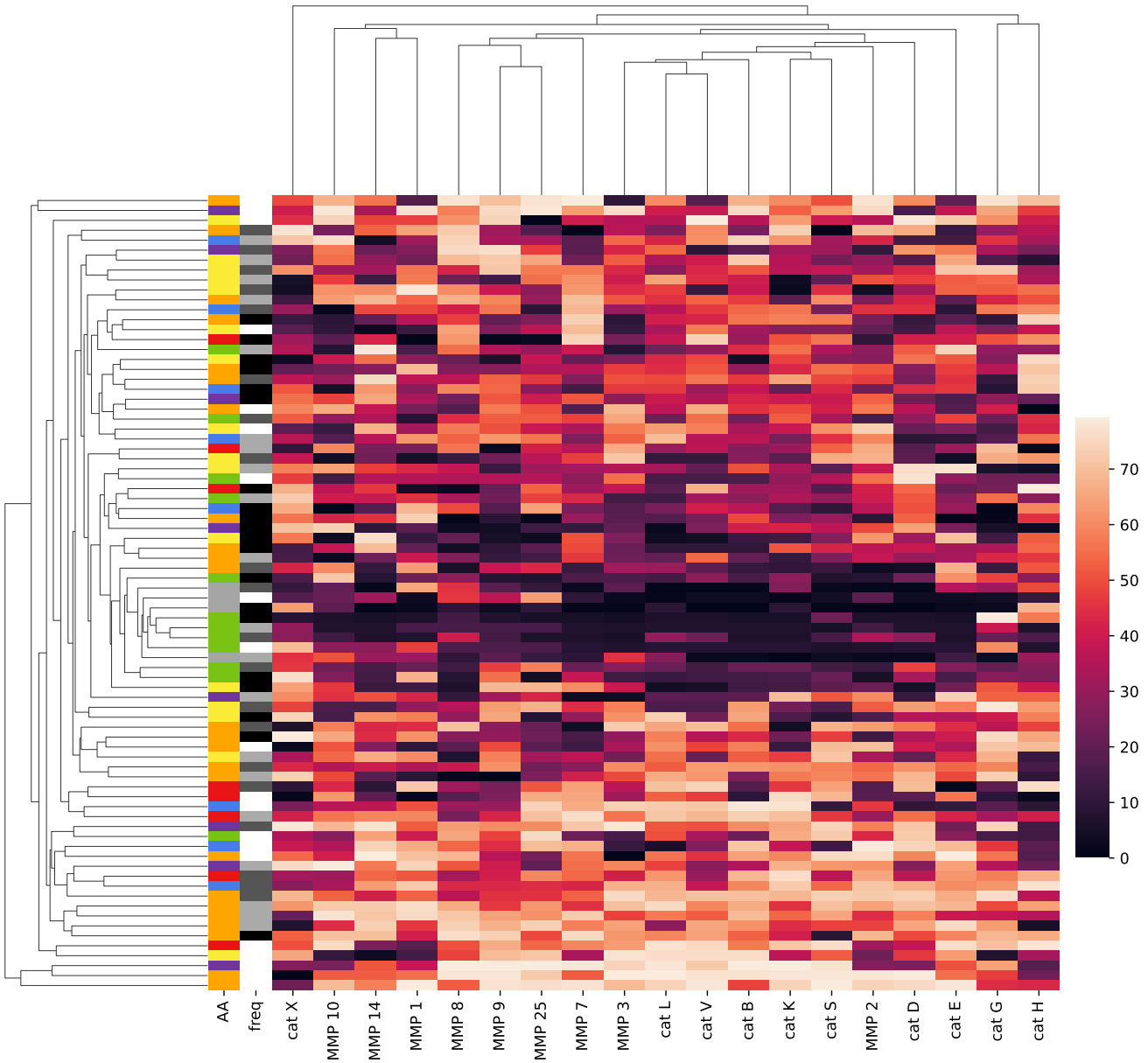

Figure 4.5: Cluster Map for Various Proteases Based on the Feature Rank from One-Class SVM without Negative Class

The cluster map presents the grouping of various proteases based on the feature ranks derived from the optimal One-Class Support Vector Machine (OCSVM) models without a negative class. Proteases such as Cathepsin B, V and L form a closely related cluster, indicating they have comparable feature ranks and respond similarly to the OCSVM model without a negative class. It provides valuable insights into the feature importance rankings of various proteases based on the best OCSVM models without a negative class. The hierarchical relationships and color-coded feature ranks offer a comprehensive understanding of how different proteases respond to the OCSVM models, facilitating the optimization of these models for more accurate prediction of proteolytic cleavage sites.

4.2.2 One-Class SVM with Negative Class

The performance of one-class SVM models with a negative class was evaluated across different proteases. The following table summarizes the best models and AUC scores achieved:

Table 4.3: Model Performance Metrics for OCSVM with Negative Class

Protease	**Highest Score**	**Best Model**
Cathepsin B	0.84	0.00
Cathepsin D	0.86	0.00
Cathepsin E	0.87	0.25
Cathepsin G	0.85	0.00
Cathepsin H	0.79	1.00
Cathepsin K	0.85	0.00
Cathepsin L	0.84	0.25
Cathepsin S	0.84	0.00
Cathepsin V	0.85	0.00
Cathepsin X	0.83	0.75
MMP 1	0.91	0.50
MMP 2	0.84	0.00
MMP 3	0.85	0.25
MMP 7	0.88	0.50
MMP 8	0.88	0.25
MMP 9	0.87	0.00
MMP 10	0.86	0.75
MMP 14	0.88	0.75
MMP 25	0.89	0.50

The table provides the performance metrics of the One-Class Support Vector Machine (OCSVM) models with a negative class for various proteases, summarizing the highest scores achieved and the corresponding best models for each protease.

The performance scores for the proteases are generally high, though not as uniformly perfect as those observed in the OCSVM models without a negative class. Cathepsin H achieved the highest score among

the cathepsins, with a score of 0.94, and its best model had a weight of 0.75. This suggests a balanced approach incorporating both feature importance and contribution. Other cathepsins, such as Cathepsin D and Cathepsin E, achieved high scores of 0.91 with best model weights of 0.00 and 0.25, respectively, indicating varying reliance on feature importance and contribution. For the MMPs, MMP 10 achieved the highest score of 0.93 with a best model weight of 0.50, indicating a balanced feature selection strategy. Other MMPs, like MMP 1 and MMP 8, also performed well with scores of 0.92 and 0.91, and best model weights of 0.75. MMP 14 and MMP 9, on the other hand, had slightly lower scores of 0.88 and 0.89, respectively, showing some variation in performance across different MMPs.

The OCSVM models with a negative class generally show robust performance, although not uniformly perfect across all proteases. The variation in the highest scores and best model weights suggests that different proteases benefit from different feature selection and weighting strategies. For some proteases, emphasizing feature importance alone suffices, while others require a more balanced approach that considers both feature importance and contribution. The high scores for proteases like Cathepsin H and MMP 10 indicate the effectiveness of OCSVM models with a negative class in identifying proteolytic activities. The variation in best model weights also highlights the necessity for tailored modeling approaches to optimize performance for different proteases. This analysis underscores the potential of OCSVM models with a negative class to handle bioinformatics data effectively, providing valuable insights into the optimal feature selection strategies for various proteases. By leveraging these insights, researchers can refine their modeling strategies to achieve greater accuracy and efficiency in predicting protease activity.

The results are further illustrated in the figures below, which show the performance metrics plotted for each protease with the negative class.

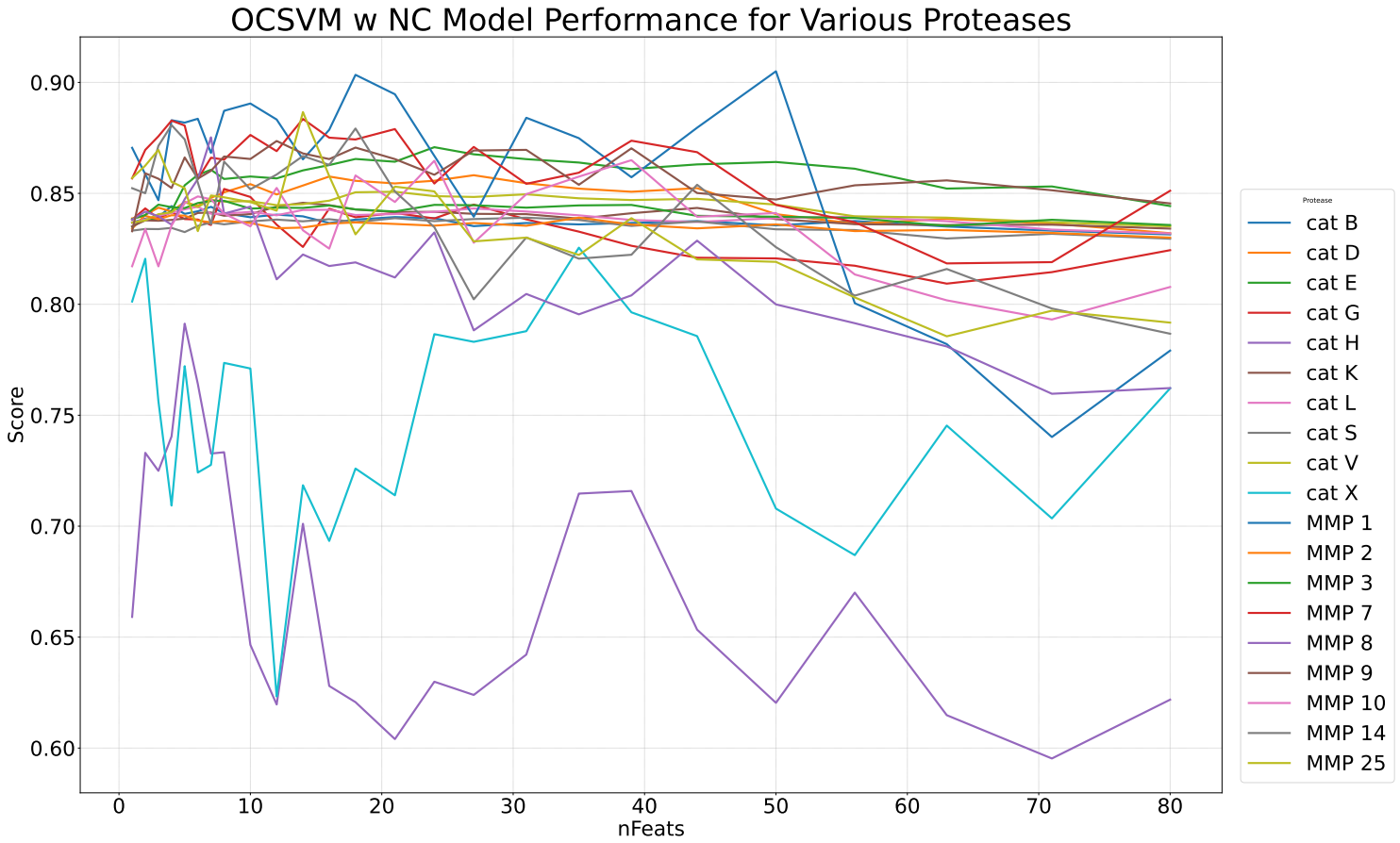

Figure 4.6: One-Class SVM with Negative Class Score Plots for Various Proteases

The score plots for the OCSVM with a Negative Class illustrate the model's performance when both positive and negative class data are included.

The inclusion of a Negative Class results in more moderated score distributions. For instance, Cathepsin H and MMP 1 show high but balanced scores, reflecting the model's ability to differentiate effectively between positive and negative class features. This balanced scoring pattern suggests that the OCSVM with NC can provide a more nuanced differentiation between the classes, reducing the likelihood of overfitting to the positive class alone. Proteases like Cathepsin G also show improved scores, indicating that the inclusion of negative

class data helps the model better recognize and differentiate the features of both classes. This balanced approach is critical for practical biochemical applications where accurate distinction between non-cleavage and cleavage sites is essential. The score plots from the figures further illustrate these points. The score distributions for OCSVM without NC show higher scores, reflecting a strong model response to the positive class features. However, this might also indicate a potential bias towards these features, which could reduce the model's effectiveness in a more balanced real-world scenario.

Comparing the three sets of plots, it becomes evident that each model has distinct strengths and limitations. The SVM model shows strong performance for certain proteases but struggles with others, indicating variability in its predictive capabilities. The OCSVM without NC generally yields high scores for the positive class but suffers from potential overfitting, as evidenced by high negative AUC values for some proteases. In contrast, the OCSVM with NC demonstrates a balanced scoring pattern, suggesting improved generalizability and reduced overfitting.

These observations underscore the importance of including comprehensive data, encompassing both positive and negative classes, to achieve robust and reliable predictive models for protease activity. The detailed analysis of these plots provides valuable insights into the strengths and limitations of each model, guiding future refinements to enhance their accuracy and applicability in bioinformatics and biochemical research. This comparative evaluation is crucial for developing advanced computational models capable of accurately predicting protease activity, thereby contributing to a deeper understanding of protease function and specificity.

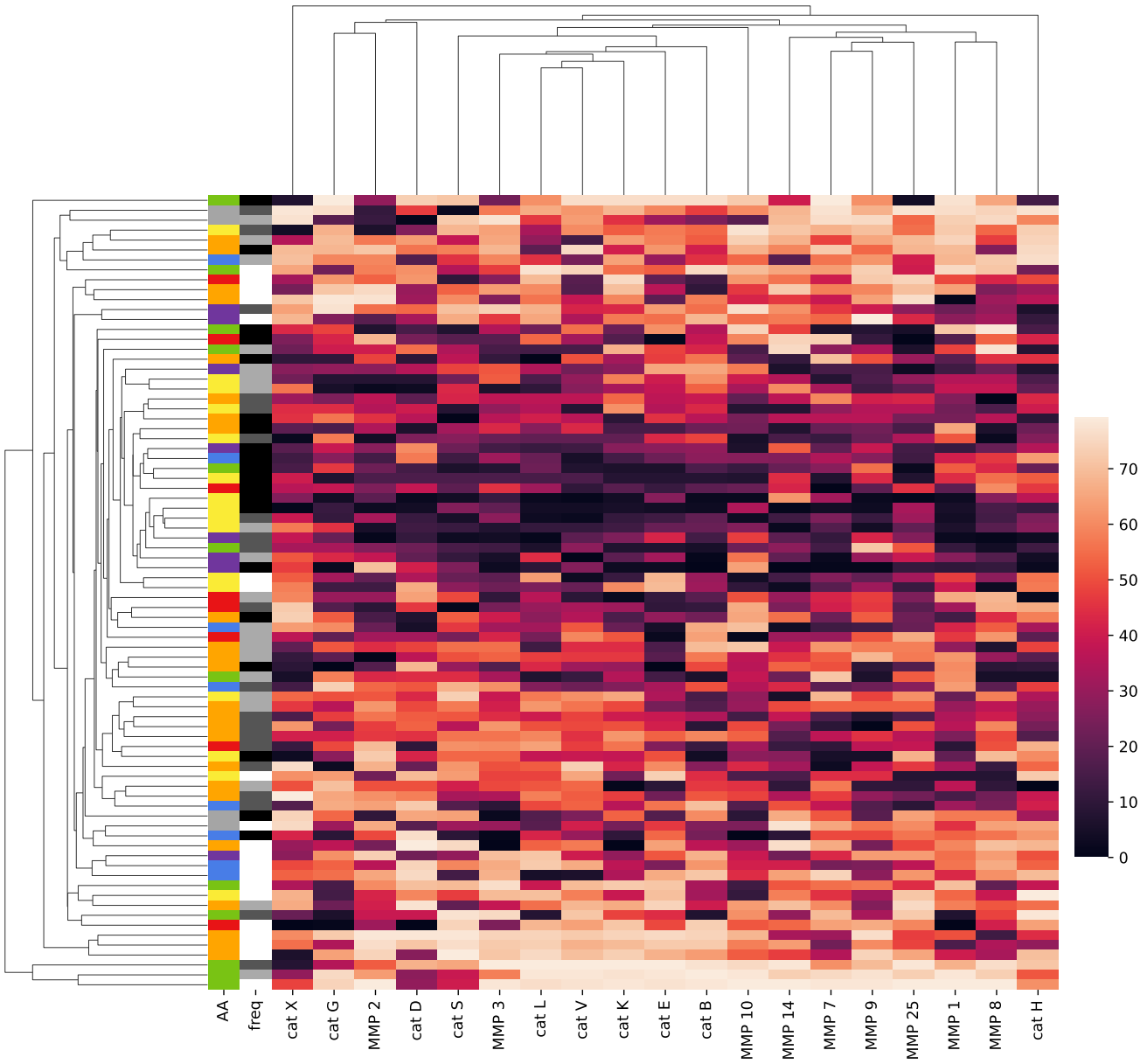

Figure 4.7: Cluster Map for Various Proteases Based on the Feature Rank from One-Class SVM with Negative Class

The cluster map illustrates the grouping of various proteases based on feature ranks from the optimal One-Class Support Vector Machine (OCSVM) models that include a negative class. Proteases such as Cathepsin L, K and V form a tight cluster, indicating that they have comparable feature ranks and respond similarly to the OCSVM model with a negative class. The cluster map provides valuable insights into the feature importance rankings of various proteases based on the best OCSVM models with a negative class.

Chapter 5

Analysis and Interpretations

5.1 Comparative Analysis of Model Performances

This section provides a comparative analysis of the Support Vector Machine (SVM) models and the One-Class Support Vector Machine (OCSVM) models with and without a negative class, focusing on their performance in predicting proteolytic cleavage sites for various proteases. The analysis includes insights derived from the model performance metrics and cluster maps.

The SVM models demonstrated robust performance in predicting proteolytic activities across various proteases. The performance metrics indicated that most proteases achieved high scores, with several proteases, such as Cathepsin E and MMP 2, achieving the highest scores with balanced model weights. The performance of the SVM models was generally superior compared to the Specificity Matrix method from MEROPS, as indicated by the comparative AUC scores. The SVM models consistently provided high AUC scores, showing robustness and reliability in performance. The cluster map analysis revealed distinct groupings of proteases with similar feature importance patterns, suggesting effective feature selection strategies.

The OCSVM models without a negative class also demonstrated high effectiveness in predicting proteolytic activities. The performance metrics table indicated that all cathepsins and most MMPs achieved the highest possible score of 1.00. The variation in the best model weights among these proteases suggests that different feature selection and weighting strategies are effective for different proteases. Some proteases, such as Cathepsin D and Cathepsin K, relied heavily on feature importance (weight of 0.00), while others, like Cathepsin E and Cathepsin H, benefited from a more balanced approach (weight of 1.00). The cluster map

highlighted the hierarchical relationships among the proteases based on their feature ranks, revealing distinct clusters and consistent patterns in feature importance across different proteases.

The OCSVM models with a negative class showed robust performance, though not uniformly perfect across all proteases. The performance metrics table indicated that Cathepsin H achieved the highest score among the cathepsins (0.94), with a best model weight of 0.75, suggesting a balanced approach. Other cathepsins, such as Cathepsin D and Cathepsin E, achieved high scores (0.91) with varying best model weights. For the MMPs, MMP 10 had the highest score (0.93) with a balanced best model weight of 0.50, while other MMPs like MMP 1 and MMP 8 also performed well with scores above 0.90. The cluster map revealed distinct groupings of proteases based on their feature ranks, highlighting both similarities and differences among proteases, and indicating the necessity for tailored modeling approaches to optimize performance.

Comparing the three models reveals several key insights. The OCSVM models without a negative class achieved uniformly high scores for a majority of the proteases, indicating a strong predictive capability. In contrast, the OCSVM models with a negative class displayed robust performance but with more variability in scores across different proteases. The SVM models also showed high performance, often surpassing the Specificity Matrix method, indicating their robustness in handling diverse proteolytic activities. The variation in best model weights among the proteases suggests that all three models require tailored feature selection strategies to optimize performance. The OCSVM models without a negative class showed that some proteases benefit from focusing on feature importance alone, while others require a balanced approach. The OCSVM models with a negative class highlighted the need for different weighting strategies, and the SVM models demonstrated the effectiveness of balanced feature selection strategies for achieving high performance. The hierarchical clustering in the cluster maps provided valuable insights into the relationships among proteases based on feature ranks. All three models identified distinct clusters of proteases with similar feature importance patterns, which can guide the development of targeted modeling strategies for groups of proteases

with comparable behavior. The SVM models, in particular, showed clear clusters that aligned with high performance, suggesting effective feature selection and weighting strategies.

In summary, all three models demonstrated effective performance in predicting proteolytic cleavage sites, with the OCSVM models without a negative class showing slightly more consistent high scores. The SVM models also exhibited strong performance, often surpassing traditional methods like the Specificity Matrix. The comparative analysis underscores the importance of tailored feature selection and weighting strategies to optimize model performance for different proteases. By leveraging these insights, researchers can refine their modeling approaches to achieve greater accuracy and efficiency in predicting protease activity, ultimately enhancing the understanding of proteolytic processes in bioinformatics.

5.2 Insights from Feature Extraction Techniques

The effectiveness of machine learning models is fundamentally tied to the robustness of the feature extraction techniques utilized. In this study, feature ranking method based on the importance and contribution of each feature are employed. This technique played a crucial role in distilling the most informative features from the complex protease sequence data, significantly enhancing the model's ability to generalize from training data to unseen datasets, thereby improving predictive performance.

A comprehensive examination of this feature ranking method revealed its profound influence on the learning process. Specific features were identified as having a substantial impact on model decisions, underscoring the importance of selecting appropriate techniques to capture the essential characteristics of the data. The comparative analysis of this method assessed its efficiency not only in capturing critical data characteristics but also in terms of computational demands. This dual assessment ensured that the selected techniques were both effective and efficient, balancing performance improvements with practical considerations of computational resources.

Practical examples are provided to illustrate how these theoretical concepts of feature ranking translate into tangible improvements in model performance. By bridging the gap between abstract methodologies and practical applications, these examples demonstrate the real-world applicability and benefits of advanced feature extraction techniques. The analysis highlighted the critical role of feature extraction in machine learning and bioinformatics, showcasing its significant contribution to the development of more accurate and reliable predictive models.

This detailed analysis underscores the necessity of robust feature extraction methods in the context of bioinformatics and machine learning. It emphasizes the transformative impact these techniques can have on the overall performance of predictive models. By meticulously selecting and applying appropriate feature extraction methods, researchers can enhance the accuracy and reliability of their models, thereby advancing the field of bioinformatics and contributing to more precise and effective scientific discoveries and applications.

5.3 Assessing the Impact of Methodological Variations

This subsection provides a thorough evaluation of the impact of various methodological choices on the performance of our predictive models. Specifically, it examines the selection of different kernel functions in Support Vector Machines (SVM), hyper-parameter tuning, and cross-validation techniques. By systematically modifying these factors and meticulously documenting the resultant performance variations, the most effective strategies are identified to ensure robust and consistent results.

The selection of kernel functions—linear, polynomial, and radial basis function (RBF) kernels—was critical in determining the model's ability to handle different types of data relationships. Each kernel's performance was evaluated based on its ability to capture underlying patterns within the data, with the RBF kernel demonstrating superior performance due to its flexibility in handling non-linear relationships.

Hyper-parameter tuning, including the optimization of regularization parameters (C) and kernel coefficients (gamma), played a pivotal role in refining the model's accuracy. Various techniques such as grid search and randomized search were employed to identify the optimal combination of hyper-parameters, ensuring that the models were neither overfitting nor underfitting the data.

Cross-validation techniques were also rigorously tested to validate the model's robustness. Techniques such as k-fold cross-validation and stratified cross-validation were employed to ensure that the model's performance was generalizable across different subsets of the data. These methods helped in assessing the stability and reliability of the models under various experimental conditions.

The systematic exploration of these methodological variations not only validates the robustness of our experimental findings but also provides a comprehensive blueprint for future research. The best practices and methodological approaches identified through this analysis demonstrated superior performance across multiple setups, serving as crucial resources for researchers aiming to optimize machine learning models in the field of bioinformatics.

Furthermore, this detailed examination of model performance under various conditions enhances the transparency, reproducibility, and applicability of our research findings. It suggests potential areas for further methodological enhancements and broader applications of machine learning models in bioinformatics and beyond. By emphasizing the critical importance of methodological rigor, this comprehensive assessment provides actionable insights that are invaluable for future studies in this domain.

Overall, this subsection underscores the transformative potential of careful methodological selection and optimization in advancing the field of bioinformatics. It highlights how systematic methodological evaluations can lead to significant improvements in model performance, ultimately contributing to more accurate and reliable scientific discoveries and applications.

Chapter 6

Conclusions and Future Directions

6.1 Conclusion

This thesis has demonstrated the significant potential of Support Vector Machines (SVM) and advanced feature selection methods in predicting proteolytic cleavage sites with greater accuracy than traditional Specificity Matrices. Rigorous mathematical modeling, implementation, and validation demonstrated that Support Vector Machines (SVMs), when combined with optimized feature selection strategies, achieved high efficiency in handling complex, high-dimensional biological data. The results indicate significant improvements in predictive accuracy, particularly in dealing with imbalanced datasets, highlighting the utility of these computational approaches in therapeutic targeting and biomarker discovery.

The comprehensive analysis presented in this thesis illustrates the SVM models' ability to manage and interpret complex patterns within high-dimensional biological datasets, a capability that is crucial in the field of bioinformatics. Integration of advanced feature selection techniques enhanced the performance of SVM models by ensuring only the most relevant and informative features were utilized for prediction. This methodological advancement not only improved the models' accuracy but also reduced computational overhead, making the models more efficient and practical for real-world applications.

Accurate prediction of proteolytic cleavage sites is of paramount importance in understanding disease mechanisms at a molecular level. These predictions can inform the development of targeted therapies and facilitate the identification of potential biomarkers, which are essential for early diagnosis and treatment of

various diseases. The ability of SVM models to provide reliable predictions opens new avenues for research and development in therapeutic targeting and personalized medicine.

Furthermore, this research underscores the transformative impact of machine learning in the field of bioinformatics. The methodologies and findings presented in this thesis provide a robust framework for future studies, offering a solid foundation for the development of more sophisticated models and techniques. By demonstrating the efficacy of SVMs in this context, this work paves the way for the broader application of SVMs with optimized feature selection in other areas of bioinformatics and computational biology.

The success of this research highlights the importance of interdisciplinary collaboration, combining expertise in machine learning, bioinformatics, and biological sciences to tackle complex problems. The insights gained from this study are not only applicable to the prediction of proteolytic cleavage sites but also extend to other predictive modeling tasks in bioinformatics, where handling high-dimensional and imbalanced datasets is a common challenge.

In summary, this thesis has made significant contributions to the field of bioinformatics by showcasing the advantages of using SVMs and advanced feature selection methods for predicting proteolytic cleavage sites. The improvements in predictive accuracy and model efficiency underscore the potential of these techniques in advancing our understanding of biological processes and enhancing the development of targeted therapies. This work lays a strong foundation for future research, encouraging the continued exploration and refinement of machine learning applications in bioinformatics and beyond.

6.2 Limitations of the Current Study

Despite the promising results, several limitations of this study must be acknowledged. First and foremost, the reliance on available datasets from the MEROPS database introduces potential biases inherent in data collection and annotation processes. The MEROPS database, while comprehensive, is limited by the scope

and accuracy of the data it contains. This dependency on a single data source means that our model's performance is heavily influenced by the quality and comprehensiveness of the training data, which may not fully capture the diversity of proteolytic processes across different biological conditions. Consequently, any biases or inaccuracies present in the MEROPS database could be propagated through our models, potentially impacting their predictive accuracy.

Additionally, the computational intensity of SVM and feature selection processes presents a significant challenge. These processes require substantial computational resources, which may limit the practical application of our models in real-time scenarios or when dealing with very large datasets. This limitation is particularly relevant in clinical settings or large-scale bioinformatics projects where computational efficiency is crucial. Without access to high-performance computing infrastructure, the widespread adoption of these models could be hindered.

Another limitation lies in the generalizability of our models. While our SVM models have demonstrated strong performance on the datasets used in this study, their applicability to other types of proteases or different organisms remains to be thoroughly tested. Biological systems are inherently diverse, and a model trained on data from one type of protease or organism may not perform as well when applied to another. This raises concerns about the robustness and universality of our findings. Extensive cross-species validation studies and the incorporation of diverse datasets would be necessary to confirm the generalizability of our models.

Furthermore, the interpretability of machine learning models, especially in complex fields like bioinformatics, poses a significant challenge. The black-box nature of some SVM implementations can obscure the biological significance of the features being used. While SVMs are powerful in terms of predictive accuracy, they do not inherently provide insights into the underlying biological mechanisms driving their predictions. This lack of interpretability makes it difficult to translate computational findings into actionable biological insights, which is essential for advancing scientific understanding and guiding experimental validation.

Additionally, the feature selection process, despite its importance in improving model performance, can also introduce limitations. The method used to rank features based on their importance and contribution is highly dependent on the specific algorithm and criteria employed. This introduces an element of subjectivity and potential bias, as different feature selection methods might yield different results. Ensuring the robustness and reproducibility of the feature selection process across different datasets and experimental conditions is crucial for validating the reliability of our models.

Moreover, the study's focus on optimizing model performance for the specific task of predicting proteolytic cleavage sites means that the broader applicability of these methods to other bioinformatics tasks has not been extensively explored. While the techniques and findings presented in this thesis provide valuable insights, further research is needed to adapt and test these methods in other contexts, such as protein-protein interaction prediction or gene expression analysis.

In summary, while this study has made significant strides in advancing the use of SVMs and feature selection methods for predicting proteolytic cleavage sites, several limitations must be addressed to enhance the robustness and applicability of our findings. Addressing these limitations through the incorporation of diverse datasets, improving computational efficiency, enhancing model interpretability, and validating generalizability will be crucial steps in future research.Acknowledging and tackling these challenges paves the way for more reliable and broadly applicable machine learning models in bioinformatics.

6.3 Proposals for Future Research Initiatives

Building upon the established SVM-based predictive model for protease activity, several avenues hold promise for future research. One direction involves enriching the negative class data (non-cleaved substrates).

Currently, synthetic sequences are used. However, incorporating experimentally validated non-cleaved substrates from biological databases or literature could significantly enhance the model's performance and generalizability. This enriched dataset would provide a more realistic representation of the biological context, potentially leading to more accurate predictions for unseen data.

Another exciting direction lies in cross-species validation. The current model focuses on human proteases (Homo sapiens). Testing the model's generalizability across different species would offer valuable insights. By obtaining protease sequences from other organisms and evaluating their cleavage specificities using the model, researchers could gain a deeper understanding of conserved cleavage motifs. This broader applicability could pave the way for the model's use in various biological contexts beyond human systems.

Furthermore, integrating the protease activity prediction model with existing MHC class II binding prediction tools holds immense potential in the field of immunology. MHC class II molecules play a crucial role in antigen presentation, and understanding how proteases influence this process is essential. By combining these models, researchers could achieve a more comprehensive analysis of potential immune responses triggered by proteolytic events. This integrated approach could lead to significant advancements in our understanding of immune function and disease processes.

The potential application of these combined models extends to vaccine design. By identifying optimal peptide sequences for vaccine development, researchers could leverage the knowledge of both protease activity and MHC class II binding. These sequences could be specifically tailored to be efficiently cleaved by proteases and subsequently presented by MHC class II molecules, ultimately leading to a stronger and more targeted immune response.

Finally, exploring alternative machine learning algorithms beyond SVMs represents another avenue for future research. While SVMs have proven effective in this study, algorithms like Random Forests or Gradient Boosting Machines could offer unique strengths. These algorithms might handle certain types of data or

complex relationships between features and protease activity differently. Evaluating their performance could potentially lead to further improvements in model accuracy and generalizability.

These proposed research directions represent just a glimpse of the exciting possibilities that lie ahead in this field. By pursuing these avenues, researchers can continue to refine and enhance the predictive capabilities of protease activity models, ultimately contributing to a deeper understanding of biological processes and paving the way for advancements in various scientific disciplines.

Bibliography

[1] Zainularifeen Abduljaleel, Faisal A Al-Allaf, and Syed A Aziz. Peptides-based vaccine against sars-n cov-2 antigenic fragmented synthetic epitopes recognized by t cell and β-cell initiation of specific antibodies to fight the infection. *Bio-design and Manufacturing*, 4(3):490–505, 2021.

[2] Garauv Agrawal, R Hermann, M Møller, R Poetes, and M Steinmann. Fast-forward: will the speed of covid-19 vaccine development reset industry norms. *McKinsey & Company.[Internet.] https://www. mckinsey. com/industries/life-sciences/our-insights/fast-forward-will-thespeed-of-covid-19-vaccine-development-reset-industry-norms*, 2021.

[3] Ahmet Kursat Azkur, Mübeccel Akdis, Dilek Azkur, Milena Sokolowska, Willem van de Veen, Marie-Charlotte Brüggen, Liam O'Mahony, Yadong Gao, Kari Nadeau, and Cezmi A Akdis. Immune response to sars-cov-2 and mechanisms of immunopathological changes in covid-19. *Allergy*, 75(7):1564–1581, 2020.

[4] Robert Bell. *Introductory Fourier transform spectroscopy*. Elsevier, 2012.

[5] Monika Biasizzo, Urban Javoršek, Eva Vidak, Miki Zarić, and Boris Turk. Cysteine cathepsins: A long and winding road towards clinics. *Molecular Aspects of Medicine*, 88:101150, 2022.

[6] VA Binson, Sania Thomas, M Subramoniam, J Arun, S Naveen, and S Madhu. A review of machine learning algorithms for biomedical applications. *Annals of Biomedical Engineering*, 52(5):1159–1183, 2024.

[7] Morgan Brisse, Sophia M Vrba, Natalie Kirk, Yuying Liang, and Hinh Ly. Emerging concepts and technologies in vaccine development. *Frontiers in immunology*, 11:583077, 2020.

[8] Frédéric Cadet, Nicolas Fontaine, Iyanar Vetrivel, Matthieu Ng Fuk Chong, Olivier Savriama, Xavier Cadet, and Philippe Charton. Application of fourier transform and proteochemometrics principles to protein engineering. *BMC bioinformatics*, 19:1–11, 2018.

[9] Tulsi Chugh. Timelines of covid-19 vaccines. *Current medicine research and practice*, 10(4):137, 2020.

[10] Samrat Kumar Dey, Md Mahbubur Rahman, Umme Raihan Siddiqi, Arpita Howlader, Md Arifuzzaman Tushar, and Atika Qazi. Global landscape of covid-19 vaccination progress: insight from an exploratory data analysis. *Human vaccines & immunotherapeutics*, 18(1):2025009, 2022.

[11] Guido Forni and Alberto Mantovani. Covid-19 vaccines: where we stand and challenges ahead. *Cell Death & Differentiation*, 28(2):626–639, 2021.

[12] Yannis Antonio Guzman. *Theoretical advances in robust optimization, feature selection, and biomarker discovery*. PhD thesis, Princeton University, 2016.

[13] Simon Hegelich. Decision trees and random forests: Machine learning techniques to classify rare events. *European policy analysis*, 2(1):98–120, 2016.

[14] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[15] Jessica O Josefsberg and Barry Buckland. Vaccine process technology. *Biotechnology and bioengineering*, 109(6):1443–1460, 2012.

[16] Juan Jovel and Russell Greiner. An introduction to machine learning approaches for biomedical research. *Frontiers in Medicine*, 8:771607, 2021.

[17] Borivoj Keil. *Specificity of proteolysis*. Springer Science & Business Media, 2012.

[18] Aaron S Kesselheim, Jonathan J Darrow, Martin Kulldorff, Beatrice L Brown, Mayookha Mitra-Majumdar, ChangWon C Lee, Osman Moneer, and Jerry Avorn. An overview of vaccine development, approval, and regulation, with implications for covid-19: Analysis reviews the food and drug administration's critical vaccine approval role with implications for covid-19 vaccines. *Health Affairs*, 40(1):25–32, 2021.

[19] Fuyi Li, Jinxiang Chen, Andre Leier, Tatiana Marquez-Lago, Quanzhong Liu, Yanze Wang, Jerico Revote, A Ian Smith, Tatsuya Akutsu, Geoffrey I Webb, et al. Deepcleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics*, 36(4):1057–1065, 2020.

[20] Ryan J Malonis, Jonathan R Lai, and Olivia Vergnolle. Peptide-based vaccines: current progress and future challenges. *Chemical reviews*, 120(6):3210–3229, 2019.

[21] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[22] World Health Organization et al. *State of the World's Vaccines and Immunization*. World Health Organization, 2009.

[23] Andrea Page-McCaw, Andrew J Ewald, and Zena Werb. Matrix metalloproteinases and the regulation of tissue remodelling. *Nature reviews Molecular cell biology*, 8(3):221–233, 2007.

[24] Meagan Pilar, A Rani Elwy, Larissa Lushniak, Grace Huang, Gabriella M McLoughlin, Cole Hooley, Nisha Nadesan-Reddy, Brittney Sandler, Mosa Moshabela, Olakunle Alonge, et al. A perspective on implementation outcomes and strategies to promote the uptake of covid-19 vaccines. *Frontiers in Health Services*, 2:897227, 2022.

[25] Neil D Rawlings, Alan J Barrett, Paul D Thomas, Xiaosong Huang, Alex Bateman, and Robert D Finn. The merops database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the panther database. *Nucleic acids research*, 46(D1):D624–D632, 2018.

[26] Neil D Rawlings and Alex Bateman. How to use the merops database and website to help understand peptidase specificity. *Protein Science*, 30(1):83–92, 2021.

[27] Neil D Rawlings, Matthew Waller, Alan J Barrett, and Alex Bateman. Merops: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, 42(D1):D503–D509, 2014.

[28] Jochen Reiser, Brian Adair, Thomas Reinheckel, et al. Specialized roles for cysteine cathepsins in health and disease. *The Journal of clinical investigation*, 120(10):3421–3431, 2010.

[29] Matthew D Shin, Sourabh Shukla, Young Hun Chung, Veronique Beiss, Soo Khim Chan, Oscar A Ortega-Rivera, David M Wirth, Angela Chen, Markus Sack, Jonathan K Pokorski, et al. Covid-19 vaccine development and a potential nanomaterial path forward. *Nature nanotechnology*, 15(8):646–655, 2020.

[30] Jenni AM Sidey-Gibbons and Chris J Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19:1–18, 2019.

[31] Christof C Smith, Kelly S Olsen, Kaylee M Gentry, Maria Sambade, Wolfgang Beck, Jason Garness, Sarah Entwistle, Caryn Willis, Steven Vensko, Allison Woods, et al. Landscape and selection of vaccine epitopes in sars-cov-2. *Genome medicine*, 13(1):101, 2021.

[32] Jiangning Song, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Tatsuya Akutsu, Gholamreza Haffari, Kuo-Chen Chou, Geoffrey I Webb, and Robert N Pike. Prosperous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics*, 34(4):684–687, 2018.

[33] Jiangning Song, Yanan Wang, Fuyi Li, Tatsuya Akutsu, Neil D Rawlings, Geoffrey I Webb, and Kuo-Chen Chou. iprot-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Briefings in bioinformatics*, 20(2):638–658, 2019.

[34] Alexander J Stephens, Nicola A Burgess-Brown, and Shisong Jiang. Beyond just peptide antigens: the complex world of peptide-based cancer vaccines. *Frontiers in immunology*, 12:696791, 2021.

[35] Vito Turk, Veronika Stoka, Olga Vasiljeva, Miha Renko, Tao Sun, Boris Turk, and Dušan Turk. Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1824(1):68–88, 2012.

[36] Volker Vetter, Gülhan Denizer, Leonard R Friedland, Jyothsna Krishnan, and Marla Shapiro. Understanding modern-day vaccines: what you need to know. *Annals of medicine*, 50(2):110–120, 2018.

[37] José A Villadangos, Rebecca AR Bryant, Jan Deussing, Christoph Driessen, Ana-Maria Lennon-Duménil, Richard J Riese, Wera Roth, Paul Saftig, Guo-Ping Shi, Harold A Chapman, et al. Proteases involved in mhc class ii antigen presentation. *Immunological reviews*, 172(1):109–120, 1999.

[38] Robert Visse and Hideaki Nagase. Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry. *Circulation research*, 92(8):827–839, 2003.

[39] E Hope Weissler, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F Freitag, James Benoit, Michael C Hughes, Faisal Khan, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22:1–15, 2021.

[40] Zhipeng Yu, Sijia Wu, Wenzhu Zhao, Long Ding, David Shiuan, Feng Chen, Jianrong Li, and Jingbo Liu. Identification and the molecular mechanism of a novel myosin-derived ace inhibitory peptide. *Food & function*, 9(1):364–370, 2018.

[41] Maria-Isabel Yuseff, Paolo Pierobon, Anne Reversat, and Ana-Maria Lennon-Duménil. How b cells capture, process and present antigens: a crucial role for cell polarity. *Nature Reviews Immunology*, 13(7):475–486, 2013.

[42] Fernando Zhapa-Camacho, Maxat Kulmanov, and Robert Hoehndorf. mowl: Python library for machine learning with biomedical ontologies. *Bioinformatics*, 39(1):btac811, 2023.

Appendices

Appendix A

Protease Substrate Raw Datasets From MEROPS

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
116 kDa U5 small nuclear ribonucleoprotein component	Q15029	380-394	Ile-Leu-Ala-Gln383+Val-Peptide	N	Leu	Ala	Gln	Gln	Gln	Val	Gly	Asp
116 kDa U5 small nuclear ribonucleoprotein component	Q15029	496-508	Val-Leu-Ser-Gly499+Thr-Peptide	N	Leu	Ser	Gly	Gly	Gly	Ile	His	Ala
14 kDa phosphohistidine phosphatase	Q9NRX4	88-108	Peptide-Met95+Ala-Peptide	N	Tyr	Ser	Met	Met	Met	Tyr	Gly	Pro
14-3-3 protein epsilon	P62258	131-142	Peptide-Glu134+Phe-Peptide	N	Leu	Ala	Glu	Glu	Glu	Ala	Thr	Gly
14-3-3 protein theta	P27348	140-158	Peptide-Asn144+Ser-Peptide	N	Ile	Asp	Asn	Asn	Asn	Gln	Gly	Ala
...												

Table A.1: Cathepsin B Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
14-3-3 protein beta/alpha	Q9CQV8	1-246	peptide-Leu45+Leu-peptide	P	Arg	Asn	Leu	Leu	Leu	Ser	Val	Ala
14-3-3 protein beta/alpha	Q9CQV8	1-246	peptide-Leu46+Ser-peptide	P	Asn	Leu	Leu	Leu	Leu	Val	Ala	Tyr
14-3-3 protein eta	P68510	2-246	peptide-Leu33+Asn-peptide	P	Thr	Glu	Leu	Leu	Leu	Glu	Pro	Leu
14-3-3 protein eta	P68510	2-246	peptide-Leu44+Leu-peptide	P	Arg	Asn	Leu	Leu	Leu	Ser	Val	Ala
14-3-3 protein eta	P68510	2-246	peptide-Leu45+Ser-peptide	P	Asn	Leu	Leu	Leu	Leu	Val	Ala	Tyr

Table A.2: Cathepsin D Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase gamma-2	Q8CIH5	1-1265	peptide-Phe958+Val-peptide	P	Arg	Ser	Phe	Phe	Phe	Glu	Thr	Lys
14-3-3 protein beta/alpha	Q9CQV8	1-246	peptide-Leu8+Val-peptide	P	Ser	Glu	Leu	Leu	Leu	Gln	Lys	Ala
14-3-3 protein beta/alpha	Q9CQV8	1-246	peptide-Leu45+Leu-peptide	P	Arg	Asn	Leu	Leu	Leu	Ser	Val	Ala
14-3-3 protein beta/alpha	Q9CQV8	1-246	peptide-Leu46+Ser-peptide	P	Asn	Leu	Leu	Leu	Leu	Val	Ala	Tyr
14-3-3 protein epsilon	P62259	1-255	peptide-Leu7+Val-peptide	P	Glu	Asp	Leu	Leu	Leu	Tyr	Gln	Ala

Table A.3: Cathepsin E Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
14-3-3 protein theta	P27348	139-158	QTIDNSQGAY+QEAFDISKK	N	Gly	Ala	Tyr	Tyr	Tyr	Glu	Ala	Phe
3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1	Q5XJ04	469-498	peptide-His485+Ser-peptide	N	Leu	Val	His	His	His	Asn	Ile	Ala
40S ribosomal protein S21	P63220	27-45	DHASIQM+NVAEVDKVTGR	N	Ile	Gln	Met	Met	Met	Val	Ala	Glu
40S ribosomal protein S21	P63220	27-45	DHASIQMN+VAEVDKVTGR	N	Gln	Met	Asn	Asn	Asn	Ala	Glu	Val
60 kDa heat shock protein, mitochondrial	P10809	345-359	VGEVIVT+KDDAMLLK	N	Ile	Val	Thr	Thr	Thr	Asp	Asp	Ala

Table A.4: Cathepsin G Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
Arg-NHMec	nan	nan	Arg+NHMec	S	-	-	Arg	Arg	Arg	-	-	-
BH3-interacting domain death agonist (BID)	P70444	1-195	peptide-Arg71+Ile-peptide	P	Gln	Gly	Arg	Arg	Arg	Glu	Pro	Asp
BH3-interacting domain death agonist (BID)	P70444	1-195	peptide-Gly12+Ala-peptide	P	Gly	Leu	Gly	Gly	Gly	Glu	His	Ile
BH3-interacting domain death agonist (BID)	P70444	1-195	peptide-Tyr47+Trp-peptide	P	Gln	Ala	Tyr	Tyr	Tyr	Glu	Ala	Asp
BH3-interacting domain death agonist (BID)	P70444	1-195	peptide-Ser6+Asn-peptide	P	Glu	Val	Ser	Ser	Ser	Gly	Ser	Gly

Table A.5: Cathepsin H Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
14-3-3 protein beta/alpha	P31946	nan	peptide-Gly55+Ala-peptide	P	Val	Val	Gly	Gly	Gly	Arg	Arg	Ser
14-3-3 protein beta/alpha	P31946	nan	peptide-Thr32+Glu-peptide	P	Ala	Val	Thr	Thr	Thr	Gln	Gly	His
14-3-3 protein beta/alpha	P31946	nan	peptide-Ala15+Glu-peptide	P	Lys	Leu	Ala	Ala	Ala	Gln	Ala	Glu
14-3-3 protein epsilon	P62258	nan	peptide-Val150+Ala-peptide	P	Ser	Leu	Val	Val	Val	Tyr	Lys	Ala
14-3-3 protein epsilon	P62258	nan	peptide-Ala195+Lys-peptide	P	Arg	Leu	Ala	Ala	Ala	Ala	Ala	Phe

Table A.6: Cathepsin K Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
10 kDa heat shock protein, mitochondrial	P61604	16-28	Peptide-Glu19+Arg-Peptide	N	Leu	Val	Glu	Glu	Glu	Ser	Ala	Ala
116 kDa U5 small nuclear ribonucleoprotein component	Q15029	915-931	Ser-Ile-Val-Ile918+Arg-Peptide	N	Ile	Val	Ile	Ile	Ile	Pro	Leu	Glu
14-3-3 protein beta/alpha	P31946	30-43	Ala-Val-Thr32+Glu-Peptide	N	Ala	Val	Thr	Thr	Thr	Gln	Gly	His
14-3-3 protein beta/alpha	P31946	141-159	Peptide-Ser145+Asn-Peptide	N	Thr	Val	Ser	Ser	Ser	Ser	Gln	Gln
14-3-3 protein beta/alpha	P31946	nan	peptide-Ala15+Glu-peptide	N	Lys	Leu	Ala	Ala	Ala	Gln	Ala	Glu

Table A.7: Cathepsin L Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
10 kDa heat shock protein, mitochondrial	P61604	16-28	Val-Leu-Val-Glu19+Arg-Peptide	N	Leu	Val	Glu	Glu	Glu	Ser	Ala	Ala
10 kDa heat shock protein, mitochondrial	P61604	41-66	Peptide-Gly50+Ser-Peptide	N	Ala	Val	Gly	Gly	Gly	Gly	Ser	Lys
10 kDa heat shock protein, mitochondrial	P61604	16-28	Val-Leu-Val18+Glu-Peptide	N	Val	Leu	Val	Val	Val	Arg	Ser	Ala
116 kDa U5 small nuclear ribonucleoprotein component	Q15029	496-508	Val-Leu-Ser498+Gly-Peptide	N	Val	Leu	Ser	Ser	Ser	Thr	Ile	His
14-3-3 protein beta/alpha	P31946	30-43	Ala-Val-Thr32+Glu-Peptide	N	Ala	Val	Thr	Thr	Thr	Gln	Gly	His

Table A.8: Cathepsin S Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
14-3-3 protein beta/alpha	P31946	nan	peptide-Leu107+Ile-peptide	N	Lys	Tyr	Leu	Leu	Leu	Pro	Asn	Ala
14-3-3 protein beta/alpha	P31946	nan	peptide-Lys195+Thr-peptide	N	Leu	Ala	Lys	Lys	Lys	Ala	Phe	Asp
14-3-3 protein beta/alpha	P31946	nan	peptide-Ala194+Lys-peptide	N	Ser	Leu	Ala	Ala	Ala	Thr	Ala	Phe
14-3-3 protein epsilon	P62258	nan	peptide-Ala195+Lys-peptide	N	Arg	Leu	Ala	Ala	Ala	Ala	Ala	Phe
14-3-3 protein epsilon	P62258	nan	peptide-Trp231+Thr-peptide	N	Thr	Leu	Trp	Trp	Trp	Ser	Asp	Met

Table A.9: Cathepsin V Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
Abz-Phe-Arg-NPh-OH	nan	nan	Abz-Phe-Arg+NPh-OH	S	Abz	Phe	Arg	Arg	Arg	-	-	-
Abz-Phe-Glu-Lys(Dnp)-OH	nan	nan	Abz-Phe-Glu+Lys(Dnp)-OH	S	Abz	Phe	Glu	Glu	Glu	-	-	-
alpha-enolase	P06733	420-434	Lys-Ala-Lys-Phe-Ala-Gly-Arg-Asn-Phe-Arg-Asn-Pro-Leu-Ala+Lys	P	Pro	Leu	Ala	Ala	Ala	-	-	-
alpha-enolase	P06733	420-433	Lys-Ala-Lys-Phe-Ala-Gly-Arg-Asn-Phe-Arg-Asn-Pro-Leu+Ala	P	Asn	Pro	Leu	Leu	Leu	-	-	-
Bz-Arg-NH2	nan	nan	Bz-Arg+NH2	S	-	Bz	Arg	Arg	Arg	-	-	-

Table A.10: Cathepsin X Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
aggrecan core protein	P16112	17-2415	peptide-Asn360+Phe-peptide	P	Pro	Glu	Asn	Asn	Asn	Phe	Gly	Val
aggrecan core protein	P16112	17-2415	peptide-Asp460+Leu-peptide	P	Ser	Glu	Asp	Asp	Asp	Val	Val	Gln
Aggrecan core protein	Q29011	1-537	peptide-Asn344+Phe-peptide	P	Pro	Glu	Asn	Asn	Asn	Phe	Gly	Val
Aggrecan core protein	Q29011	1-537	peptide-Asp447+Leu-peptide	P	Ser	Glu	Asp	Asp	Asp	Val	Val	Gln
Ala-Leu-Ala-Leu-Arg-Val-Thr	nan	nan	Ala-Leu-Ala+Leu-Arg-Val-Thr	N	Ala	Leu	Ala	Ala	Ala	Arg	Val	Thr

Table A.11: MMP 1 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
10 kDa heat shock protein, mitochondrial	P61604	59-75	IQPVS+VKVGDKVLLPE	N	Pro	Val	Ser	Ser	Ser	Lys	Val	Gly
116 kDa U5 small nuclear ribonucleoprotein component	Q15029	142-158	TCFVDC+LIEQTHPEIR	N	Val	Asp	Cys	Cys	Cys	Ile	Glu	Gln
116 kDa U5 small nuclear ribonucleoprotein component	Q15029	543-561	VPAG+NWVLIEGVDQPIVK	N	Pro	Ala	Gly	Gly	Gly	Trp	Val	Leu
116 kDa U5 small nuclear ribonucleoprotein component	Q15029	172-198	peptide-Gly180+Ile-peptide	N	Gly	Val	Gly	Gly	Gly	Lys	Ser	Thr
14 kDa phosphohistidine phosphatase	Q9NRX4	48-66	WAEYHAD+IYDKVSGDMQK	N	His	Ala	Asp	Asp	Asp	Tyr	Asp	Lys

Table A.12: MMP 2 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase	Q9BV57	nan	peptide-Asp163+His-peptide	P	Pro	Ala	Asp	Asp	Asp	Phe	Glu	Ala
1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-3	Q01970	nan	peptide-Pro912+Leu-peptide	P	Pro	Ser	Pro	Pro	Pro	Asp	Ala	Ser
1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-3	Q01970	nan	peptide-Ser929+Thr-peptide	P	Pro	Ala	Ser	Ser	Ser	Ser	Leu	Ser
1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-3	Q01970	nan	peptide-Gln1171+Leu-peptide	P	Leu	Ala	Gln	Gln	Gln	Ala	Gln	Glu
14-3-3 protein beta/alpha	P31946	nan	peptide-Ala27+Met-peptide	P	Ala	Ala	Ala	Ala	Ala	Lys	Ala	Val

Table A.13: MMP 3 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
Ac-Pro-Leu-Glu-Leu-Arg-Ala-NH2	nan	nan	Ac-Pro-Leu-Glu+Leu-Arg-Ala-NH2	S	Pro	Leu	Glu	Glu	Glu	Arg	Ala	NH2
aggrecan core protein	P16112	17-2415	peptide-Asn360+Phe-peptide	P	Pro	Glu	Asn	Asn	Asn	Phe	Gly	Val
aggrecan core protein	P16112	17-2415	peptide-Asp460+Leu-peptide	P	Ser	Glu	Asp	Asp	Asp	Val	Val	Gln
Aggrecan core protein	Q29011	1-537	peptide-Asn344+Phe-peptide	P	Pro	Glu	Asn	Asn	Asn	Phe	Gly	Val
Aggrecan core protein	Q29011	1-537	peptide-Asp447+Leu-peptide	P	Ser	Glu	Asp	Asp	Asp	Val	Val	Gln

Table A.14: MMP 7 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
aggrecan core protein	P16112	17-2415	peptide-Asn360+Phe-peptide	P	Pro	Glu	Asn	Asn	Asn	Phe	Gly	Val
aggrecan core protein	P16112	17-2415	peptide-Glu392+Ala-peptide	P	Glu	Gly	Glu	Glu	Glu	Arg	Gly	Ser
aggrecan core protein	P16112	17-2415	peptide-Asp460+Leu-peptide	P	Ser	Glu	Asp	Asp	Asp	Val	Val	Gln
Aggrecan core protein	Q29011	1-537	peptide-Asn344+Phe-peptide	P	Pro	Glu	Asn	Asn	Asn	Phe	Gly	Val
Aggrecan core protein	Q29011	1-537	peptide-Asp447+Leu-peptide	P	Ser	Glu	Asp	Asp	Asp	Val	Val	Gln

Table A.15: MMP 8 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
A disintegrin and metalloproteinase with thrombospondin motifs 4 precursor	O75173	1-837	peptide-Arg212+Phe-peptide	P	Ala	Lys	Arg	Arg	Arg	Ala	Ser	Leu
Ac-Pro-Leu-Gly-Leu-Arg-Ser-Lys	nan	nan	Ac-Pro-Leu-Gly-Leu+Arg-Ser-Lys	S	Leu	Gly	Leu	Leu	Leu	Ser	Lys	-
Ac-Pro-Leu-Gly-Leu-Arg-Ser-Lys	nan	nan	Ac-Pro-Leu-Gly-Leu-Arg+Ser-Lys	S	Gly	Leu	Arg	Arg	Arg	Lys	-	-
Ac-Pro-Leu-Ser-Leu-Arg-Ser-Lys	nan	nan	Ac-Pro-Leu-Ser-Leu+Arg-Ser-Lys	S	Leu	Ser	Leu	Leu	Leu	Ser	Lys	-
Ac-Pro-Leu-Ser-Leu-Arg-Ser-Lys	nan	nan	Ac-Pro-Leu-Ser-Leu-Arg+Ser-Lys	S	Ser	Leu	Arg	Arg	Arg	Lys	-	-

Table A.16: MMP 9 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
14-3-3 protein beta/alpha	Q9CQV8	21-43	peptide-Ala27+Met-peptide	N	Ala	Ala	Ala	Ala	Ala	Lys	Ala	Val
14-3-3 protein gamma	P61982	20-42	peptide-Ala26+Met-peptide	N	Ala	Ala	Ala	Ala	Ala	Lys	Asn	Val
14-3-3 protein sigma	O70456	1-248	peptide-Phe25+Met-peptide	N	Ala	Ala	Phe	Phe	Phe	Lys	Ser	Ala
182 kDa tankyrase-1-binding protein	P58871	307-402	peptide-Leu390+Leu-peptide	N	Pro	Gln	Leu	Leu	Leu	Thr	Glu	Gly
Acidic leucine-rich nuclear phosphoprotein 32 family member A	O35381	138-150	LLPQV+MYLDGYDR	N	Pro	Gln	Val	Val	Val	Tyr	Leu	Asp

Table A.17: MMP 10 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
Acrogranin	P28799	18-593	peptide-Ala359+Leu-peptide	P	Pro	Gln	Ala	Ala	Ala	Lys	Arg	Asp
aggrecan core protein	P16112	17-2415	peptide-Asn341+Phe-peptide	P	Pro	Glu	Asn	Asn	Asn	Phe	Gly	Val
aggrecan core protein	P16112	17-2415	peptide-Asp440+Leu-peptide	P	Ser	Glu	Asp	Asp	Asp	Val	Val	Gln
aggrecan core protein	P16112	17-2415	peptide-Gln353+Thr-peptide	P	Thr	Val	Gln	Gln	Gln	Val	Thr	Trp
amyloid beta A4 protein	P05067	18-770	peptide-Asn579+Met-peptide	P	Leu	Ala	Asn	Asn	Asn	Ile	Ser	Glu

Table A.18: MMP 14 Substrate Data

Substrate Name	Uniprot	Residue range	Cleavage Site	Cleavage type	P4	P3	P2	P1	P1'	P2'	P3'	P4'
45 kDa calcium-binding protein	Q9BRK5	37-362	peptide-Leu83+Gly-peptide	P	Val	Phe	Leu	Leu	Leu	Lys	Asp	Leu
45 kDa calcium-binding protein	Q9BRK5	37-362	peptide-Leu87+Gly-peptide	P	Lys	Asp	Leu	Leu	Leu	Gly	Phe	Asp
actin, gamma-enteric smooth muscle	P63267	2-376	peptide-Ala20+Gly-peptide	P	Cys	Lys	Ala	Ala	Ala	Phe	Ala	Gly
actin, gamma-enteric smooth muscle	P63267	2-376	peptide-Arg29+Ala-peptide	P	Ala	Pro	Arg	Arg	Arg	Val	Phe	Pro
actin, gamma-enteric smooth muscle	P63267	2-376	peptide-Leu105+Leu-peptide	P	Pro	Thr	Leu	Leu	Leu	Thr	Glu	Ala

Table A.19: MMP 25 Substrate Data

Appendix B

Blocks Substitution Matrix (blosum 100)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0	8	-3	-4	-5	-2	-2	-3	-1	-4	-4	-4	-2	-3	-5	2	1	-1	-6	-5	-2
1	-3	10	-2	-5	-8	0	-2	-6	-1	-7	-6	3	-4	-6	-5	-3	-3	-7	-5	-6
2	-4	-2	11	1	-5	-1	-2	-2	0	-7	-7	-1	-5	-7	-5	0	-1	-8	-5	-7
3	-5	-5	1	10	-8	-2	2	-4	-3	-8	-8	-3	-8	-8	-5	-2	-4	-10	-7	-8
4	-2	-8	-5	-8	14	-7	-9	-7	-8	-3	-5	-8	-4	-4	-8	-3	-3	-7	-6	-3
5	-2	0	-1	-2	-7	11	2	-5	1	-6	-5	2	-2	-6	-4	-2	-3	-5	-4	-5
6	-3	-2	-2	2	-9	2	10	-6	-2	-7	-7	0	-5	-8	-4	-2	-3	-8	-7	-5
7	-1	-6	-2	-4	-7	-5	-6	9	-6	-9	-8	-5	-7	-8	-6	-2	-5	-7	-8	-8
8	-4	-1	0	-3	-8	1	-2	-6	13	-7	-6	-3	-5	-4	-5	-3	-4	-5	1	-7
9	-4	-7	-7	-8	-3	-6	-7	-9	-7	8	2	-6	1	-2	-7	-5	-3	-6	-4	4
10	-4	-6	-7	-8	-5	-5	-7	-8	-6	2	8	-6	3	0	-7	-6	-4	-5	-4	0
11	-2	3	-1	-3	-8	2	0	-5	-3	-6	-6	10	-4	-6	-3	-2	-3	-8	-5	-5
12	-3	-4	-5	-8	-4	-2	-5	-7	-5	1	3	-4	12	-1	-5	-4	-2	-4	-5	0
13	-5	-6	-7	-8	-4	-6	-8	-8	-4	-2	0	-6	-1	11	-7	-5	-5	0	4	-3
14	-2	-5	-5	-5	-8	-4	-4	-6	-5	-7	-7	-3	-5	-7	12	-3	-4	-8	-7	-6
15	1	-3	0	-2	-3	-2	-2	-2	-3	-5	-6	-2	-4	-5	-3	9	2	-7	-5	-4
16	-1	-3	-1	-4	-3	-3	-3	-5	-4	-3	-4	-3	-2	-5	-4	2	9	-7	-5	-1
17	-6	-7	-8	-10	-7	-5	-8	-7	-5	-6	-5	-8	-4	0	-8	-7	-7	17	2	-5
18	-5	-5	-5	-7	-6	-4	-7	-8	1	-4	-4	-5	-5	4	-7	-5	-5	2	12	-5
19	-2	-6	-7	-8	-3	-5	-5	-8	-7	4	0	-5	0	-3	-6	-4	-1	-5	-5	8

Table B.1: BLOSUM 100 Substitution Matrix