COMPARATIVE ANALYSIS OF NEGATIVE SAMPLING STRATEGIES IN PROTEIN-PROTEIN INTERACTION PREDICTION

ZEHRA KESEMEN

ÖZYEĞİN UNIVERSITY

JULY, 2025

COMPARATIVE ANALYSIS OF NEGATIVE SAMPLING STRATEGIES IN PROTEIN-PROTEIN INTERACTION PREDICTION

by
Zehra Kesemen

Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in
Computer Science

Advisor: Assoc. Prof. Reyhan Aydoğan
Coadvisor: Asst. Prof. İlknur Karadeniz

Graduate School of Science and Engineering
Özyeğin University
İstanbul

July, 2025

COMPARATIVE ANALYSIS OF NEGATIVE SAMPLING STRATEGIES IN PROTEIN-PROTEIN INTERACTION PREDICTION

Approved by:

Assoc. Prof. Reyhan Aydoğan, Advisor
Department of Artificial Intelligence and Data Engineering
Özyeğin University

Prof. Hasan Fehmi Ateş
Department of Artificial Intelligence and Data Engineering
Özyeğin University

Asst. Prof. İlknur Karadeniz, Co-advisor
Department of Artificial Intelligence and Data Engineering
Özyeğin University

Prof. Olcay Taner Yıldız
Department of Artificial Intelligence and Data Engineering
Özyeğin University

Prof. Arzucan Özgür
Department of Computer Engineering
Boğaziçi University

Approval Date: May 15, 2025

DECLARATION OF ORIGINALITY

I hereby declare that I am the sole author of this thesis and that this is the true copy of my thesis, including the final revisions, approved by my thesis committee. All data and information have been obtained, produced, and presented in accordance with the rules of research ethics and principles of academic honesty. As required by these rules, to the best of my knowledge I have acknowledged ideas, thoughts, and any copyrighted material in accordance with the standard referencing rules. I certify that any part of this thesis has not been submitted for a degree or diploma in another educational institution.

Zehra Kesemen

ABSTRACT

The accurate prediction of protein-protein interactions (PPIs) between host and pathogen proteins is essential for understanding viral infection mechanisms. However, a significant challenge is the lack of experimentally verified negative interactions (i.e., non-interacting pairs), which makes the development of effective negative sampling strategies essential. This thesis presents a comparative analysis of negative sampling strategies, specifically cluster-based methods. It proposes a novel Clustering-based Negative Sampling (CNS) method that aims to minimize the incorrect selection of negative interactions as positive interactions (i.e., experimentally confirmed interacting pairs). Furthermore, a structured evaluation methodology, Reliability of Sampling to Avoid False Examples (R-SAFE), is introduced to quantify the accuracy of these negative sampling methods. Experimental evaluations using four publicly available virus-host interaction datasets show that the characteristics of the datasets play a crucial role in the reliability of the negative sampling strategies. Therefore, a decision tree model is used to recommend the most appropriate negative sampling strategy based on the dataset features. The results emphasize the importance of negative sampling strategies to improve the predictive reliability of PPI models.

Keywords: Host-pathogen interactions, Negative sampling strategy, Machine learning methods, Binary classification, Evaluation methodology

ÖZET

Konakçı ve patojen proteinleri arasındaki protein-protein etkileşimlerinin (PPI) doğru bir şekilde tahmin edilmesi, viral enfeksiyon mekanizmalarının anlaşılması açısından büyük önem taşımaktadır. Ancak, deneysel olarak doğrulanmış negatif etkileşimlerin eksikliği (etkileşmeyen protein çiftleri), bu alanda karşılaşılan temel zorluklardan biridir. Bu eksiklik, etkili negatif örnekleme stratejilerinin geliştirilmesini gerekli kılmaktadır. Bu tez kapsamında, özellikle kümeleme tabanlı yaklaşımlar olmak üzere çeşitli negatif örnekleme stratejileri karşılaştırmalı olarak analiz edilmiştir. Çalışmada, negatif etkileşimlerin yanlışlıkla pozitif etkileşimler olarak seçilme riskini (deneysel olarak doğrulanmış etkileşimli çiftler) en aza indirmeyi amaçlayan yeni bir Kümeleme Tabanlı Negatif Örnekleme (CNS) yöntemi önerilmektedir. Ayrıca, bu stratejilerin doğruluğunu nicel olarak değerlendirebilmek amacıyla, Yanlış Örneklerden Kaçınmak için Örneklemenin Güvenilirliği (R-SAFE) adını taşıyan yapılandırılmış bir değerlendirme metodolojisi geliştirilmiştir. Dört halka açık virüs-konakçı etkileşim veri seti üzerinde gerçekleştirilen deneysel çalışmalar, veri setlerinin yapısal özelliklerinin negatif örnekleme stratejilerinin güvenilirliği üzerinde belirleyici bir etkisi olduğunu ortaya koymaktadır. Bu doğrultuda, veri seti özelliklerine dayalı olarak en uygun negatif örnekleme stratejisini önermek için bir karar ağacı modeli kullanılmıştır. Elde edilen sonuçlar, PPI modellerinin öngörüsel güvenilirliğini artırmak için negatif örnekleme stratejilerinin önemini vurgulamaktadır.

Anahtar Kelimeler: Konakçı-patojen etkileşimleri, Negatif örnekleme stratejisi, Makine öğrenmesi yöntemleri, İkili sınıflandırma, Değerlendirme yöntemi

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Reyhan Aydoğan. Her knowledge, guidance, and support were truly valuable in shaping the direction and content of this thesis. Despite the challenges posed by the fact that I came from a different field, she was always understanding and guided me at every stage. She never refused to support me, even under challenging circumstances, and was always available when I needed help. She made me part of her research group and created a supportive academic and personal environment. Thanks to the opportunities and resources she generously provided.

I would also like to thank my co-supervisor, Asst. Prof. İlknur Karadeniz, for kindly agreeing to participate in this study and supporting my work throughout the thesis process. Her expertise in bioinformatics and thoughtful guidance significantly shaped the scientific depth of this research. I had the opportunity to share an important academic experience with her during the conference, and her presence and encouragement made the journey even more enjoyable and meaningful. Working with two dedicated women researchers has been a valuable experience and opportunity.

I wholeheartedly thank my dear Ozan, who has been by my side every step of the way and whose love, patience and unwavering faith in me have been a constant source of strength throughout this journey.

My heartfelt thanks to my dear family who have always been there for me with their support, patience, strength and belief in me.

I would also like to thank the Scientific and Technological Research Council of Türkiye (TÜBİTAK) for their financial support provided under the 2210-C National MSc Scholarship Program for Priority Areas in Science and Technology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

ACC	Accuracy
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BPE	Byte Pair Encoding
CNS	Clustering-based Negative Sampling
F1	F1-Score
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GAM	Generalized Additive Model
GraphSAGE	Graph Sample and Aggregate
HPIDB	Host–Pathogen Interaction Database
LR	Logistic Regression
ML	Machine Learning
MI	Molecular Interaction
NPV	Negative Predictive Value
P.C.	Positive Count
PCA	Principal Component Analysis
PHISTO	Pathogen–Host Interaction Search Tool
PPV	Positive Predictive Value (Precision)
PPI	Protein-Protein Interaction

PU	Positive-Unlabeled
RF	Random Forest
R-SAFE	Reliability of Sampling to Avoid False Examples
SP	Specificity
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate

P	Set of known positive host-pathogen protein-protein interactions
U	Set of unknown (unlabeled) host-pathogen protein pairs
N	Set of selected negative samples
T	Combined set of positive and unlabeled samples ($T = P \cup U$)
r	Positive-to-negative ratio
k	Number of clusters
C_i	The i^{th} cluster
$C_{i\text{-center}}$	Centroid (center) of cluster C_i
s_i	Number of negative samples to be selected from cluster C_i
$\|X\|$	Cardinality (number of elements) of set X
r_i	Ratio of unlabeled to total samples in cluster C_i
$d(\cdot, \cdot)$	Distance function (e.g., Euclidean, Cosine, Canberra)
u	An unlabeled sample
m	Index of the closest cluster centroid to sample u

1. INTRODUCTION

Predictive modeling is central to modern data analytics and enables systems to distinguish between positive and negative instances in a variety of applications. Advances in machine learning such as deep neural networks and transformation models have improved the ability to learn from complex data and improve the accuracy and scalability of binary classification tasks. A crucial component of predictive modeling is the creation of high-quality training datasets. High-quality data allows the models to generalize effectively by mapping the underlying distribution of the target task. This facilitates convergence during training and improves prediction performance on unseen data. Conversely, datasets with mislabeled instances, sampling biases, or poorly defined negative classes can propagate errors, distort learned representations, and ultimately compromise the reliability and interpretability of the models.

A significant challenge in computational biology is the absence of well-defined negative labels [1]. For example, experimental databases typically provide only positive interaction pairs for protein-protein interaction (PPI) prediction, particularly in species such as pathogen-host systems. However, predictive models also need negative examples—pairs of proteins that do not interact—in order to learn effective decision boundaries. Negative sampling is a widely used technique in scenarios where only positive labels are available or where negative labels are scarce [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Instead of labeling all possible negatives, which is often not possible due to the vastness of the input space, negative sampling selects a subset of examples that are assumed not to belong to the positive class. This approach reduces the computational effort, improves the class balance and enables the model to learn meaningful discriminative patterns. Negative sampling has already been used successfully in various fields, e.g. natural language processing, recommender systems and computational biology. A notable example is the word2vec model [21] in natural language processing,

where negative sampling is used to efficiently train word embeddings by distinguishing target words from randomly sampled ”noise” words, thereby reducing computational complexity while preserving semantic relations in the learned vector space. These successes have inspired the adaptation of similar strategies in other domains with sparse or weakly labeled data.

In biological applications, as in other fields, negative sampling is a crucial step, as reliable negative markers are often not available, making direct training of supervised models impossible. Negative sampling offers a practical solution by allowing researchers to simulate negative examples under certain assumptions [22]. In the context of predicting protein-protein interactions (PPI) between pathogens and their hosts, which are crucial for understanding infection mechanisms, developing targeted therapies and drug discovery, one of the key requirements for training predictive models is the generation of datasets containing both positive and negative interaction pairs. While positive pairs are usually derived from curated experimental data, negative examples (i.e. protein pairs that are assumed not to interact) are often not available and have to be generated artificially.

The quality of the negative examples has a decisive influence on the performance of the model. Poorly selected negative examples, especially those containing unknown positive examples, lead to label noise, that can affect accuracy and limit generalizability. Moreover, in imbalanced datasets, where negative examples far outweigh positives, deep learning models are prone to bias toward predicting the majority class. To mitigate this, strategies like sampling or weighting negative samples are commonly employed to ensure balanced learning. These approaches are especially important in the context of pathogen-host protein interaction prediction, where datasets are often sparse and noisy.

To overcome the challenge of creating reliable negative data sets, several negative sampling strategies have been proposed, including random sampling [2, 3, 4, 5, 6, 7, 8,

9, 10, 11, 12, 13], dissimilarity-based techniques [16, 17, 18, 19, 20, 23], and clustering-based methods[14, 15]. The most commonly used approach assumes that randomly selected unlabeled protein pairs do not interact and can therefore be labeled as negative. While this method is simple and computationally efficient, it carries the risk of classifying true but unlabeled interactions as negative, leading to noise in the training data. Such misclassification may affect both the biological credibility of the dataset and the reliability of the prediction models, as it leads to systematic biases. In response to these limitations, alternative strategies have been developed to reduce the incorrect selection of positive samples as negatives. Dissimilarity-based methods avoid protein pairs that are similar to known interactors based on sequence or functional features, under the assumption that such pairs are more likely to interact. Similarly, clustering-based approaches extend this principle by grouping proteins according to structural, physicochemical or functional similarity and selectively drawing negative samples from different or biologically unrelated clusters. These methods aim to more accurately reflect true non-interacting pairs and improve the reliability of downstream predictions.

However, the systematic evaluation and effectiveness of these sampling strategies, which may often vary between datasets with different characteristics such as sequence composition or functional annotations, are not well studied in the context of host-pathogen PPI prediction. Moreover, few studies have conducted a systematic comparison of these techniques under standardized evaluation criteria. In this study, we systematically investigate the effects of five negative sampling strategies namely random sampling, dissimilarity-based approaches, Wei's method, Wang's method, and recently proposed clustering-based CNS method on the prediction of host-pathogen PPIs. We present Reliability of Sampling to Avoid False Examples (R-SAFE), a framework for estimating the risk of sampling unknown positives and evaluating the biological and computational impact of different sampling methods. Using four benchmark datasets, we evaluate the

impact of each strategy on different data splits and random seeds and analyze both the predictive performance and the spatial distribution of mislabeled samples in the feature space. To further enhance biological validity, we extend the R-SAFE framework by incorporating the Negatome dataset, which provides experimentally validated non-interacting protein pairs. Our results show that the performance of negative sampling strategies varies significantly depending on the dataset and experimental setup. While certain methods perform well in specific contexts, no single strategy outperforms the others in all scenarios. Furthermore, we adopt a machine learning-based approach to identify the sampling strategy that is likely to produce the most effective selection based on the characteristics of the dataset and the corresponding most appropriate sampling strategy based on the experimental results conducted in this study. Thus, we can deduce the key features of the dataset and select the optimal sampling strategy for the underlying dataset.

1.1 Contributions of the Thesis

This thesis addresses an important challenge in predicting interactions between viruses and human proteins: reliably selecting negative samples (non-interacting protein pairs). In this area, positive interactions (experimentally confirmed interacting protein pairs) are recorded in public databases. Most protein pairs remain unlabeled (with no known interaction status). Therefore, negative samples must be generated artificially from these unlabeled pairs. This step is crucial because errors in selecting negative samples, e.g., labeling positive interactions as negative, can affect the prediction reliability of the model.

The first contribution of this thesis is the development of a novel negative sampling method called Clustering-based Negative Sampling (CNS). CNS applies a hierarchical clustering approach to group protein pairs based on their structural similarities and selects negative samples from the unlabeled data using different selection strategies. This way,

negative samples are chosen more carefully, reducing the risk of mistakenly selecting positive interactions as negatives.

The second important contribution introduces an evaluation framework called Reliability of Sampling to Avoid False Examples (R-SAFE). In contrast to traditional evaluation approaches that focus solely on predictive performance, R-SAFE simulates unknown positive interactions and quantifies how often different sampling methods falsely classify them as negative. Building on the findings of R-SAFE, the decision tree model is used to recommend the most appropriate negative sampling strategy for a given dataset. This model uses dataset features such as amino acid composition, sequence similarity, physicochemical properties, and sequence diversity to guide the selection process, thus supporting a data-driven and adaptive approach to negative sampling.

Finally, the proposed methods are evaluated through extensive experiments using four publicly available virus-host interaction datasets. The results show that the performance of negative sampling strategies strongly depends on the characteristics of the data set. These results emphasize the need for context-specific adaptive sampling strategies rather than relying on a single universal solution.

1.2 Organization of the Thesis

This work is divided into six main chapters. Chapter 1 contains the introduction and the context of the research. Chapter 2 provides an overview of the relevant literature on existing negative sampling methods and their limitations. Chapter 3 outlines the biological background, problem definition, foundations of machine learning approaches, and benchmark methods relevant to the thesis. Chapter 4 presents the proposed clustering-based negative sampling (CNS) strategy in detail, including the methodology and the evaluations. Chapter 5 presents and validates the Reliability of Sampling to Avoid False Examples (R-SAFE) evaluation framework, the chapter incorporates the Negatome dataset

with experimentally validated non-interacting protein pairs. Chapter 6 also describes the decision tree-based selection method for recommending the optimal sampling strategy. Chapter 7 concludes the thesis by summarizing the main findings and contributions.

2. RELATED WORK

Protein-Protein Interaction (PPI) are central to many biological processes, including signaling, metabolism and disease mechanisms. Experimental methods to detect PPIs are valuable but often costly, labor-intensive and limited in scope. Therefore, computational prediction of PPIs has become essential for accelerating the discovery of interactions, guiding experiments, and advancing applications in drug discovery and systems biology.

Recent advances in machine learning have made it possible to study these interactions on a large scale. However, the reliability of predictive models is highly dependent on the quality of the training data, especially the construction of negative samples. In recent years, the growing number of experimentally validated protein-protein interactions (PPIs) has led to the development of several comprehensive databases cataloging interaction data across different organisms. Notable examples include DIP [24], MINT [25], BioGRID [26], IntAct [27] and STRING [28]. These repositories provide important resources for training PPI prediction models and benchmarking their performance. While these databases primarily focus on experimentally confirmed interactions, the inclusion of high-quality negative datasets(i.e. non-interacting protein pairs) is equally important to improve classification accuracy and model reliability [29].

One of the major challenges in PPI prediction is the lack of experimentally validated negative examples. Since true non-interacting protein pairs are rarely identified and reported, distinguishing between truly non-interacting pairs and those with unknown interaction status remains difficult. A rare curated resource for negative interactions is the Negatome database [30], which contains about $2,000$ non-interacting protein pairs derived from the literature and structural analyzes of protein complexes. However, the limited size of this dataset makes it unsuitable for training data-intensive models, especially for deep learning architectures that require extensive training examples. Therefore,

the generation of synthetic but meaningful negative examples remains an active area of research.

With the advent of machine learning, several strategies have been developed to address the lack of negative data in the prediction of biological interactions. These include random sampling, dissimilarity-based approaches, and more recent techniques such as Positive-Unlabeled (PU) learning and knowledge graph-based methods. A comprehensive review of techniques for predicting virus–host interactions concludes that while existing models are still limited in their ability to capture the full complexity of these networks, ongoing improvements in machine learning are rapidly advancing the field.

The most widely used negative sampling strategy is random negative sampling, proposed by Shen et al. [31] and is based on earlier work by Chen et al. [32]. In this approach, all possible protein pairs are constructed and known interacting pairs are removed. The remaining pairs are considered as negative candidates from which a random subset is selected. Although this method is computationally efficient and easy to implement, it carries the risk that undetected positive pairs are considered negative pairs, introducing noise into the labeling. This noise can hinder learning, reduce prediction sensitivity and inflate performance metrics. Despite these limitations, random sampling remains a popular basis due to its ease of application and has been used in numerous studies [2, 3, 4, 33, 6, 7, 8, 9, 10, 11, 12, 13].

To enhance biological relevance, Guo et al. [34] proposed a heuristic approach that leverages subcellular localization data obtained from Swiss-Prot.[1]. Assuming that proteins located in different cellular compartments are unlikely to interact, they constructed negative samples by linking proteins from different compartments. The reliability of the data was ensured by filtering out proteins with ambiguous annotations (e.g. “putative”

[1] http://www.expasy.org/sprot/

or “hypothetical”) and categorizing the remaining proteins into eight different subcellular groups. Careful pairing rules were applied to avoid ambiguous combinations, such as those involving doubly localized proteins. Several studies [35, 36] have extended this strategy and demonstrated its utility in generating biologically meaningful negative data sets.

Another prominent strategy is dissimilarity-based negative sampling, which aims to avoid selecting negatives that are too similar to known positives. This technique, first introduced by Eid et al. [37] in the DeNovo method, selects negative samples based on sequence similarity thresholds. For example, if two viral proteins are highly similar and one of them is known to interact with a human protein, pairing of the other viral protein with the same human protein is avoided to reduce the risk of false negative results. This method has been used in several recent studies [16, 17, 18, 19, 20, 38]. Neumann et al. [22] point out, however, that although these strategies reduce the number of false negatives in the training data, they may also oversimplify the classification task. They argue that applying similarity constraints to the test set inflates the performance metrics, leading to misleading conclusions about the generalization ability of a model.

To overcome these limitations, clustering-based strategies for negative sampling have gained traction. Wei et al. [14] used MiniBatchKMeans clustering to reduce the probability of selecting false negatives when generating negative samples for miRNA-disease association prediction. Wang et al. [15] also proposed several strategies to address data imbalances using clustering and nearest neighbor-based approaches. Their K-Means clustering-based method selects representative samples proportional to the clusters to improve the distributional balance in the training data.

Negative sampling also plays a crucial role in other areas of machine learning such as natural language processing, recommender systems and graph learning. Yang et al. studied various negative sampling techniques and highlighted their effects on class

imbalance, training efficiency and embedding optimization. Their results emphasize the importance of strategically selecting negative samples to improve the robustness of the model. Similarly, Yang et al. analyzed the theoretical effects of different sampling distributions on graph learning tasks and proposed adaptive strategies to improve link prediction and node classification performance.

Recent techniques have extended negative sampling by incorporating broader machine learning paradigms. Semi-supervised learning methods, such as semi-supervised SVMs, improve prediction accuracy by using unlabeled data. Self-training approaches also improve generalization by iteratively augmenting the training sets with pseudo-labeled negatives with high confidence, as shown in compound–protein interaction studies.

Knowledge graph-based methods offer another promising direction. Frameworks such as OntoProtein embed biological ontologies into protein representations and use knowledge-based negative sampling to avoid biologically meaningful examples. Moreover, recent work on embedding knowledge graphs has emphasized the importance of explicit negative information. TrueWalks, for example, integrates known negative relations to improve the prediction of biomedical connections, emphasizing the role of structured negative information in representation learning.

In addition to the above strategies, positive and unlabeled learning (PU) offers a compelling solution to the challenge of limited negative data in predicting biological interactions. PU learning trains models using only known positive instances and a pool of unlabeled examples that may contain both positive and negative pairs. This approach has become very popular in bioinformatics as it is suitable for domains where there are few labeled negative examples. Recent theoretical work has introduced specialized loss functions to reduce bias and overfitting in PU environments and enable integration with deep learning methods [39]. PU learning has been effectively applied to tasks such as disease–gene association and gene function prediction [40]. While PU learning reduces

the need for curated negative labels, it often suffers from instability, sensitivity to class priors, and problems in reliably estimating the true negative distribution. In contrast, well-designed negative sampling strategies allow explicit control over class balance and incorporation of domain-specific biological heuristics. These advantages make negative sampling a more feasible and interpretable solution for many tasks, especially in the prediction of protein–protein and virus–host interactions, where biological knowledge can guide the selection of likely non-interacting pairs.

All of these studies emphasize the critical role of negative sampling in supervised learning tasks, especially in domains where reliable negative labels are scarce. Carefully designed strategies are important not only for creating more realistic training datasets, but also for fair and interpretable model evaluation. In particular, in the context of PPI prediction, selecting reliable negative samples is crucial for accurately discriminating between interacting and non-interacting protein pairs, thereby significantly increasing the models' biological relevance and predictive power. Despite the recent advances, the prediction of PPI between host and pathogen still lacks a systematic evaluation of negative sampling strategies.

3. PROBLEM DEFINITION AND BACKGROUND

This chapter reviews the essential biological concepts underlying protein-protein interactions, focusing on host-pathogen systems. We then formalize the problem of predicting such interactions as a binary classification task, introduce the notation for positive and unlabeled interaction sets, and describe the range of supervised models and evaluation metrics employed throughout this study. Finally, we present the feature extraction, selection, and clustering techniques that form the methodological foundation for our negative sampling strategies.

3.1 Biological Background

Understanding the biological foundations of proteins and their structural organization is essential for modeling protein-protein interactions computationally. This section provides an overview of the fundamental properties of proteins and the hierarchical nature of their structures.

3.1.1 Proteins

Proteins are complex molecular substances that perform crucial tasks for life in all living organisms. They consist of long chains of amino acids that act as their building blocks. Proteins are made up of 20 unique amino acids, and the specific sequence of these amino acids determines the unique structure and function of each protein. This sequence allows proteins to fold into complex three-dimensional shapes, allowing them to interact with other molecules in highly specific and effective ways.

3.1.2 Protein Structure

Proteins have a complex structure that is divided into four levels: primary, secondary, tertiary, and quaternary structure. The simplest level is the primary structure,

which consists of a linear sequence of amino acids linked by peptide bonds. The next level is the secondary structure, which involves local folding within the polypeptide chain, often forming alpha helices and beta sheets. Tertiary structure refers to the overall three-dimensional shape of a single polypeptide chain. In contrast, quaternary structure describes how multiple polypeptide chains interact when the protein has more than one chain [41]. While higher-level structures provide more insight into a protein's function, primary sequence data is commonly used in studies because it is easier to analyze and readily available.

3.1.3 Pathogen-Host Interactions

Pathogen-host interactions describe the complex relationships between a host organism, e.g., a human, an animal, or a plant, and a pathogenic microorganism such as bacteria, viruses, fungi, or parasites that cause disease. At the molecular level, these interactions often involve specific protein-protein interactions between the host and the pathogen. The structures of these proteins are of crucial importance as they determine how the proteins on both sides combine and interact. These protein-protein interactions are crucial for the pathogen to attach to the host, penetrate its cells, and evade its defenses, as well as for the host to recognize and fight the pathogen.

In bioinformatics, protein-protein interactions (PPIs) are studied to understand how proteins communicate and coordinate in biological systems. PPIs are the physical contact between proteins necessary for the execution of cellular functions. In the context of pathogen-host interactions, mapping PPIs helps uncover how pathogens alter host processes at the molecular level and how hosts resist these changes. Computational methods are key in predicting, analyzing, and modeling these interactions, especially when experimental validation is limited or difficult. Researchers can uncover infection mechanisms, reconstruct interaction networks, and identify key proteins involved in disease processes

by studying the interaction between host and pathogen proteins. Computational analysis of pathogen-host PPIs also helps uncover patterns that support the discovery of new biological insights and contribute to advances in vaccine development, drug target identification, and systems biology.

3.2 Problem Definition

The set of host proteins, denoted $P_{Host}(i)$, is defined as the collection of individual host proteins represented by the elements $x_1, x_2, \ldots, x_i$, where i is an integer between 1 and N_{Host}, inclusive. The parameter N_{Host} represents the total number of host proteins in the set. This is formulated as follows:

$$P_{Host}(i) = \{x_1, x_2, \ldots, x_i \mid 1 \leq i \leq N_{Host}\}.$$

Here, each x_i represents a specific host protein, and the set includes all host proteins up to the N_{Host}^{th} element. Similarly, the set $P_{\text{Pathogen}}(j)$ includes proteins associated with pathogen proteins and is represented by the elements $y_1, y_2, \ldots, y_j$, where j ranges from 1 to N_{Pathogen}. This is expressed as follows:

$$P_{Pathogen}(j) = \{y_1, y_2, \ldots, y_j \mid 1 \leq j \leq N_{Pathogen}\}$$

Adjust $N_{Pathogen}$ and N_{Host} based on the specific number of proteins relevant to the dataset. The interaction denoted by I characterizes the relationship between a host and a protein, and it can be defined as follows:

$$I(x_i, y_j) = \begin{cases} 1, & \text{if there is an interaction between } x_i \text{ and } y_j \\ 0, & \text{otherwise} \end{cases}$$

We define the sets of positive and unlabeled protein-protein interactions by denoting them with the notations P and U, respectively:

$$P = \{(x_i, y_j) \mid I(x_i, y_j) = 1, (i, j) \in (N_{\text{host}} \times N_{\text{Pathogen}})\} \tag{3.1}$$

$$U = \{(x_i, y_j) \mid I(x_i, y_j) = 0, (i, j) \in (N_{\text{host}} \times N_{\text{Pathogen}})\} \tag{3.2}$$

where $(x_i, y_j) \in (P_{Host} \times P_{Pathogen})$. In other words, unlabeled protein interactions are all possible interactions that are not positive. The negative samples are denoted by N and the total number of negative samples ($|N|$) selected from the unlabeled samples is determined by multiplying the number of positive samples by the specified ratio of positive to negative, i.e.,

$$|N| = \text{Number of } P \times \text{positive-to-negative ratio.}$$

Our goal is to determine a reliable subset N, which is derived from the set U.

3.3 Formulating PPI Prediction as a Binary Classification Problem

The prediction of pathogen-host protein-protein interactions (PPI) can be formulated as a binary classification problem, where each protein pair is assigned one of two possible labels: interacting or non-interacting [42]. This binary framework simplifies the task, enabling the use of classification methods to determine whether a specific host and pathogen protein pair interacts.

Formally, given a host protein $x_i \in P_{Host}$ and a pathogen protein $y_j \in P_{Pathogen}$, their interaction is represented as a pair (x_i, y_j). A binary classifier assigns a label through a function $I(x_i, y_j) \in \{0, 1\}$, where $I(x_i, y_j) = 1$ indicates an interaction and $I(x_i, y_j) = 0$ indicates no interaction.

3.4 Supervised Classification Models

This section describes the supervised learning models used in this thesis for classification tasks.

3.4.1 Decision Tree

Decision trees are intuitive and interpretable models that recursively partition the feature space by selecting features that best partition the data according to some criterion,

such as Gini impurity or information gain. The algorithm evaluates all available features at each node and identifies the feature and threshold that lead to the best split based on the selected criterion. The data is then split into subsets according to this split, and the process is repeated recursively for each subset. This recursive partitioning continues until a stopping condition is met, e.g., reaching a maximum tree depth, reaching pure leaf nodes, or too few samples to justify further partitioning. The resulting tree structure represents a set of decision rules that can be easily visualized and interpreted.

3.4.2 Random Forest

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and random feature selection at each split [43]. By averaging the predictions of many trees, Random Forest reduces variance, improves robustness, and mitigates overfitting. In addition to strong predictive performance, it also provides feature importance scores that offer valuable insights into the contribution of each feature.

3.4.3 Generalized Additive Models with Interactions (GA^2M)

Generalized additive models with interactions (GA^2M) extend traditional Generalized Additive Model (GAM)s by modeling the effects of individual traits and selected pairwise interactions [44]. The model structure is defined as follows:

$$\text{prediction} = \beta_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j)$$

where f_i represents univariate functions and f_{ij} represents bivariate interaction functions learned from the data. This architecture balances the model's expressiveness and interpretability, making GA^{2}Ms suitable for tasks where transparent and understandable decision-making is required.

3.4.4 Logistic Regression

Logistic regression is a simple but effective classification model that estimates the probability of class membership using the logistic function [45]. It models the log-likelihood of the outcome as a linear combination of the input features. Therefore, it is well suited to problems where the relationship between features and outcomes is approximately linear. Despite its simplicity, logistic regression is a solid foundation often used in biological classification tasks for benchmarking purposes.

3.4.5 Support Vector Machines (SVM)

Support Vector Machine (SVM) are robust classifiers that determine the optimal hyperplane between two classes by maximizing the span between them [46]. In cases where the data is not linearly separable, SVM uses kernel functions (e.g., radial basis functions or polynomial kernels) to project the data into a higher dimensional space where linear separation is possible. By focusing only on support vectors (critical examples closest to the decision boundary) SVM models achieve strong generalization capabilities with efficient memory usage.

3.5 Evaluation Metrics for Classification

Machine Learning (ML) provides a powerful toolkit for analyzing complex biological data and has become increasingly important in computational biology. By learning patterns from data, ML algorithms can make predictions or decisions without being explicitly programmed. Among the major learning paradigms, supervised, unsupervised, and reinforcement learning, supervised learning is most commonly used in biological prediction tasks, including predicting protein-protein interactions (PPI). In supervised learning, models are trained on labeled data to learn associations between input features and target outcomes. Evaluating the performance of classification models requires metrics

that reflect correctness and reliability. This thesis uses the following evaluation metrics to assess model performance across different negative sampling strategies and datasets.

To define these metrics, we refer to the confusion matrix, which consists of the following terms:

- **True Positive (TP) (True Positive)**: Number of correctly predicted positive instances.
- **True Negative (TN) (True Negative)**: Number of correctly predicted negative instances.
- **False Positive (FP) (False Positive)**: Number of negative instances incorrectly predicted as positive.
- **False Negative (FN) (False Negative)**: Number of positive instances incorrectly predicted as negative.

Based on these definitions, the evaluation metrics are computed as follows:

- **Accuracy (ACC)**: Proportion of correctly predicted instances out of all instances.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Positive Predictive Value (Precision) (PPV) (Positive Predictive Value / Precision)**: Proportion of true positive predictions out of all predicted positives.

$$\text{PPV} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity / True Positive Rate)**: Proportion of true positive predictions out of all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score (F1)**: Harmonic mean of precision and recall, establishing a balance between the two.

$$\text{F1} = 2 \times \frac{\text{PPV} \times \text{Recall}}{\text{PPV} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

- **Specificity (SP) (Specificity / True Negative Rate)**: Proportion of true negatives out of all actual negatives.

$$\text{SP} = \frac{TN}{TN + FP}$$

- **Negative Predictive Value (NPV)**: Proportion of true negative predictions out of all predicted negatives.

$$\text{NPV} = \frac{TN}{TN + FN}$$

- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: Area under the Receiver Operating Characteristic curve, which reflects the model's ability to distinguish between positive and negative classes. This is computed by plotting the True Positive Rate (TPR) against the false positive rate (False Positive Rate (FPR)) at various threshold settings.

In the context of predicting PPI, Recall and F1 scores are particularly emphasized, as failure to detect true interactions may result in biologically important connections being overlooked.

3.6 Feature Selection and Reduction of Dimensionality

Feature selection and dimensionality reduction techniques are crucial for enhancing model performance, decreasing overfitting, and improving interpretability in high-dimensional biological datasets, such as protein-protein interaction data. An overview of the methods employed in this study, including the Chi-Square (χ^2) test, SelectKBest, and Principal Component Analysis (PCA), is given in this section.

3.6.1 Principal Component Analysis (PCA)

An unsupervised linear dimensionality reduction technique called Principal Component Analysis (PCA) converts the original feature space into a new set of orthogonal axes, known as principal components. These components are ranked according to the amount of variance they capture from the original data and are obtained by computing the eigenvectors of the data's covariance matrix.

The first principal component (Component 1, PC1) represents the direction with the highest variance in the data set and captures the most informative patterns. The second principal component (Component 2, PC2) is orthogonal to PC1 and captures the second largest variance. By projecting the data onto the space defined by the first two components, one can often visualize class separability and identify clusters or outliers in the data.

3.6.2 Feature Selection Methods

Univariate feature selection methods evaluate each feature independently concerning the target variable and rank them based on an evaluation criterion. These methods are often used in supervised learning tasks to retain only the most relevant features. The SelectKBest technique was used in combination with the Chi-Square test, which assesses the statistical dependence between individual features and the target. SelectKBest is a widely used feature selection technique that selects the best k features according to a given statistical test. It evaluates each feature individually and ranks them according to their score so that only the most informative features can be selected for model training. This method is widely used due to its simplicity and flexibility in combination with different evaluation functions such as ANOVA, mutual information, or chi-square.

The Chi-Square test is a statistical method for measuring the dependency between two categorical variables. In connection with the selection of characteristics, it evaluates

how much the observed frequencies of a characteristic deviate from the expected frequencies under the assumption of independence from the target variable. A higher χ^2 value indicates a greater relevance of the characteristic. This test is particularly suitable for discretely assessed characteristics and is often used in SelectKBest.

3.7 Clustering Techniques

This thesis uses clustering algorithms to cluster protein embeddings that represent sequence features. Grouping proteins into clusters facilitates the selection of negative samples from different groups and thus reduces the risk of selecting biologically similar or interacting protein pairs.

3.7.1 K-Means Clustering

K-Means clustering, proposed by MacQueen in 1967, groups data points into k clusters by minimizing the distance between the points and their assigned cluster centers. In this algorithm, k random cluster centers are first selected. Each data point is assigned to the nearest center based on a chosen distance metric, usually Euclidean distance. Once assigned, the cluster centers are updated by calculating the average points assigned to each cluster. These steps of assigning and updating are repeated iteratively until the cluster centers stabilize, and no significant changes occur.

3.7.2 MiniBatch K-Means Clustering

MiniBatch K-Means [47] is a faster and more scalable variant of the standard K-Means algorithm. Instead of using the entire data set to update the cluster centers at each iteration, MiniBatch K-Means randomly selects small batches of data points. Each point in the mini-batch is assigned to the nearest cluster center, and the centers are incrementally updated based on these assignments. This mini-batch strategy significantly reduces computation time and memory requirements and allows the algorithm to process large

data sets efficiently. Although the resulting clusters can be somewhat inaccurate compared to the clusters generated with standard K-Means, MiniBatch K-Means achieves a favorable trade-off between the clustering quality and computational efficiency.

3.7.3 Agglomerative Clustering

Agglomerative clustering is a hierarchical method in which each data point is initially treated as a single cluster [48]. At each step, the two closest clusters are identified and merged based on a specific distance metric and linkage method, such as simple, complete, average, or Ward's linkage. As the merging continues, a hierarchical, tree-like structure called a dendrogram emerges, representing how the clusters combine at different levels of similarity. The dendrogram can be truncated at a selected threshold to determine the desired number of clusters.

3.8 Feature Representation Techniques

To convert amino acid sequences into representative numerical vectors, we apply natural language processing techniques that capture structural and contextual patterns, as described in the following section.

3.8.1 Byte-Pair Encoding (BPE)

Byte Pair Encoding (BPE) [49] is a method initially developed for data compression and later adapted for sequence tokenization. It simplifies protein sequences by merging frequently occurring adjacent amino acid pairs into new tokens, thus capturing common motifs while reducing sequence complexity. BPE operates iteratively by identifying and merging the most frequent substrings, progressively updating the vocabulary at each step. This process facilitates handling large and complex protein datasets by creating a compact and informative token representation. This thesis applies BPE tokenization to host and pathogen protein sequences separately. Separate vocabularies are constructed

for each group to capture better the specific sequence patterns and characteristics unique to host and pathogen proteins.

3.8.2 Doc2Vec Embeddings

Doc2Vec, which was introduced as an extension of Word2Vec [21], is a neural embedding method designed to learn fixed-length vector representations of entire sequences rather than individual elements. Unlike traditional word embeddings, which represent each token separately, Doc2Vec captures the entire semantic and structural information of a sequence in a single continuous vector.

In the context of PPI prediction, protein sequences are treated as documents and amino acids or motifs as words. In this thesis, motif-based tokenization was first applied to protein sequences using BPE to capture frequent local patterns. These tokenized sequences were then input to Doc2Vec models to learn sequence-level embeddings. Each protein sequence has been embedded in a 128-dimensional vector representation, which makes it possible to extract meaningful features of proteins and support tasks such as classification and interaction prediction. In the preprocessing stage, the amino acid sequences of all proteins in the training dataset are first divided into smaller sub-sequences using the Byte Pair Encoding (BPE) algorithm. This process is carried out separately for host proteins P_{Host} and pathogen proteins $P_{Pathogen}$, resulting in two different vocabularies that reflect the specific patterns of each organism.

Next, all protein sequences are used to build two separate text corpora, one for host proteins and one for pathogen proteins. These corpora are then used to train the Doc2Vec model, which transforms each protein sequence into a fixed-size numerical vector that captures the semantic and contextual information of the amino acid composition. Let $x_i \in \mathbb{R}^{128}$ denote the embedding vector of the i-th host protein, and $y_j \in \mathbb{R}^{128}$ denote the embedding vector of the j-th pathogen protein. Both x_i and y_j are learned embeddings

obtained through the Doc2Vec model. For each potential host-pathogen pair (x_i, y_j), the corresponding feature vector is constructed by concatenating their embeddings. This concatenated vector, denoted as $[x_i \| y_j] \in \mathbb{R}^{256}$, serves as the feature representation of the protein pair. These combined vectors are then used as input for the clustering algorithms applied during the negative sampling process.

3.9 Graph-Based Learning Approaches

To incorporate both protein sequence features and interaction network topology, we employ Graph Sample and Aggregate (GraphSAGE) as a graph-based embedding method in our study. GraphSAGE [50] is a scalable method for generating node embeddings in large graphs. Instead of training on the entire graph, it learns how to aggregate features from a node's local neighborhood, enabling inductive generalization to unseen nodes. GraphSAGE generates hybrid node embeddings in this thesis by integrating protein sequence information with graph topology. Protein sequences are first embedded into fixed-length vectors using Doc2Vec, capturing functional and evolutionary characteristics. These sequence embeddings are then combined with the local topological structure of the interaction graph through GraphSAGE's neighborhood aggregation process. As a result, the final node embeddings capture both the intrinsic properties of protein sequences and their relational context within the graph, providing a richer and biologically meaningful representation for downstream prediction tasks.

3.10 Benchmark Methods

This section outlines the benchmark methods used to evaluate the impact of negative sampling strategies on model performance in protein-protein interaction (PPI) prediction tasks throughout this thesis. The strategies considered include the widely adopted random sampling approach and dissimilarity-based sampling strategy, as well as two

clustering-based methods introduced in previous studies. In particular, we analyze two clustering-based methods from the literature: Wang's method [51], designed for protein interaction site prediction, and Wei's method [14], proposed for miRNA-disease association prediction. These methods serve as baselines for evaluating the effectiveness of our proposed sampling strategy.

3.10.1 Random Sampling Method

The random sampling strategy is a basic method widely used in PPI studies, first introduced by Shen *et al.* [52]. In this approach, all possible protein pairs are generated by combining every protein in the dataset. From these pairs, the positive examples that are experimentally verified as interacting proteins are removed. The remaining pairs are considered potentially non-interacting protein pairs. Negative samples are randomly selected from this pool of assumed non-interacting pairs.

3.10.2 Dissimilarity-based Sampling Strategy:

This strategy enhances negative interaction selection by leveraging sequence similarity among viral proteins, introduced in the Denovo study [37]. The process begins by computing the global alignment BitScore between a target viral protein and all other viral proteins using the Needleman-Wunsch algorithm with the BLOSUM30 matrix. Outliers are removed to improve reliability. The remaining BitScores are then normalized, and their complement is used to determine dissimilarity distances. Suppose the dissimilarity between the target viral protein and another viral protein is greater than the threshold of 0.8. In that case, the human proteins that interact with the other viral protein are considered unlikely to interact with the target viral protein. These proteins are added to the negative sample pool. Finally, a random selection process extracts a specified number of negative samples.

3.10.3 Wei's Sampling Strategy

The Wei's sampling strategy, proposed by Wei *et al.* [14], adopts a clustering-based method that merges both positive (P) and unlabeled (U) protein pairs into a single dataset $T = P \cup U$. The combined dataset T is partitioned into K clusters $\{C_1, C_2, \ldots, C_K\}$ using the MiniBatchKMeans clustering algorithm. For each cluster C_i, the number of unlabeled samples $|U_i|$ and the number of positive samples $|P_i|$ are computed. The unlabeled ratio for each cluster is defined as:

$$r_i = \frac{|U_i|}{|U_i| + |P_i|}$$

Clusters are then sorted in descending order based on r_i, producing a ranking $\{Cr_1, Cr_2, \ldots, Cr_K\}$, where Cr_i denotes the i^{th} cluster in the ranked list. This ranking prioritizes clusters with a higher proportion of unlabeled pairs, under the assumption that such clusters are more likely to contain true negatives. To form the negative sample set N, unlabeled samples are extracted from the top-ranked clusters in order, until the desired number of negative samples $|N|$ is reached. This process ensures that samples are selected from the most unlabeled-dense regions first:

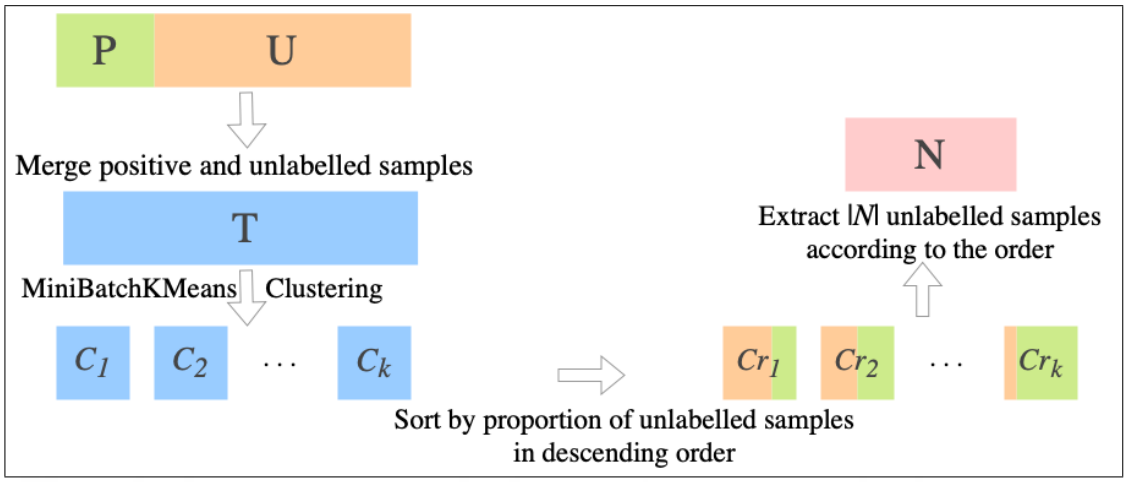
$$N = \text{Top}_{|N|} \left(\bigcup_{i=1}^{K} U_i \mid \text{sorted by } r_i \right)$$

An illustration of this process is shown in Figure 3.1, where the full dataset T is clustered and clusters are ranked based on the ratio of unlabeled to total samples. The top $|N|$ unlabeled samples are then selected from the ranked clusters accordingly.

3.10.4 Wang's Sampling Strategy

This sampling strategy, introduced by Wang *et al.* [51], applies the K-means clustering algorithm to the set of unlabeled protein pairs U, which contains all unknown interaction candidates. The goal is to divide U into K to select proportionally sample negative

Figure 3.1: *Wei's Method.*



examples from each cluster.

Let U be the full set of unlabeled samples, and N be the total number of negative samples to be selected. After clustering, each cluster C_k contains a subset of unlabeled pairs. Denoting the number of unlabeled samples in cluster C_k as $|C_k|$, the number of negative samples to be selected from that cluster, s_k, is computed proportionally as:

$$s_k = \frac{|C_k|}{|U|} \cdot |N|$$

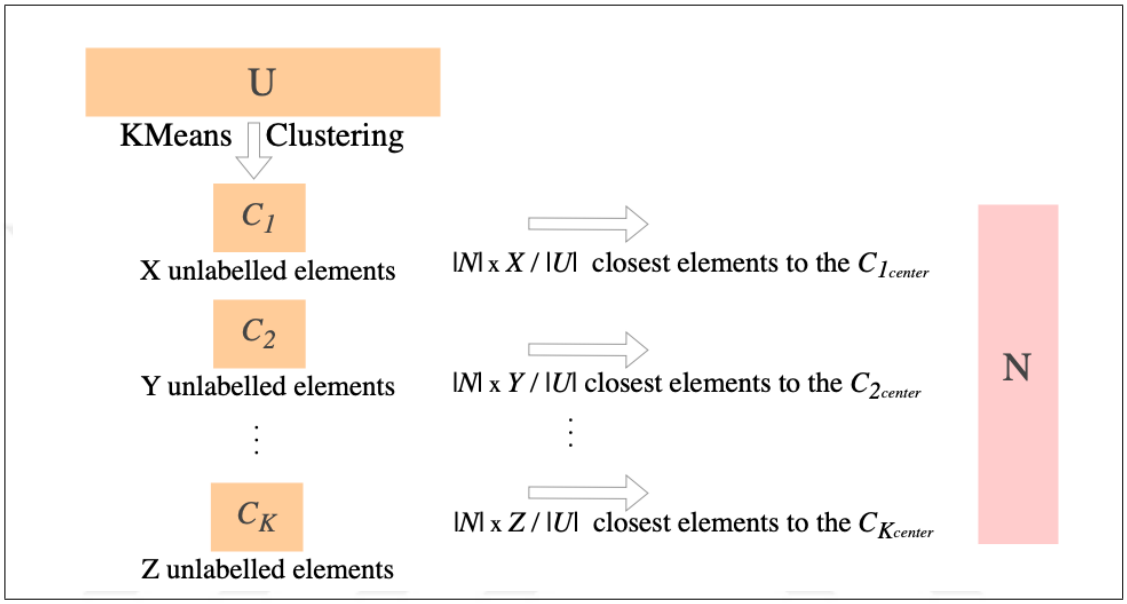
This proportional strategy ensures that the number of selected negative samples from each cluster reflects its relative size in the unlabeled set. The total number of selected samples from all clusters satisfies:

$$\sum_{k=1}^{K} \frac{|C_k| \cdot |N|}{|U|} = |N|$$

This formulation guarantees that exactly $|N|$ negative samples are selected across all clusters. Within each cluster, the s_k unlabeled samples that are closest to the cluster centroid $C_{k\text{-center}}$ are selected as negative examples based on Euclidean distance in the embedding space.

The method is illustrated in Figure 3.2, where the unlabeled set U is partitioned into clusters $C_1, C_2, \ldots, C_K$ containing $X, Y, \ldots, Z$ elements respectively. Each cluster contributes to the negative set N proportionally, and the closest elements to the corresponding centroids are selected.

Figure 3.2: *Wang's Method.*



3.11 Distance Metrics

To quantify the relationship between each unlabeled vector u and the set of cluster centroids $\{C_1, C_2, \ldots, C_K\}$, we compute their pairwise distances using the following distance metrics. Here, $u \in \mathbb{R}^m$ denotes an unlabeled vector, and $C_{k\text{-center}} \in \mathbb{R}^m$ denotes the centroid of the k-th cluster, where m is the number of features (dimensions).

Euclidean distance measures the straight-line distance between two points in m-dimensional space as shown in Equation 3.3. It is commonly used in clustering due to its simplicity and geometric interpretation.

$$d_{\text{euclidean}}(C_{k\text{-center}}, u) = \sqrt{\sum_{i=1}^{m} (C_{k\text{-center}}[i] - u[i])^2} \tag{3.3}$$

Cosine distance is derived from cosine similarity and measures the dissimilarity between two vectors based on the angle between them as shown in Equation 3.4. It is equal to $1 - \cos(\theta)$, where θ is the angle between the vectors. A value closer to 0 indicates higher similarity, while values closer to 1 reflect greater dissimilarity. The dot product in the numerator measures the alignment between the vectors, while the denominator normalizes them by their magnitude.

$$d_{\text{cosine}}(C_{k\text{-center}}, u) = 1 - \frac{C_{k\text{-center}} \cdot u}{\|C_{k\text{-center}}\|_2 \cdot \|u\|_2} = 1 - \frac{\sum_{i=1}^{m} C_{k\text{-center}}[i] \cdot u[i]}{\sqrt{\sum_{i=1}^{m} C_{k\text{-center}}[i]^2} \cdot \sqrt{\sum_{i=1}^{m} u[i]^2}} \tag{3.4}$$

Canberra distance computes normalized absolute differences across all dimensions as shown in Equation 3.5. It is more sensitive to small changes in lower-valued dimensions and has been shown to perform well in biological data analysis [53, 54].

$$d_{\text{canberra}}(C_{k\text{-center}}, u) = \sum_{i=1}^{m} \frac{|C_{k\text{-center}}[i] - u[i]|}{|C_{k\text{-center}}[i]| + |u[i]|} \tag{3.5}$$

4. PROPOSED NEGATIVE SAMPLING STRATEGY

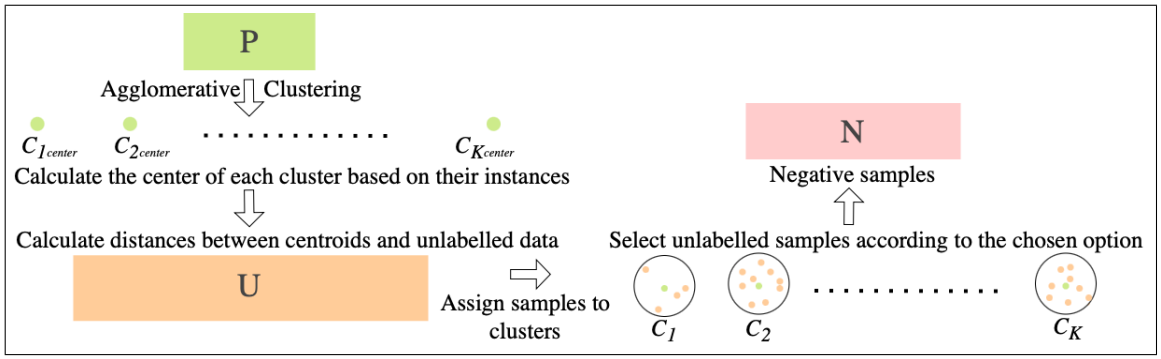
In this chapter, we present a novel cluster-based approach for negative sampling. Then, we explore existing cluster-based negative sampling methods in the literature to compare them with the widely used random sampling method. The intuition of our method is to avoid selecting an unknown positive sample as a negative sample as much as possible. We aim to select unknown samples with different characteristics by forming clusters of positive samples and selecting a few examples from each cluster. To see how the similarity of the negative samples to the positive samples is affected, we investigate four different alternative selection mechanisms in which the samples are selected from the clusters depending on their distance from the centroid of the positive sample clusters. These mechanisms include selecting the closest samples, the farthest samples, samples selected uniformly from cluster centroids, and selecting both the closest and the farthest samples. We empirically compare our approach with the existing cluster-based selection approaches, the random sampling method, and dissimilarity-based method.

4.1 Clustering-based Negative Sampling Strategy (CNS)

This section presents the proposed *Clustering-based Negative Sampling (CNS)* strategy, which aims to improve the reliability of negative sample selection from unlabeled protein pairs. CNS consists of a four-step process that combines clustering and distance-based techniques to reduce the risk of selecting unknown positive interactions as negative examples. As outlined in Algorithm 4.2, the CNS methodology follows a four-step process combining clustering and distance-based selection. The overall workflow of the CNS methodology is illustrated in Figure 4.1, which outlines the sequential steps of the process: (i) clustering of known positive samples, (ii) assignment of unlabeled samples to the nearest cluster centroids, (iii) computation of distances within each cluster, and (iv) selection of negative samples based on predefined strategies. Each step is described

in detail below.

Figure 4.1: *The workflow of CNS.*



- **Step 1: Clustering of Positive Samples and Centroid Calculation**

  To reduce randomness in the sampling process, we apply *Agglomerative Clustering* [48] to the set of known positive protein-protein interactions. Agglomerative Clustering is an unsupervised data mining technique used to create a hierarchy of clusters. It operates in a bottom-up manner, where each sample starts as its cluster, and clusters are progressively merged based on similarity, resulting in a hierarchy. Unlike K-means, Agglomerative clustering is deterministic, meaning it produces the same result for repeated runs on the same data. This consistency enhances the explainability and reproducibility of the clustering process. First, the set of positive samples P is divided into k clusters, denoted as $C_1, C_2, \ldots, C_k$, such that:

$$P = \bigcup_{n=1}^{k} C_n$$

  For each cluster C_n, we calculate the *centroid* $C_{n\text{-center}}$, which is the average vector of all samples in that cluster. These centroids serve as representative points summarizing the location and structure of positive examples in the feature space.

- **Step 2: Assignment of unlabeled Samples to the Closest Centroid**

Inspired by the principles of K-means clustering, this step assigns each *unlabeled protein pair* $u \in U$ to the cluster whose centroid is most similar, based on a selected distance metrics as in described 3.11. This proximity-based assignment helps group unlabeled samples with positive samples they are most similar to in the feature space.

Let $C_{k\text{-center}} \in \mathbb{R}^m$ denote the centroid vector of the k^{th} cluster, and $u \in \mathbb{R}^m$ denote the feature vector of an unlabeled host-pathogen protein pair. Both vectors are obtained by concatenating the Doc2Vec embeddings of the host and pathogen protein sequences.

For each unlabeled sample $u \in U$, the distance to all cluster centroids $C_{k\text{-center}}$, where $k \in \{1, \ldots, K\}$, is computed using the selected distance metric. The sample is then assigned to the cluster whose centroid is closest:

$$m = \arg \min_{k \in \{1,\ldots,K\}} d(C_{k\text{-center}}, u)$$

Once the closest centroid is identified, the sample u is added to the corresponding cluster C_m, thereby updating its membership:

$$C_m = C_m \cup \{u\}$$

Note that while centroids $C_{k\text{-center}}$ remain fixed (since they are computed only from positive interactions) cluster memberships C_k are dynamically updated by including assigned unlabeled samples.

This exhaustive assignment procedure ensures that each unlabeled sample is grouped with the positive interactions it most closely resembles in the feature space, providing a meaningful structure for subsequent distance-based ranking and negative sampling.

- **Step 3: Measuring Distances Within Clusters**

  Following the assignment of unlabeled samples to their respective clusters, the next step involves computing the distance between each unlabeled sample and the centroid of the cluster to which it has been assigned. This calculation provides a quantitative measure of the similarity (or dissimilarity) between the unlabeled sample and the representative structure of the known positive interactions within that cluster.

  The same distance metric employed during the assignment phase (such as Euclidean distance, Canberra distance, or cosine distance) is used consistently in this stage to maintain methodological coherence. The resulting distance values are subsequently utilized in the next step to rank the unlabeled samples within each cluster, thereby facilitating the selection of the most suitable candidates to be used as negative samples.

- **Step 4: Selection of Unlabeled Samples**

  In the final phase of the CNS methodology, unlabeled samples assigned to each cluster are ranked based on their distance to the corresponding cluster centroid. A selected distance metrics, such as Euclidean, Canberra, or cosine distance, is used to compute this proximity as described in Section 3.11. Within each cluster, the samples are sorted in descending order of distance, such that those farthest from the centroid appear later in the ranking. This ordering reflects each sample's dissimilarity to the representative positive samples in that cluster.

  To ensure that the selection of negative samples is balanced across clusters, the number of samples chosen from each cluster is made proportional to its size relative to the overall set of unlabeled samples. Let C_i denote the i^{th} cluster, and U_i represent the number of unlabeled samples in C_i. Let $|U|$ be the total number of unlabeled samples across all clusters, and N the total number of negative samples

to be selected. Then:

$$N = |\text{Positive Samples}| \times \text{Positive-to-Negative Ratio}$$

To account for the representation of sample pairs, a scaling factor is applied to the denominator. The number of negative samples to be selected from cluster C_i, denoted as s_i, is computed as:

$$s_i = \frac{U_i}{|U|} \cdot N \tag{4.1}$$

This proportional allocation ensures that each cluster contributes to the negative sample set in accordance with its size, promoting fairness and reducing sampling bias. As a sanity check, the total number of selected samples across all clusters satisfies:

$$\sum_{i=1}^{k} s_i = \sum_{i=1}^{k} \left(\frac{U_i}{|U|} \cdot N \right) = \frac{N}{|U|} \sum_{i=1}^{k} U_i = \frac{N}{|U|} \cdot |U| = N \tag{4.2}$$

Thus, the sum of s_i over all clusters equals the total number of negative samples N, verifying the correctness of the proportional allocation scheme.

To investigate how sample similarity affects negative sampling quality, four alternative selection strategies are explored. For each strategy, s_i samples are selected from cluster C_i as follows:

- **Closest Selection:** Chooses the s_i samples that are nearest to the cluster centroid.
- **Farthest Selection:** Chooses the s_i samples that are farthest from the centroid.
- **Closest and Farthest Selection:** Selects $\frac{s_i}{2}$ samples from the closest and $\frac{s_i}{2}$ from the farthest ends of the ranked list.

Figure 4.2: *Clustering-based Negative Sampling (CNS) algorithm.*

Require:
P: Set of known positive samples
U: Set of unlabeled samples
k: Number of clusters
$d(\cdot, \cdot)$: Distance metric (e.g., Euclidean, Canberra, Cosine)
N: Total number of negative samples to select
Strategy: Selection strategy (Closest, Farthest, Both, Uniform)
Ensure:
N selected negative samples from U
Step 1: Clustering of positive samples
Apply Agglomerative Clustering on P to form k clusters: $C_1, C_2, ..., C_k$
Compute centroids $C_{1\text{-center}}, ..., C_{k\text{-center}}$
Step 2: Assign unlabeled samples to clusters
for all $u \in U$ **do**
    Compute distances $d(C_{i\text{-center}}, u)$ for all $i = 1, \ldots, k$
    Assign u to the cluster C_m with minimum distance
    $C_m \leftarrow C_m \cup \{u\}$
end for
Step 3: Compute distances within clusters
for all $u \in C_i$ **do**
    Compute distance $d(C_{i\text{-center}}, u)$
end for
Step 4: Select negative samples
for all clusters C_i **do**
    Compute $s_i = \frac{|U_i|}{|U|} \cdot N$ ▷ Number of negatives from cluster i
    Rank unlabeled samples in C_i by distance to $C_{i\text{-center}}$
    if *Strategy* = Closest **then**
        Select s_i samples with smallest distances
    else if *Strategy* = Farthest **then**
        Select s_i samples with largest distances
    else if *Strategy* = Both **then**
        Select $\frac{s_i}{2}$ closest and $\frac{s_i}{2}$ farthest samples
    else if *Strategy* = Uniform **then**
        Select equal parts from top, middle, and bottom of the ranked list
    end if
end for
return Selected negative samples from all clusters

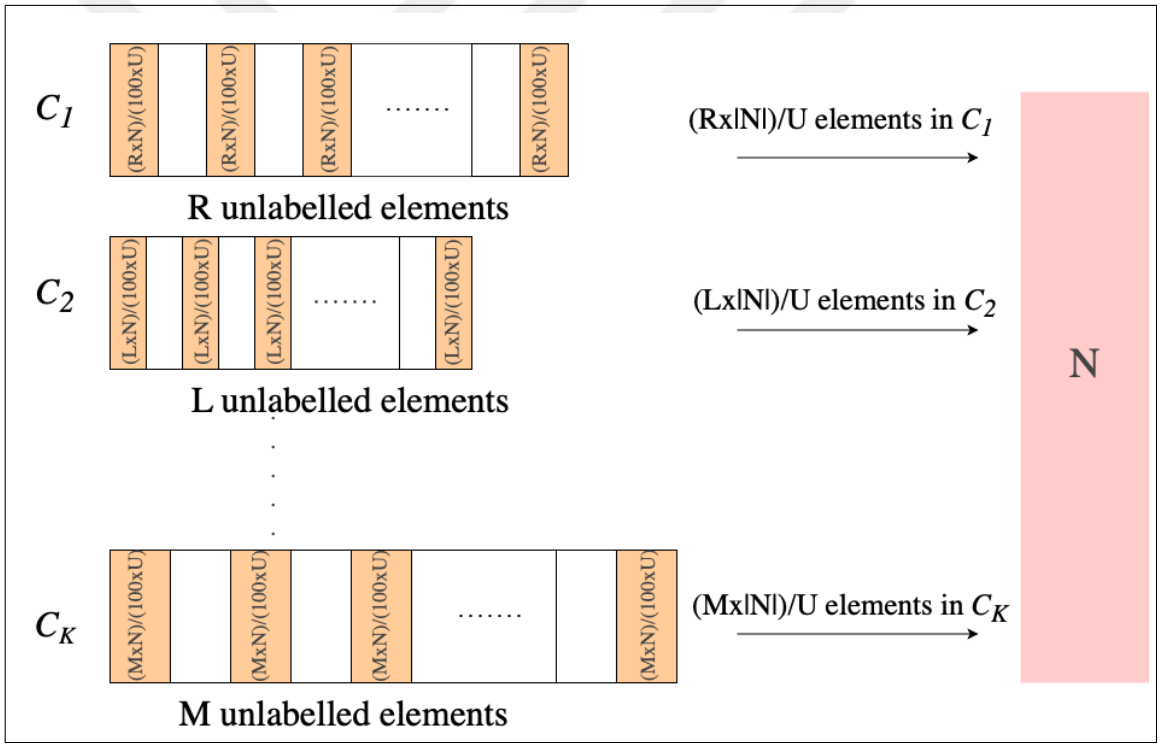
- **Uniform Selection:** Divides s_i into three equal parts: one-third from the top (most similar), one-third from the middle (moderately similar), and one-third from the bottom (most dissimilar) sections of the sorted list. This approach captures a more

diverse sample distribution while preserving proportional representation across clusters. An illustration of this selection mechanism is provided in Figure 4.3. The sets $R, M, \ldots, L$ represent unlabeled samples within each cluster, where $R + M + \ldots + L = |U|$. The equation

$$\frac{R \cdot |N|}{|U|} + \frac{M \cdot |N|}{|U|} + \ldots + \frac{L \cdot |N|}{|U|} = |N|$$

ensures proportional selection from each set, promoting balance and reliability across the full unlabeled dataset.

Figure 4.3: *The methodology of uniform selection.*



These strategies allow us to test how the relative position of unlabeled samples within their clusters (i.e., near or far from positive examples) influences their reliability as negative samples. This also supports a deeper understanding of the relationship between sample similarity and sampling quality.

4.2 Experimental Setup and Evaluation

To assess the effectiveness of the proposed negative sampling strategies and their impact on protein-protein interaction prediction, we conducted a comprehensive set of experiments across four datasets, five sampling settings, and four classification models.

4.2.1 Protein Interaction Datasets

We used four publicly available pathogen-host protein-protein interaction datasets to evaluate the effectiveness of the proposed sampling strategies under different biological conditions.

- **Phisto dataset.** The main dataset used in this study is the Pathogen–Host Interaction Search Tool (PHISTO)[1]) dataset [55], which includes a comprehensive collection of $39,544$ positive interactions between pathogens and hosts. This dataset includes $6,571$ viral proteins, $1,715$ human proteins, and $11,229,721$ unlabeled interactions. PHISTO contains interactions involving $3,879$ unique host proteins and 578 unique pathogen proteins.

- **Tsukiyama's dataset.** The PPI datasets used by Tsukiyama *et al.* [19] were obtained from the Host–Pathogen Interaction Database (HPIDB) [56]. In their preprocessing, interactions with a Molecular Interaction (MI) score below 0.3 were excluded to retain high-confidence PPIs. Redundant interactions were removed using CD-HIT with a sequence identity threshold of 95%, and only proteins with sequence lengths between amino acid residues 30 and 1000 were selected. This resulted in a dataset comprising $22,383$ positive interactions involving $5,882$ human and 996 viral proteins, along with $6,365,665$ unlabeled protein pairs.

- **Human-Virus dataset.** The human-virus interaction dataset, originally compiled

[1] `http://www.phisto.org`

in the DeepViral study by Liu-Wei *et al.* [33], was used as a benchmark dataset in the DeepTrio study [57]. In this study, we adopt the same dataset, which contains $8,929$ positive interactions with $3,363$ human proteins and 360 viral proteins and $1,201,751$ unlabeled protein pairs.

- **Herpes dataset.** This dataset, curated by [38], was compiled from five public molecular interaction databases: IntAct, BioGRID, HPIDB, VirHostNet, and Virus-Mentha. Herpes dataset consists of $9,339$ positive protein-protein interactions (PPIs) and $2,232,723$ unlabeled interactions involving seven herpesvirus types (HHV-1 to HHV-6 and HHV-8) by removing non-physical, redundant, and non-standard interactions. It includes $3,879$ human proteins and 578 viral proteins.

4.2.2 Positive-to-Negative Ratio

In protein-protein interaction (PPI) studies, the positive-to-negative ratio refers to the proportion of non-interacting protein pairs (negatives) to interacting protein pairs (positives) within a dataset. The choice of the positive-to-negative ratio can significantly affect the model performance and evaluation reliability. Therefore, handling this ratio is essential for building robust and biologically meaningful PPI prediction models.

A balanced ratio, such as $1:1$, can improve the sensitivity of the model in recognizing true interactions but may not reflect the real-world distribution, where non-interacting pairs vastly outnumber interacting ones. In biological data, positive samples are typically much fewer than negative samples. This class imbalance can bias machine learning models toward predicting the majority class (negatives) more accurately, potentially leading to poor performance in the minority class (positives). Some studies adopt a $1:1$ ratio during training to mitigate class imbalance and improve model learning, although this setting does not accurately represent the actual distribution of interactions in biological systems.

Conversely, using a highly unbalanced ratio better captures biological reality but can make models biased toward predicting non-interactions. That is why a positive-to-negative ratio of $1:10$ or higher is standard in many PPI studies. In this thesis, a $1:10$ positive-to-negative ratio is adopted during training and evaluation.

4.2.3 Evaluation Protocol

To evaluate the performance of the prediction models in a reliable and generalizable way, we performed a 5-fold cross-validation. In this approach, the entire data set is divided into five equally sized subsets. In each iteration, four subsets are used for training and the remaining subset is reserved for testing. This process is repeated five times, ensuring that each subset is used once as a test set. The final performance metrics are obtained by averaging the results across all five foldings. This technique reduces the variance associated with a single split into training and testing and provides a more robust assessment of model performance.

4.2.4 Experimental Framework

This experimental framework builds on the PPI prediction architecture proposed by Koca et al. [8], that integrates sequence-based embeddings and topological information to predict host-pathogen-protein-protein interactions. In this study, we adopt their general modeling pipeline and extend it by systematically evaluating and improving the negative sampling step, which is the main contribution of this thesis.

- **Preparation of the input data:** The raw amino acid sequences of host and pathogen proteins are first preprocessed by applying Byte Pair Encoding (BPE), which segments the sequences into frequent subunits. Separate BPE vocabularies are created for each group of organisms. The tokenized sequences are then used to train a Doc2Vec model that embeds each protein into a 128-dimensional vector. For each

protein pair, the embeddings of the host and pathogen proteins are concatenated into a single 256-dimensional vector representing the interaction.

- **Integration of graph information with GraphSAGE:** To enrich sequence-based representations with topological context, the GraphSAGE framework [50] is used. It aggregates features of neighboring nodes in the protein interaction graph and generates hybrid embeddings that contain both sequence and network information.

- **Generation of negative samples (Our contribution):** One of the most critical challenges in PPI prediction is the lack of experimentally verified negative interactions. In this thesis, we focus on evaluating and improving negative sampling strategies that select reliable non-interacting pairs from unlabeled data. This step represents the main contribution of our work.

  We compare widely used sampling strategies such as random sampling, dissimilarity-based method and existing clustering-based methods (e.g., Wang and Wei's approaches) with our proposed method, clustering-based negative sampling (CNS). All sampling methods are applied after generating the embedding to ensure that the decisions are based on meaningful representations. A fixed 1:10 ratio of positives to negatives is used in all settings, reflecting biological imbalance and allowing fair comparisons between strategies.

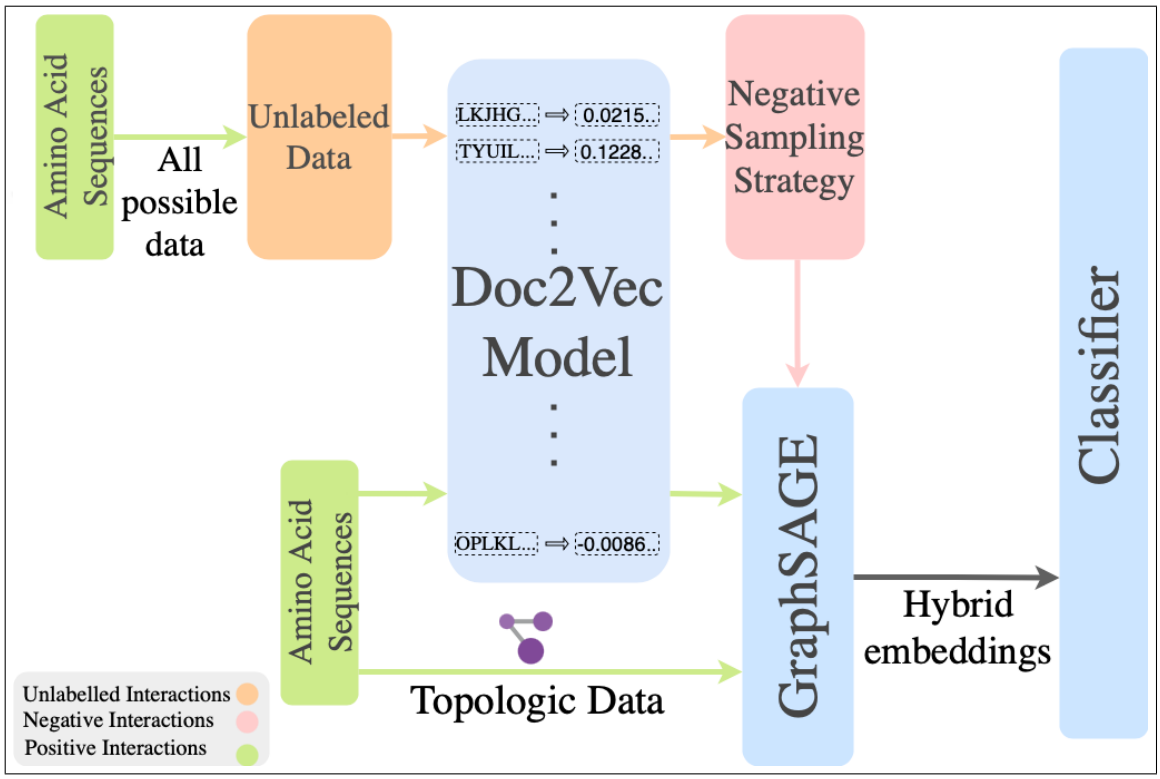
- **Creation of the classification models:** Four classifiers are trained with the labeled data: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Generalized Additive Model with pairwise interactions (GA^2M). These models were selected for their robustness, interpretability and ability to process high-dimensional biological data.

- **Model Evaluation:** To evaluate the performance of the model, we use 5-fold cross-validation, in which the dataset is divided into five parts and each part is used once

as a test set. This approach provides a reliable estimate of the average performance.

The performance of the model is measured using common classification metrics. In particular, we report recall to reflect the ability to correctly identify true interactions, F1-score to balance precision and recall.

Figure 4.4: *Step-by-step architecture of the PPI prediction framework*



4.3 Results

This section presents the experimental results of evaluating negative sampling strategies under various configurations. We analyze how factors such as distance metrics, number of clusters, and the choice of classifier affect the performance of the proposed and baseline methods across multiple datasets.

4.3.1 *Effect of Distance Metrics on Method's Performance*

The performance of the proposed sampling methods may vary depending on parameters such as the number of clusters, distance metrics, and selection mechanisms. We conducted further analyses to understand the impact of these parameters. First, the impact of distance metrics on the performance of our method was evaluated using the PHISTO dataset. In this phase, the GA^{2}M classifier was selected aligning with the work of Koca *et al.* [8]. Fig. 4.5 shows the recall metric (i.e., the percentage of true positive predictions compared to the confirmed positive samples) where the y-axis represents the selection criterion, the x-axis shows the number of clusters (k) and the type of distance metric used. As can be seen in Figures 4.5 and 4.6, the choice of distance metric has significant effects on recall and F1 scores. Detailed results for all cluster sizes and distance metrics are provided in Table A.1. Regardless of the number of clusters, it can be seen that the Canberra distance performs better than other methods with the farthest selection. Based on these results, the Canberra distance was used as the distance metric for our method in the further experimental studies below.

Figure 4.5: *Performance metrics of GA^{2}M on the PHISTO dataset: Recall scores.*

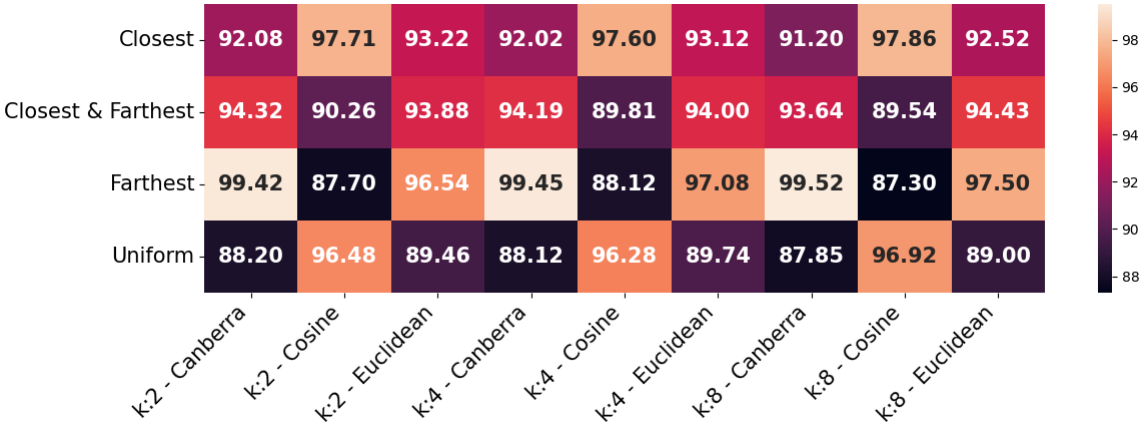


Figure 4.6: *Performance metrics of GA^{2}M on the PHISTO dataset: F1 scores.*

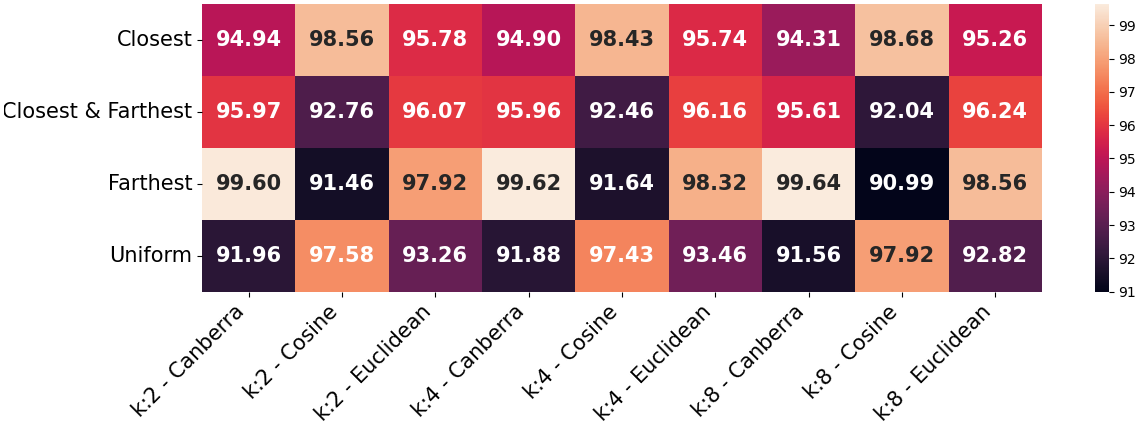


Table 4.1: *The F1 and recall values in percent for the PHISTO dataset with different cluster numbers.*

Cluster Numbers	$k = 2$		$k = 4$		$k = 8$	
	F1	Recall	F1	Recall	F1	Recall
Closest	**94.94**	**92.08**	94.90	92.02	94.31	91.20
Closest & Farthest	**95.97**	**94.32**	95.96	94.19	95.61	93.64
Farthest	99.60	99.42	99.62	99.45	**99.64**	**99.52**
Uniform	**91.96**	**88.20**	91.88	88.12	91.56	87.85
Wang's method	95.87	93.36	**96.18**	**93.95**	94.28	91.58
Wei's method	97.73	95.81	97.84	96.08	**99.22**	**98.90**

4.3.2 Effect of Cluster Numbers

We analyzed the effects of different cluster numbers, including 2, 4, and 8, on the PHISTO dataset. As can be seen in Table 4.1, the optimal cluster number varied depending on the method. For example, the best performance for the Closest, Closest and Farthest, and Uniform methods was achieved with 2 clusters, while the Farthest method performed best with 8 clusters. Similarly, Wang's method achieved the highest F1 and recall values with 4 clusters, while Wei's method showed the best results with 8 clusters. Consequently, the optimal number of clusters for each method was fixed based on their performance in the following analyses.

4.3.3 Analysis of Classifier Performance

In this section, we assess the performance of different classifiers with the sampling strategies. Table 4.2 shows the F1 and recall values of each classifier. We used the classifiers aligned with Koca's study [8], including Random Forest (RF), Logistic Regression (LR), SVM and Generalized Additive 2 Model (GA^{2}M) as explained in Section 3.4. The results show that the proposed approach with the farthest selection slightly outperforms the other strategies. Of the four classifiers, the Random Forest classifier outperformed the others, showing slightly better results than GA^{2}M. Therefore, we advocated using the Random Forest. The detailed performance results across different cluster sizes and distance metrics for each dataset and classifier are provided in Tables A.1 (GA^{2}M), A.3 (RF), A.4 (LR), and A.5 (SVM).

Table 4.2: *The F1 and recall values in percent for the PHISTO dataset with different classifiers and the optimal cluster number for each method.*

Classifiers	GA^{2}M		LR		SVM		RF	
	F1	Recall	F1	Recall	F1	Recall	F1	Recall
Closest	**94.94**	**92.08**	92.76	89.30	94.33	91.28	94.90	91.46
Closest & Farthest	95.97	**94.32**	93.42	91.18	95.97	94.20	**96.34**	94.05
Farthest	**99.64**	**99.52**	99.49	99.24	99.58	99.40	**99.64**	99.44
Uniform	91.96	**88.20**	88.60	83.67	91.30	87.05	**92.36**	87.72
Wang's method	96.18	**93.95**	94.00	90.76	95.45	92.96	**96.44**	93.91
Wei's method	99.22	98.90	98.88	98.37	98.99	98.36	**99.36**	**99.10**
Random sampling	73.32	67.94	62.34	54.44	70.97	63.06	**79.39**	**73.52**
Dissimilarity-based	71.78	66.45	54.78	46.00	69.07	60.36	78.81	72.50

4.3.4 Performance Across Other Datasets

To evaluate the effectiveness of the sampling strategies across different pathogen-host protein-protein interaction datasets, we further analyze the performance of the Random Forest classifier on three additional datasets, as described in Section 4.2.1. Table 4.3 presents the results for cluster numbers 2, 4, and 8 across all datasets. Based on the

results in Table 4.3, Wei's method consistently achieved the highest performance across most datasets and cluster numbers for Random Forest. In the PHISTO dataset, our method with the farthest selection outperformed other methods and reached an F1 score of 99.64% with Random Forest at $k = 8$. Detailed performance tables for each dataset (PHISTO, DeepTrio, Tsukiyama, and Herpes) and cluster configuration are provided in Tables A.3, A.8, A.12, and A.14.

Table 4.3: *Performance of the Random Forest classifier across all datasets with different cluster numbers.*

Datasets	**Cluster Numbers**	**Farthest**		**Wang's Method**		**Wei's Method**		**Random Sampling**		**Diss.-based**	
		F1	**Recall**	**F1**	**Recall**	**F1**	**Recall**	**F1**	**Recall**	**F1**	**Recall**
PHISTO	$k = 2$	**99.62**	**99.30**	95.96	93.01	97.15	94.71	79.39	73.52	78.81	72.50
	$k = 4$	**99.62**	**99.40**	96.44	93.91	97.39	95.36				
	$k = 8$	**96.64**	**99.44**	94.53	90.94	99.36	99.10				
Tsukiyama et al.	$k = 2$	96.02	92.68	96.17	93.06	**96.62**	**93.78**	68.84	59.80	66.44	56.63
	$k = 4$	95.85	92.62	96.42	93.44	**96.50**	**93.39**				
	$k = 8$	96.06	92.62	93.78	89.24	**97.60**	**95.54**				
Human-Virus	$k = 2$	92.34	87.04	92.12	86.52	**95.10**	**91.36**	72.32	64.16	65.47	55.60
	$k = 4$	92.06	86.88	89.76	83.53	**96.54**	**93.68**				
	$k = 8$	92.31	87.12	89.00	82.48	**97.52**	**95.74**				
Herpes	$k = 2$	94.57	90.86	95.11	91.32	**96.39**	**93.39**	66.70	54.12	70.57	60.19
	$k = 4$	94.67	90.58	**94.90**	**90.90**	94.67	90.54				
	$k = 8$	94.52	90.62	92.33	86.82	**97.97**	**96.41**				

It is worth noting that there is no superior sampling strategy resulting in terms of prediction accuracy in all datasets. On the other hand, the results show that all cluster-based sampling approaches outperformed random sampling and dissimilarity-based methods. Therefore, we suggest that the clustering-based method is preferable to the random sampling and dissimilarity-based methods in this domain.

5. PROPOSED EVALUATION METHODOLOGY

To the best of our knowledge, none of the current studies investigates the risk of considering unlabeled positive instances as negative and its impact on the performance of the model. To address this gap, we propose an evaluation method in which certain positive samples are intentionally categorized as unlabeled (unknown) instances. This design enables an analysis of the extent to which positive instances are mistakenly included in the negative sample set by various sampling strategies. Furthermore, it allows us to assess the impact of such mislabeling on the predictive performance of the model, particularly its ability to correctly identify unknown positive interactions.

5.1 R-SAFE: Reliability of Sampling to Avoid False Examples

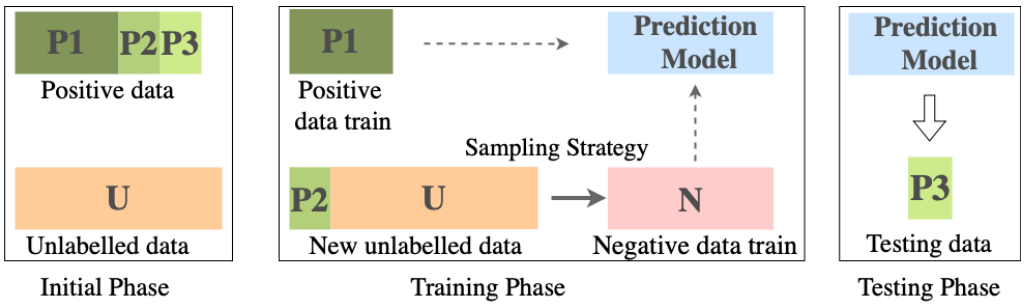
We developed a structured R-SAFE method to evaluate the performance of different negative sampling strategies to predict host-pathogen protein-protein interactions (PPI). Our method involves intentionally masking a set of known positive interactions to imitate them as *unknown* during the evaluation process. This allows us to test how well each negative sampling strategy avoids selecting these imitated unknown positives as part of the negative set. By doing so, we can directly assess the ability of each strategy to correctly exclude true positives from being misclassified as negatives. This approach is crucial to understanding which strategies are the most effective in minimizing false negatives while maintaining the overall accuracy of the PPI predictions. A detailed explanation of each phase in the evaluation process is shown in Figure 5.1.

The methodology begins with a carefully curated dataset of experimentally verified positive interactions, which are split into three subsets: $P1$, $P2$, and $P3$. Each subset plays a unique role in ensuring the robustness of the evaluation:

- **Training Set** ($P1$): This subset provides the positive interactions that the model

uses for training. By including these known positive interactions and selected negatives in the training phase, we aim to build an understanding of the model for interaction patterns. This allows the model to effectively distinguish between positive interactions and non-interacting pairs.

Figure 5.1: *Proposed Evaluation Methodology.*



- **Imitated Unknown Positives** ($P2$): The $P2$ subset plays a key role in simulating unknown positive interactions. By incorporating $P2$ into the unlabeled set, we imitate real-world scenarios in which some true interactions have not yet been experimentally confirmed. This allows us to evaluate how effectively each sampling strategy avoids misclassifying these unknown positives as negatives.

- **Test Set** ($P3$): This subset is reserved exclusively for the final model evaluation after the training phase. By excluding $P3$ from the training process, we enable an unbiased evaluation of the model, as it helps measure how well the model generalizes to interactions that were not included during training.

After dividing the data, we proceed to construct the unlabeled set, U, by pairing all proteins from the training set $P1$ and removing any pairs that are known to interact. This process generates a set of unlabeled samples by excluding the positive interaction pairs from $P1$. In this phase, we add $P2$ to the unlabeled set U to imitate unknown positive interactions and reflect the uncertainty present in real-world datasets, where some true

interactions remain unverified due to the lack of experimental evidence. This extended unlabeled set allows us to observe whether each sampling strategy mistakenly classifies unknown positive samples as negatives.

After merging U and $P2$, the negative samples N are chosen from this set with the negative sampling strategies outlined in Section 3.10. The choice of sampling method for the generation of N has a direct impact on the model learning process. If the sampling strategy selects true negatives without introducing false positives from $P2$, the model can distinguish more clearly between positive and unlabeled pairs. In contrast, a strategy that tends to select unknown positives as negatives can lead to biases that reduce the sensitivity of the model and its overall predictive accuracy.

Finally, the model is then trained using the positive samples $P1$ and the negative samples N, and its performance is evaluated in the test set $P3$ using key metrics to assess the effectiveness of each sampling strategy:

- **P.C. ratio (%)**: This metric represents the ratio of imitated positive samples from $P2$ that are incorrectly selected as negatives during the sampling process. The Positive Count (P.C.) ratio offers a normalized perspective by expressing this number as a percentage of the total size of $P2$. A lower P.C. ratio indicates a more effective and reliable sampling strategy, as it reduces the likelihood of contaminating the negative set with actual or potential positives (i.e., false negatives). It is calculated as follows:

$$\textbf{P.C. ratio (\%)} = \frac{\text{Number of incorrectly selected positives from } P2}{\text{Total number of samples in } P2} \times 100$$

This ratio allows for a fair comparison across different datasets or sampling scenarios with varying sizes of $P2$, providing a more interpretable measure.

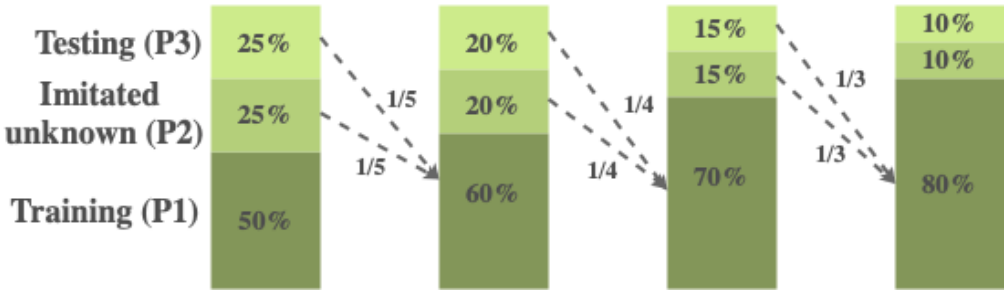
- **Prediction Scores on P3**: The performance of the model on known positive interactions in the test set $P3$ is evaluated using the recall, as described in Section 3.5. These metrics reflect the overall effectiveness of the sampling strategy in supporting reliable prediction performance.

5.2 Evaluation of Proposed Approaches

To evaluate the R-SAFE model across different data distributions, we defined four incremental split ratios: $80:10:10$, $70:15:15$, $60:20:20$, and $50:25:25$, representing the proportions of training, test, and unlabeled sets, respectively. That enables us to systematically examine the effect of the split ratio. The 80:10:10 split emphasizes learning by allocating a larger portion of data to training, thereby increasing the number of positive interactions available to the model. In contrast, the $50:25:25$ split assigns more data to the test and unlabeled sets, allowing for a broader evaluation of the model performance on data not seen during training. These varying splits provide a comprehensive perspective on the predictive capabilities of the model under different conditions. The process began with the $50:25:25$ split as the baseline, and subsequent splits were incrementally constructed by moving a defined portion of random samples from the test and unlabeled sets, as illustrated in Figure 5.2. By systematically increasing the training set while shrinking the evaluation portions, we reduce the effect of the randomness of the split. For instance, if we create independently the difference in the performance of the data distribution instead of the split ratio. To ensure the results are not biased by a single random split, the dataset-splitting process was repeated 10 times using different random seeds. Consequently, we construct 10 different combinations for each $P1:P2:P3$.

Adopting the evaluation methodology proposed in Section 5.1, we analyzed the existing negative sampling strategies (Section 3.10) for each dataset described in Section 4.2.1. The main performance metric is P.C. ratio (i.e., the ratio of false negatives

Figure 5.2: *The figure illustrates the data splitting process.*



that imitate unknown positives ($P2$) incorrectly selected as negatives). A low P.C. ratio indicates a method's ability to avoid the false labeling of true interactions, which is essential for reliable model training. Accordingly, Table 5.1 shows the results of the benchmark methods consisting of five negative sampling strategies: Dissimilarity-based method, Random sampling, Wang's method, CNS, and Wei's method. Experiments were conducted on four benchmark host-pathogen PPI datasets: Herpes, Tsukiyama's dataset, Human-Virus dataset, and PHISTO. Each method was tested with four different data split ratios: $50:25:25$, $60:20:20$, $70:15:15$, and $80:10:10$, which represent the proportions of training, test, and unlabeled sets, respectively. The P.C. ratios reported in the table represent the average proportion (in percent), and standard deviation of incorrectly labeled negatives in each data set, calculated over 10 independent runs with different random seeds for each data split.

Table 5.1: *P.C. ratios (% $\pm$ standard deviation) for all datasets across different split ratios and methods*

Dataset	Split R.	Dis. Method	Random Samp.	Wang Method	CNS	Wei Method
Herpes	50:25:25	14.07% $\pm$ 0.01	**9.94% $\pm$ 0.03**	10.32% $\pm$ 0.03	15.07% $\pm$ 0.01	38.63% $\pm$ 0.16
	60:20:20	9.54% $\pm$ 0.03	8.44% $\pm$ 0.02	**8.07% $\pm$ 0.02**	11.39% $\pm$ 0.03	30.92% $\pm$ 0.16
	70:15:15	6.98% $\pm$ 0.02	6.36% $\pm$ 0.01	**6.30% $\pm$ 0.02**	8.64% $\pm$ 0.02	19.04% $\pm$ 0.11
	80:10:10	**3.96% $\pm$ 0.01**	4.26% $\pm$ 0.01	4.05% $\pm$ 0.01	5.79% $\pm$ 0.01	7.32% $\pm$ 0.04
Tsukiyama	50:25:25	16.29% $\pm$ 0.01	8.93% $\pm$ 0.02	**5.78% $\pm$ 0.01**	7.34% $\pm$ 0.01	23.74% $\pm$ 0.08
	60:20:20	11.62% $\pm$ 0.03	6.87% $\pm$ 0.02	**4.42% $\pm$ 0.01**	6.59% $\pm$ 0.02	14.45% $\pm$ 0.05
	70:15:15	7.98% $\pm$ 0.02	5.26% $\pm$ 0.01	**3.45% $\pm$ 0.01**	4.84% $\pm$ 0.01	11.07% $\pm$ 0.03
	80:10:10	5.06% $\pm$ 0.01	3.51% $\pm$ 0.01	**2.40% $\pm$ 0.00**	3.32% $\pm$ 0.01	5.94% $\pm$ 0.02
Human-Virus	50:25:25	29.88% $\pm$ 0.03	19.04% $\pm$ 0.06	**17.70% $\pm$ 0.05**	18.46% $\pm$ 0.01	18.55% $\pm$ 0.06
	60:20:20	22.16% $\pm$ 0.07	15.14% $\pm$ 0.04	13.82% $\pm$ 0.04	14.97% $\pm$ 0.04	**13.05% $\pm$ 0.05**
	70:15:15	15.94% $\pm$ 0.05	11.44% $\pm$ 0.03	10.84% $\pm$ 0.03	11.49% $\pm$ 0.03	**7.15% $\pm$ 0.02**
	80:10:10	9.91% $\pm$ 0.03	7.67% $\pm$ 0.02	6.95% $\pm$ 0.02	7.56% $\pm$ 0.02	**5.15% $\pm$ 0.02**
PHISTO	50:25:25	18.43% $\pm$ 0.01	8.61% $\pm$ 0.02	11.76% $\pm$ 0.03	**0.85% $\pm$ 0.00**	13.69% $\pm$ 0.12
	60:20:20	13.38% $\pm$ 0.04	6.88% $\pm$ 0.02	8.94% $\pm$ 0.03	**0.86% $\pm$ 0.00**	73.65% $\pm$ 0.54
	70:15:15	8.90% $\pm$ 0.02	5.31% $\pm$ 0.01	6.74% $\pm$ 0.02	**0.69% $\pm$ 0.00**	11.47% $\pm$ 0.18
	80:10:10	5.26% $\pm$ 0.01	3.54% $\pm$ 0.01	4.62% $\pm$ 0.01	**0.53% $\pm$ 0.00**	3.31% $\pm$ 0.02

At first sight, it can be observed that there is no single negative sampling strategy outperforming others in terms of the chosen metric. While CNS selects the least number of false negatives (i.e. the lowest P.C. ratio) for all split ratios in the PHISTO dataset, Wang's method has the lowest P.C. values in the Tsukiyama's dataset. For the Human-Virus dataset, except $50:25:25$ ratio, Wei's method outperforms other strategies regarding P.C. ratio. Interestingly, there is no regular pattern in the Herpes dataset. At a ratio of $50:25:25$, the random strategy incorrectly selects the least positive samples as negative. On the other hand, Wang's method outperforms others for $60:20:20$ and $70:15:15$ split ratios, whereas Dissimilarity-based method yields the best performance in terms of P.C. ratio for the $80:10.10$ ratio. It would be interesting to investigate which strategy is best suited to which type of datasets. This research question will be studied elaborately in Section 6.

After analyzing the performance of negative sampling strategies in terms of the number of incorrectly selected negative samples defined as P.C. ratio, we have studied their effect on the prediction performance of the classifier. For this purpose, we trained the Random Forest Classifier to predict whether a given pair interacts or not, as explained in Section 4.3.3, and reported the recall value on the test set $P3$, which consists of only known positives. The recall values reported in the Table 5.2 are the average values along with their corresponding standard deviations ($\pm$) calculated over 10 independent runs with different random seeds.

Overall, the Random Forest algorithm achieves the highest average recall values when we use Wei's method in negative sampling for almost all datasets and splits except the PHISTO dataset. Notably, it reached recall values of 0.941 ± 0.037 on Herpes, 0.911 ± 0.052 on Tsukiyama, and 0.928 ± 0.037 on Human-Virus at the $80:10:10$ split. On the other hand, the classifier yields the best recall value on the PHISTO dataset (e.g., 0.971 ± 0.018 at $80:10:10$ split) when CNS sampling is employed, except for the $50:$

$25:25$ split setting. For the PHISTO and Human-Virus datasets, the recall performances are aligned with the former results of the negative sampling strategies in terms of P.C. ratios (i.e., high recall with low P.C. ratio). Surprisingly, the recall values obtained in the Herpes and Tsukiyama datasets are conflicting with the results regarding P.C. ratio. That is, although the ratio of false negatives (i.e., P.C. ratio) is high, it could yield a high recall on the test set ($P3$). For example, although Wei's method has one of the highest P.C. ratios, it achieved the highest recall value of 0.911 ± 0.052 in the Tsukiyama's dataset with a ratio of $80:10:10$. This requires further investigation. Somehow, the learning algorithm learns a more general model with the wrong negative samples and performs better in its prediction. This could stem from the complexity of the classifier model and/or the variety of sample distributions in the datasets.

Table 5.2: *Average Recall-scores (± standard deviation) for different negative sampling strategies across benchmark datasets and data splits.*

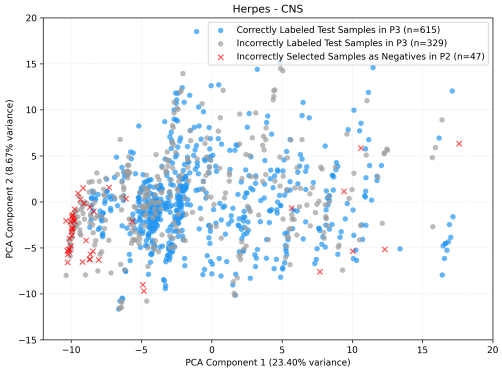
Method /Split R.	Herpes	Tsukiyama's dataset	Human-Virus	PHISTO
Dissimilarity-based/50	0.212 ± 0.013	0.238 ± 0.012	0.478 ± 0.040	0.475 ± 0.007
Random Sampling/50	0.223 ± 0.008	0.361 ± 0.006	0.301 ± 0.011	0.587 ± 0.005
Wang's Method/50	0.717 ± 0.063	0.823 ± 0.020	0.613 ± 0.009	0.861 ± 0.015
CNS/50	0.600 ± 0.018	0.751 ± 0.008	0.653 ± 0.022	0.887 ± 0.024
Wei's Method/50	**0.793** ± 0.162	**0.879** ± 0.081	**0.895** ± 0.099	**0.936** ± 0.058
Dissimilarity-based/60	0.234 ± 0.018	0.273 ± 0.010	0.498 ± 0.042	0.508 ± 0.007
Random Sampling/60	0.239 ± 0.008	0.379 ± 0.006	0.323 ± 0.007	0.602 ± 0.005
Wang's Method/60	0.750 ± 0.047	0.839 ± 0.026	0.631 ± 0.020	0.885 ± 0.012
CNS/60	0.640 ± 0.024	0.765 ± 0.010	0.681 ± 0.032	**0.912** ± 0.031
Wei's Method/60	**0.872** ± 0.101	**0.890** ± 0.073	**0.923** ± 0.023	0.896 ± 0.048
Dissimilarity-based/70	0.261 ± 0.011	0.301 ± 0.010	0.536 ± 0.052	0.534 ± 0.007
Random Sampling/70	0.257 ± 0.015	0.393 ± 0.005	0.340 ± 0.013	0.618 ± 0.003
Wang's Method/70	0.786 ± 0.029	0.859 ± 0.025	0.676 ± 0.027	0.890 ± 0.013
CNS/70	0.669 ± 0.025	0.789 ± 0.011	0.697 ± 0.016	**0.950** ± 0.025
Wei's Method/70	**0.812** ± 0.129	**0.904** ± 0.048	**0.915** ± 0.078	0.945 ± 0.045
Dissimilarity-based/80	0.291 ± 0.023	0.323 ± 0.007	0.533 ± 0.033	0.557 ± 0.007
Random Sampling/80	0.267 ± 0.017	0.409 ± 0.011	0.367 ± 0.020	0.632 ± 0.009
Wang's Method/80	0.817 ± 0.028	0.869 ± 0.019	0.719 ± 0.030	0.888 ± 0.010
CNS/80	0.701 ± 0.042	0.798 ± 0.018	0.732 ± 0.016	**0.971** ± 0.018
Wei's Method/80	**0.941** ± 0.037	**0.911** ± 0.052	**0.928** ± 0.037	0.944 ± 0.045

To better understand the unexpected increase in recall despite higher P.C. ratios in

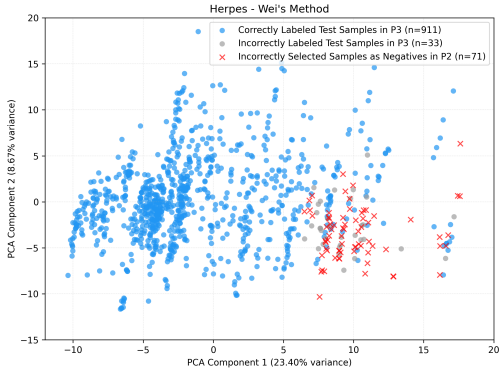
the Herpes and Tsukiyama's datasets, we performed a visual analysis of the distribution of mislabeled samples, as shown in Figures 5.3 and 5.5. Although lower P.C. ratios are generally expected to result in higher recall, our observations on the different datasets show that this relationship may not always exist, especially for the Herpes and Tsukiyama's datasets. To investigate this inconsistency between recall scores and the incorrectly labeled negatives ($P.C.$), we analyzed the results focusing on the $80:10:10$ experimental setting based on Tables 5.1 and 5.2 . We performed a detailed visual analysis of the spatial distribution of $P2$ and $P3$ samples using PCA projections (Figure 5.3–5.6).

Figure 5.3: *PCA visualization for the Herpes dataset using different sampling strategies.*

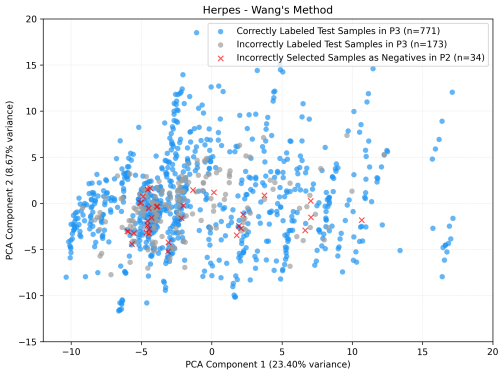
(a): *CNS*
Recall: 0.651



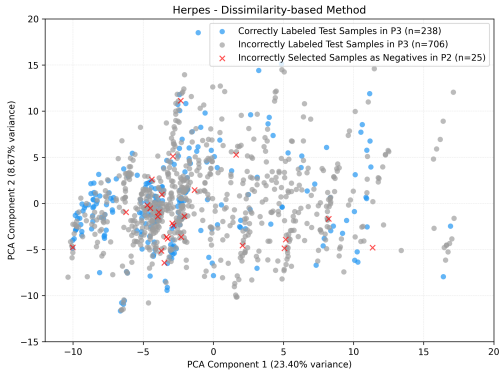
(b): *Wei*
Recall: 0.965



(c): *Wang*
Recall: 0.817



(d): *Diss.-based*
Recall: 0.252



In Herpes dataset, Wei's and CNS sampling strategies show distinct spatial patterns from each other in selecting unknown positives ($P2$) as negatives. Wei's method identifies those samples in regions where the intensity of the positive test samples in $P3$ is relatively lower than that of the areas chosen by CNS. In contrast, CNS tends to select unknown positive samples as negatives within or near to dense positive regions in test set ($P3$), increasing the risk of false positives. Moreover, in Wei's sampling strategy, the mislabeled instances tend to cluster within a specific region, whereas in CNS, we observe that these errors not only concentrate in a particular area but also slightly distribute across multiple distinct regions. Although Wang's sampling strategy selects fewer unknown positives as negative samples, the mislabeled instances it does select tend to lie near the center of the data distribution where the positive test samples ($P3$) are concentrated. This increases misclassification in the test set, which in turn explains why Wei's method achieves a higher recall score. As seen in Figure 5.3, the mislabeled unknown positive samples of dissimilarity-based method are distributed all over the sample space, resulting in more false negatives in test set.

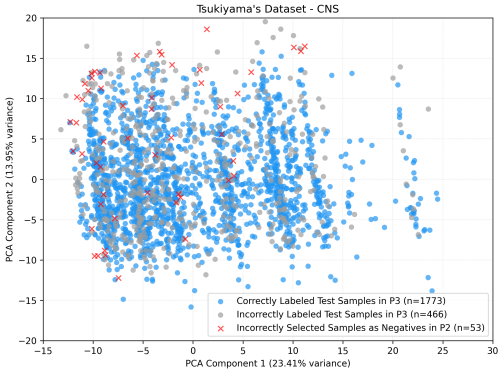
In the Tsukiyama's dataset, Wei's sampling strategy differs from the typical pattern observed in other datasets. As seen in Figure 5.4, the mislabeled $P2$ samples are not located in isolated regions but are clustered within the center of the samples with low density on test samples ($P3$). Therefore, the number of false-negative samples is not much more than expected, resulting in a lower recall value. Other strategies including CNS, Wang's, and dissimilarity-based exhibit the similar pattern as explained above.

In the Human-Virus dataset, the P.C. ratios and recall values are nearly consistent. Wei's and CNS strategies achieve high recall scores; however, their underlying selection patterns differ notably. Wei's method selects mislabeled samples from a more isolated and densely clustered region, while negative sample selection in CNS spreads over an expanse area. In the Phisto dataset, there is a clear correspondence between the mislabeled samples

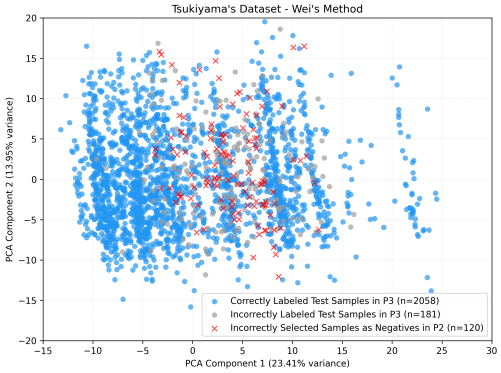
and the recall value. In particular, for the CNS and Wei's methods, the mislabeled samples tend to cluster in sparse regions that are far from the dense regions of positive test samples, making it easier for the model to distinguish them, resulting in higher recall values. In contrast, in the Wang's and dissimilarity-based methods, the mislabeled samples overlap with dense positive regions, leading to more confusion in the test set.

Figure 5.4: *PCA visualization for the Tsukiyama dataset using different sampling strategies.*

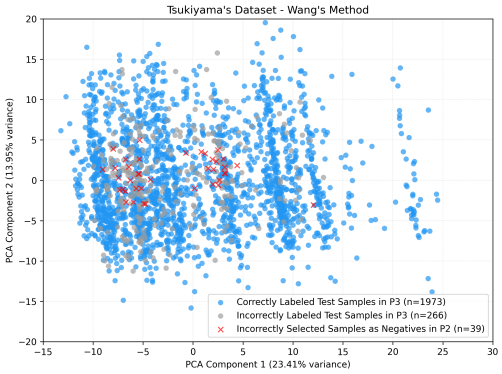
(a): *CNS*
Recall: 0.792



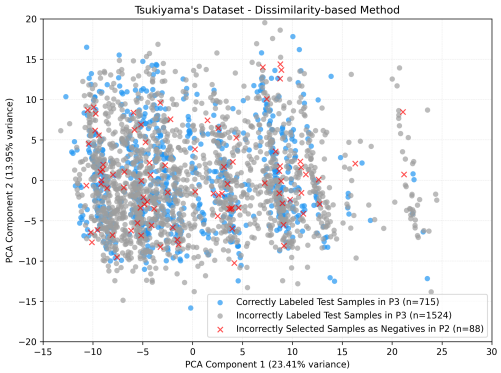
(b): *Wei*
Recall: 0.912



(c): *Wang*
Recall: 0.881



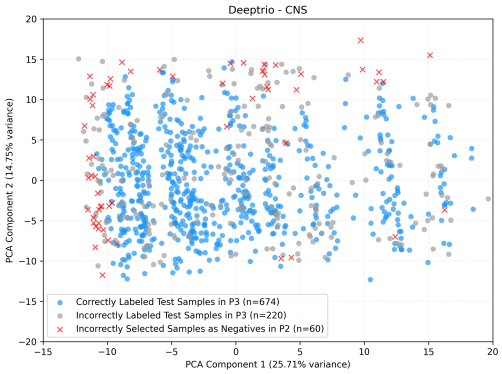
(d): *Diss.-based*
Recall: 0.319



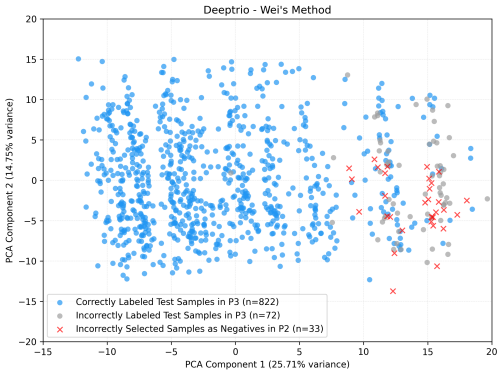
Overall, the results indicate that the distribution of incorrectly labeled negatives as positives affects the performance of the model as well as their amount. Sampling strategies that position mislabeled instances in isolated or low-density regions tend to preserve class separability and improve recall. On the other hand, an overlap with dense positive

Figure 5.5: *PCA visualization for the Human-Virus dataset using different sampling strategies.*

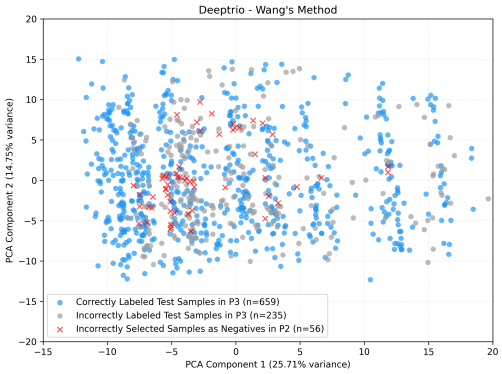
(a): *CNS*
Recall: 0.754



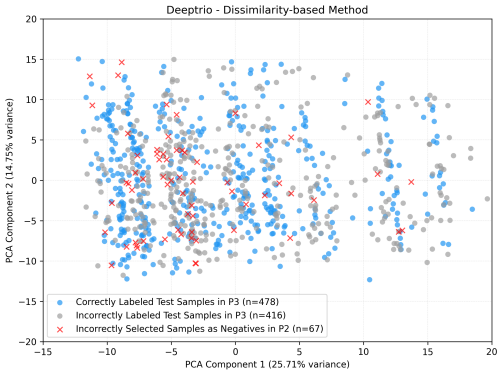
(b): *Wei*
Recall: 0.919



(c): *Wang*
Recall: 0.737



(d): *Diss.-based*
Recall: 0.535



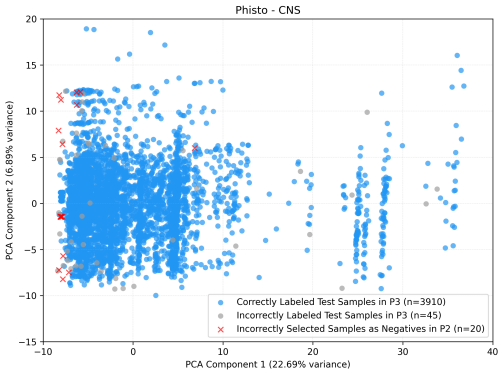
regions increases the probability of misclassification. These results show the crucial role of data distribution in determining how sampling strategies affect classification results.

5.3 The Study of Localisation-based Sampling Strategy

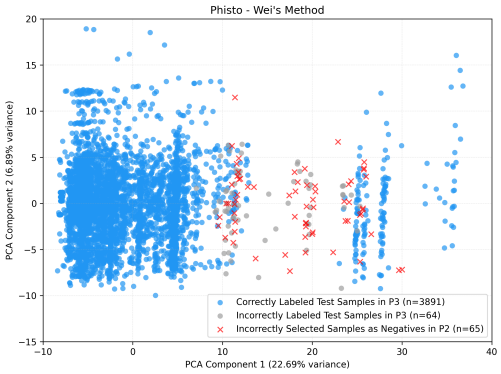
As part of our extended analysis, we introduced an additional negative sampling strategy based on subcellular localization. This approach is based on the biological assumption that proteins located in different subcellular compartments are unlikely to interact, as discussed by Guo et al. [34]. However, its applicability is limited to proteins with reliable localization annotations. We noticed that there are some proteins in our datasets

Figure 5.6: *PCA visualization for the PHISTO dataset using different sampling strategies.*

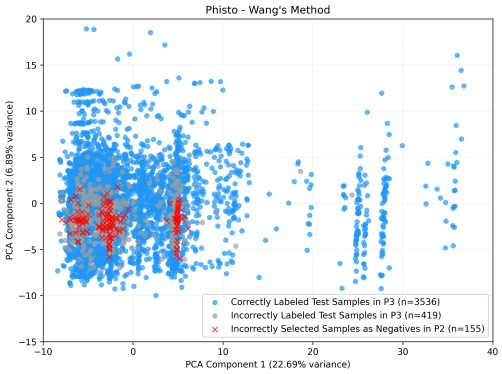
(a): *CNS*
Recall: 0.989



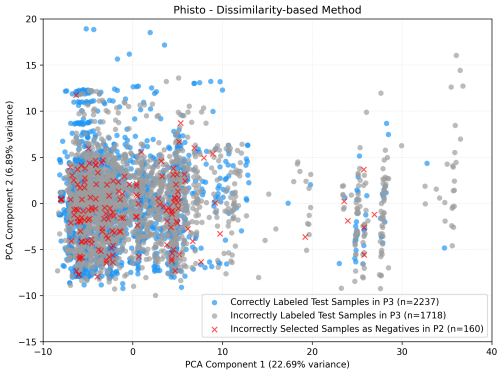
(b): *Wei*
Recall: 0.984



(c): *Wang*
Recall: 0.894



(d): *Diss.-based*
Recall: 0.566



where their localization anotations were not available. Therefore, we cannot directly compare with the benchmark methods in the previous section.

This localization-based strategy was introduced to reduce the likelihood of false negatives by avoiding pairs that may be colocalized. Negative samples were generated by pairing proteins from different localization categories, assuming that such pairs represent non-interacting proteins. To implement this method, subcellular localization annotations for the proteins in the positive dataset were retrieved from the Swiss-Prot database[1]. Protein pairs with incomplete or missing localization data were excluded from this procedure.

[1] `http://www.expasy.org/sprot/`

Proteins for which no localization information was available or annotated as "putative" or "hypothetical" were excluded to ensure the reliability of compartment assignment. The remaining proteins were categorized into eight primary subcellular compartments: Cytoplasm, Nucleus, Mitochondrion, Endoplasmic Reticulum, Golgi Apparatus, Peroxisome, Vacuole, and Cytoplasm $\&$ Nucleus. The resulting cross-compartment protein pairs were considered biologically implausible and served as negative samples. Data sets that were filtered according to this criterion are marked with an asterisk throughout the study (e.g., Herpes* instead of Herpes). In these filtered versions, the number of proteins was reduced compared to the original sets:

- Herpes*: $3,086$ human and 441 viral proteins ($\sim 79-76\%$ of the original).
- Tsukiyama*: $5,205$ human and 912 viral proteins ($\sim 89-92\%$ of the original).
- Human-Virus*: $2,896$ human and 287 viral proteins ($\sim 86-80\%$ of the original).
- PHISTO*: $5,595$ human and $1,423$ viral proteins ($\sim 85-83\%$ of the original).

To ensure a fair and consistent evaluation, we applied the R-SAFE method to retrain and evaluate the best-performing methods from Tables 5.1 and 5.2 using the filtered datasets. These methods were selected based on their performance in terms of recall and P.C. and evaluated under the same experimental setting, the results of which are shown in Table 5.3. In this table, "Best Method (P.C.-Based)" refers to the method with the lowest P.C. ratio from Table 5.1. "Best Method (Recall-Based)" refers to the method with the highest Recall value from Table 5.2. "Sub. Loc. Method" reports the results of the localization-based strategy.

The results demonstrate that the sampling strategy based on subcellular localization achieves the lowest average P.C. ratios in both the Herpes* and Tsukiyama* datasets across all data splits, indicating that it avoids selecting unknown positives as negatives.

However, it underperforms in terms of recall when compared to the best-performing methods. In the Herpes* dataset, the recall value remains relatively low (e.g. 0.485 ± 0.017 with a split of $80 : 10 : 10$) while the highest recall value of another method reaches 0.901 ± 0.085. Similarly, the localization-based method in the Tsukiyama* data set achieves 0.795 ± 0.011, falling short of 0.916 ± 0.047 achieved by the best-performing approach.

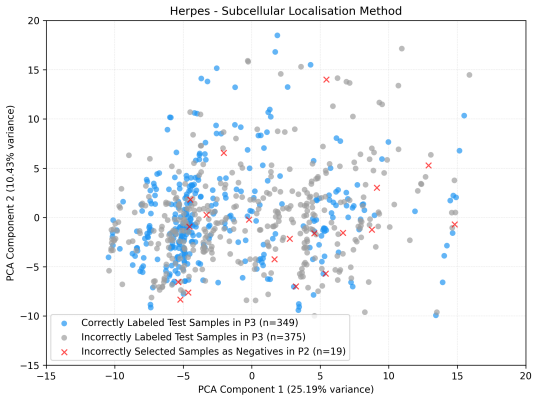
Table 5.3: *Performance comparison of three negative sampling strategies on filtered datasets where subcellular localization information was available.*

Dataset	Split Ratio	Best Method (P.C.-Based)		Best Method (Recall-Based)		Sub. Loc. Method	
		Recall	P.C. ratio	Recall	P.C. ratio	Recall	P.C. ratio
Herpes*	50:25:25	0.244 ± 0.011	%2.50	**0.771 ± 0.173**	%9.90	0.404 ± 0.019	**%2.40**
	60:20:20	0.793 ± 0.032	%3.18	**0.859 ± 0.107**	%8.57	0.433 ± 0.024	**%2.59**
	70:15:15	0.797 ± 0.041	%3.52	**0.890 ± 0.104**	%6.70	0.461 ± 0.024	**%2.81**
	80:10:10	0.296 ± 0.028	%3.96	**0.901 ± 0.085**	%6.67	0.485 ± 0.017	**%2.49**
Tsukiyama*	50:25:25	0.790 ± 0.032	%1.34	**0.835 ± 0.094**	%4.87	0.738 ± 0.009	**%1.31**
	60:20:20	0.834 ± 0.030	%1.63	**0.858 ± 0.085**	%6.07	0.762 ± 0.010	**%1.34**
	70:15:15	0.852 ± 0.023	%2.00	**0.892 ± 0.068**	%5.89	0.777 ± 0.011	**%1.30**
	80:10:10	0.871 ± 0.013	%2.49	**0.916 ± 0.047**	%5.39	0.795 ± 0.011	**%1.35**
Human-Virus*	50:25:25	0.529 ± 0.013	%4.66	**0.868 ± 0.135**	**%3.69**	0.541 ± 0.082	%5.10
	60:20:20	**0.907 ± 0.072**	**%3.33**	**0.907 ± 0.072**	**%3.33**	0.543 ± 0.061	%5.51
	70:15:15	**0.936 ± 0.018**	**%3.16**	**0.936 ± 0.018**	**%3.16**	0.560 ± 0.050	%6.29
	80:10:10	**0.886 ± 0.077**	**%4.04**	**0.886 ± 0.077**	**%4.04**	0.568 ± 0.036	%6.74
PHISTO*	50:25:25	**0.895 ± 0.037**	**%0.35**	0.891 ± 0.056	%3.69	0.658 ± 0.014	%1.81
	60:20:20	**0.930 ± 0.041**	**%0.49**	**0.930 ± 0.041**	**%0.49**	0.681 ± 0.007	%1.95
	70:15:15	**0.947 ± 0.029**	**%0.68**	**0.947 ± 0.029**	**%0.68**	0.702 ± 0.006	%2.11
	80:10:10	**0.963 ± 0.021**	**%0.85**	**0.963 ± 0.021**	**%0.85**	0.720 ± 0.007	%2.36

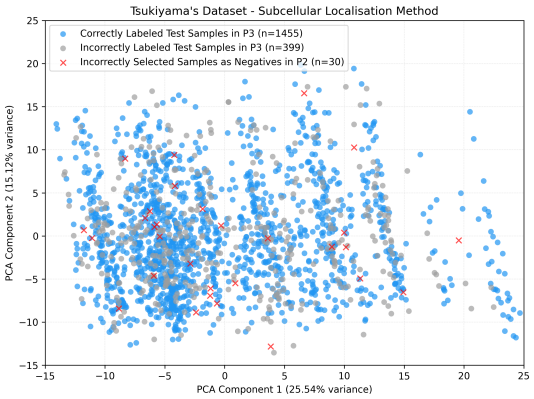
The PCA visualizations of the subcellular localization method, as shown in Figure 5.7, illustrate the distribution of incorrectly selected negative samples ($P2$) and their impact on model performance across the Human-Virus*, PHISTO*, Herpes*, and Tsukiyama* datasets. In all datasets, these incorrectly selected negative samples are typically distributed within and close to dense regions of positive test samples. This spatial overlap makes it difficult for the model to distinguish between positive and negative classes, leading to ambiguity during the learning process. As a result, despite its low P.C. ratios, the subcellular localization method provides lower recall performance compared to other sampling strategies. This also shows that not only the number of incorrectly selected negatives (P.C.), but also their spatial position within the data structure plays a crucial role in the performance of the model. On the other hand, the best-performing method achieves

Figure 5.7: *PCA visualization of the subcellular localisation-based negative sampling method across datasets.*

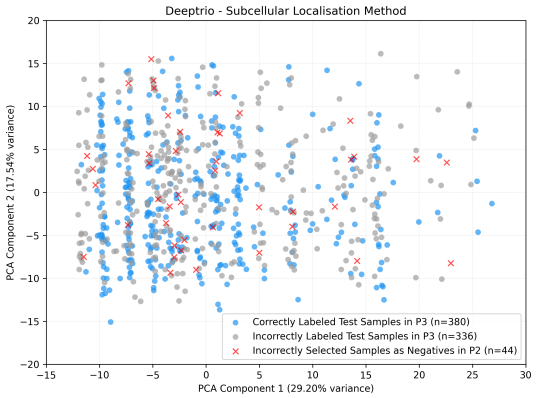
(a): *Herpes*
Recall: 0.482



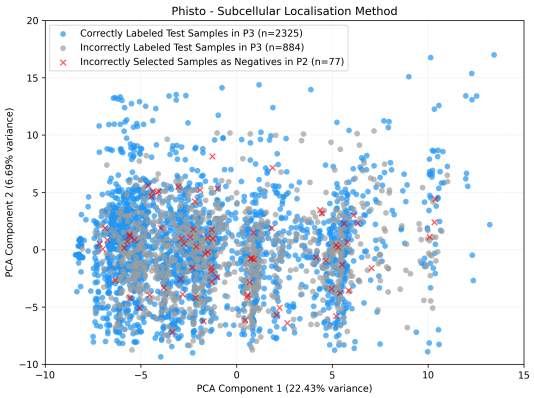
(b): *Tsukiyama*
Recall: 0.785



(c): *Human-Virus*
Recall: 0.531



(d): *PHISTO*
Recall: 0.718



strong results in terms of both P.C. ratios and recall, indicating a consistent and reliable sampling strategy in both the PHISTO* and Human-Virus* datasets. For example, in PHISTO*, the same method yields the lowest P.C. ratios across all splits (e.g., 0.85% at $80:10:10$) while achieving the highest recall value, such as 0.963 ± 0.021 in the same ratio. A similar pattern is observed in Human-Virus*, where the best method not only minimizes false negatives (e.g., 4.04% P.C. ratio at $80:10:10$), but also achieves high recall, such as 0.886 ± 0.077. These results show that in both the PHISTO* and Human-Virus* datasets, the same negative sampling strategy performs best on both evaluation

criteria, P.C. ratio and recall, suggesting a strong correspondence between reliability and predictive performance in these datasets.

To sum up, the evaluation shows two different patterns. In Human-Virus* and PHISTO*, some strategies achieve both high recall and low P.C. ratio, indicating the potential for effective and reliable sampling due to a clearer separation between correctly and incorrectly interactions. In contrast, Herpes and Tsukiyama have a noticeable conflict between predictive performance and sampling reliability. Although the subcellular localization method provides promising results in some datasets, it is not fully applicable in this study due to the incomplete localization annotations. Therefore, the sampling strategy based on subcellular localization was excluded from further evaluation due to its limited applicability in benchmark datasets.

5.4 Extended Analysis: Evaluating Negative Sampling Reliability Using the Negatome Dataset

A fundamental challenge in protein-protein interaction (PPI) prediction is the construction of a biologically reliable negative class. While positive interactions can be experimentally verified and are available in curated databases, the absence of interaction between protein pairs is rarely documented with the same level of confidence. As a result, most PPI prediction studies rely on artificially generated or randomly sampled pairs as negative examples, introducing the risk of including *unknown positives*—interactions that are biologically valid but have not yet been discovered. Such mislabeling can significantly degrade model performance and generalization.

To address this issue, we conducted an **extended validation experiment** leveraging the *Negatome* dataset [30], a manually curated resource that contains protein pairs experimentally demonstrated **not to interact**. Unlike conventional random sampling or

dissimilarity-based sampling strategies, Negatome offers a biologically grounded benchmark for assessing the reliability of negative sampling. It includes human protein pairs verified through literature curation and structural analysis, and thus serves as a high-confidence reference for true negative examples in PPI prediction.

The extended methodology is illustrated in Figure 5.8, which depicts the evaluation framework incorporating Negatome data. This protocol consists of three main phases. In the **initial phase**, the positive interaction set is divided into three disjoint subsets: $P1$ for training the model, $P2$ to simulate unknown positives by merging it into the unlabeled pool, and $P3$ reserved exclusively for final testing. Similarly, the Negatome dataset is partitioned into $N1$, which is included in the unlabeled pool during training, and $N2$, which is reserved for testing as high-confidence non-interacting pairs.

During the **training phase**, a candidate pool of unlabeled interactions is created by merging $P2$ and $N1$, representing a realistic mixture of true negatives (from Negatome) and unknown positives (from $P2$). A negative sampling strategy is then applied to this pool to select negative samples for training, allowing the evaluation of the method's ability to avoid mislabeling unknown positives as negatives while leveraging the reliability of Negatome data. Finally, in the **testing phase**, the trained model is evaluated on a test set comprising $P3$ (experimentally verified positive interactions) and $N2$ (experimentally verified negative interactions).

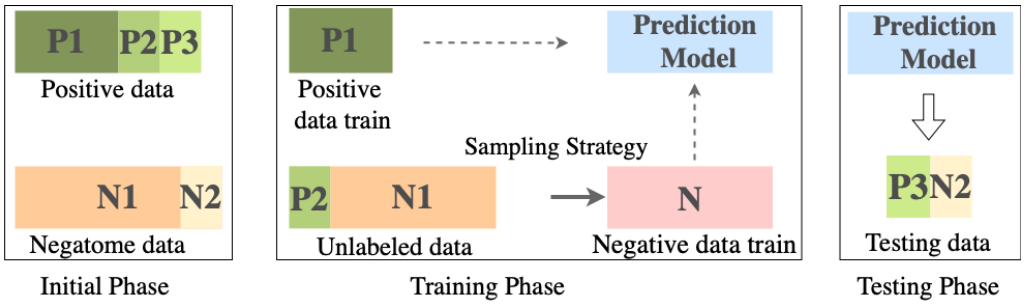


Figure 5.8: *Evaluation methodology incorporating Negatome data.*

Initially, a total of $2,164$ protein-protein interaction pairs were obtained from the

Negatome dataset. To ensure that these pairs did not contain any biologically relevant interactions, the unique protein identifiers from these pairs were extracted. Each of these proteins was then queried in the IntAct database [27], which includes experimentally validated protein-protein interactions. As a result, $1,648$ positive interactions involving these proteins were identified in IntAct.

To evaluate the performance of the negative sampling strategies, the positive interaction set was divided into three disjoint subsets: $P1$ for training, $P2$ to simulate unknown positives, and $P3$ for testing. This approach follows the methodology proposed in R-SAFE (in Section 5.2). To enable a controlled evaluation across varying training data sizes, experiments were conducted using four different split configurations(i.e. $P1$, $P2$, $P3$): $50:25:25$, $60:20:20$, $70:15:15$, and $80:10:10$. For example, in the $80:10:10$ split, 80% of the positive interactions were allocated to $P1$, 10% to $P2$, and the remaining 10% to $P3$. To create a balanced test set containing both positive and high-confidence negative samples, a subset of the Negatome dataset called $N2$ was chosen to have the same number of samples as the positive test set $P3$. This means that for every data split, the size of $N2$ always selected the number of $P3$. The remaining Negatome pairs, after removing $N2$, were assigned to $N1$ and included in the training phase as part of the unlabeled data pool. Including these Negatome pairs in the unlabeled pool simulates realistic scenarios where some negatives might be mislabeled or uncertain, which challenges the negative sampling strategy to effectively avoid selecting unknown positives. This approach ensured that the test set included unseen positive interactions ($P3$) and experimentally validated negative interactions ($N2$) without any overlap with training data. This entire data partitioning and evaluation procedure was repeated using 10 different random seeds to ensure the robustness and reliability of the results across multiple randomized splits.

In line with the evaluation methodology employed in the previous R-SAFE (in the

Section 5.1, the Random Forest classifier was used consistently as the prediction model across all experiments. The evaluation results presented in Table 5.4 reveal important insights into the effectiveness of benchmark negative sampling strategies within the context of protein-protein interaction prediction using the Negatome dataset. This table represent the mean and standard deviation calculated over 10 different random seeds to ensure statistical robustness. All metrics were computed on the held-out test set, providing an unbiased assessment of each negative sampling strategy's performance under varying data split ratios.

Table 5.4: *Performance comparison (mean ± std) in Negatome dataset.*

Split Ratio	Method	P.C. Ratio (%)	Recall	Accuracy	F1-score	ROC-AUC
50:25:25	Dissimilarity-based	%0.75 ± 0.21	0.232 ± 0.072	0.516 ± 0.087	0.320 ± 0.060	0.502 ± 0.119
	CNS Method	**%0.07 ± 0.11**	0.319 ± 0.060	0.617 ± 0.025	0.452 ± 0.059	0.759 ± 0.040
	Random Method	%0.29 ± 0.34	0.148 ± 0.012	0.564 ± 0.007	0.253 ± 0.017	0.641 ± 0.020
	Wang's Method	%0.61 ± 0.29	0.350 ± 0.063	**0.636** ± 0.029	0.488 ± 0.064	**0.775** ± 0.021
	Wei's Method	%8.04 ± 6.28	**0.551 ± 0.234**	0.619 ± 0.060	**0.576 ± 0.072**	0.666 ± 0.055
60:20:20	Dissimilarity-based	%0.79 ± 0.36	0.227 ± 0.067	0.517 ± 0.078	0.318 ± 0.063	0.470 ± 0.099
	CNS	**%0.09 ± 0.14**	0.357 ± 0.087	**0.628** ± 0.034	0.485 ± 0.074	0.752 ± 0.063
	Random Sampling	%0.33 ± 0.39	0.170 ± 0.020	0.574 ± 0.011	0.285 ± 0.029	0.653 ± 0.021
	Wang's Method	%0.63 ± 0.41	0.334 ± 0.036	0.618 ± 0.019	0.466 ± 0.034	**0.771** ± 0.027
	Wei's Method	%7.10 ± 6.00	**0.673 ± 0.252**	0.552 ± 0.073	**0.587 ± 0.071**	0.644 ± 0.068
70:15:15	Dissimilarity-based	%0.77 ± 0.40	0.227 ± 0.056	0.540 ± 0.066	0.329 ± 0.060	0.532 ± 0.104
	CNS	**%0.16 ± 0.28**	0.415 ± 0.106	**0.655** ± 0.042	0.539 ± 0.094	0.775 ± 0.068
	Random Sampling	%0.48 ± 0.49	0.173 ± 0.022	0.577 ± 0.013	0.289 ± 0.033	0.656 ± 0.026
	Wang's Method	%1.01 ± 0.72	0.359 ± 0.061	0.634 ± 0.035	0.493 ± 0.064	**0.789** ± 0.026
	Wei's Method	%8.39 ± 6.39	**0.619 ± 0.191**	0.612 ± 0.074	**0.609 ± 0.054**	0.652 ± 0.075
80:10:10	Dissimilarity-based	%1.27 ± 0.83	0.246 ± 0.044	0.577 ± 0.071	0.368 ± 0.062	0.578 ± 0.080
	CNS	**%0.12 ± 0.25**	0.471 ± 0.120	**0.670** ± 0.044	0.580 ± 0.090	**0.796** ± 0.031
	Random Sampling	%0.42 ± 0.57	0.175 ± 0.032	0.580 ± 0.015	0.293 ± 0.045	0.674 ± 0.026
	Wang's Method	%1.33 ± 0.85	0.367 ± 0.053	0.636 ± 0.034	0.501 ± 0.055	0.781 ± 0.041
	Wei's Method	%10.96 ± 4.70	**0.510 ± 0.101**	0.654 ± 0.032	**0.591 ± 0.068**	0.654 ± 0.044

In this table, the performance of different negative sampling methods on the Negatome dataset is compared, with a focus on how mistakenly selecting unknown positive samples as negatives (indicated by the P.C. Ratio) affects model generalization. Clustering-based methods, particularly the CNS approach, consistently outperform random sampling across all split ratios in terms of recall and F1-score metrics. This superiority is attributable to their more precise selection of true negative samples, as evidenced by the substantially lower P.C. Ratios, which quantify the proportion of mislabeled positives mistakenly included in the negative training set. For example, under the $70:15:15$ split, CNS achieves a P.C. Ratio of only $0.16\% \pm 0.28\%$, accompanied by a recall of

0.415 ± 0.106 and an F1-score of 0.539 ± 0.094, while Random Sampling records higher P.C. Ratios ($0.48\% \pm 0.49\%$) yet considerably lower recall (0.173 ± 0.022) and F1-score (0.289 ± 0.033). These findings indicate that clustering-based methods provide a more reliable negative set, reducing the mislabeling of unknown positives.

Interestingly, methods such as Wei's, despite their elevated P.C. Ratios (e.g., 8.39% $\pm 6.39\%$ in 70:15:15 split), achieve higher recall (0.619 ± 0.191) and F1-scores (0.609 ± 0.054) compared to other methods with lower P.C. Ratios. This observation suggests that mistakenly labeling a subset of true positives as negatives may, in some cases, improve the generalizability of the model, thereby leading to better performance in the test set. However, this effect appears to vary across datasets, indicating that the impact of such mislabeling is context-dependent and warrants further investigation.

6. SELECTING THE OPTIMAL NEGATIVE SAMPLING METHOD

After evaluating and comparing negative sampling strategies across multiple benchmark datasets (i.e., datasets containing only known positive protein-protein interactions such as PHISTO, DeepTrio, Tsukiyama, and Human-Virus), we found that the best negative sampling method varies depending on the dataset's structure. Accordingly, we propose characterizing each dataset and machine learning setting to learn which negative sampling is the most appropriate for those settings.

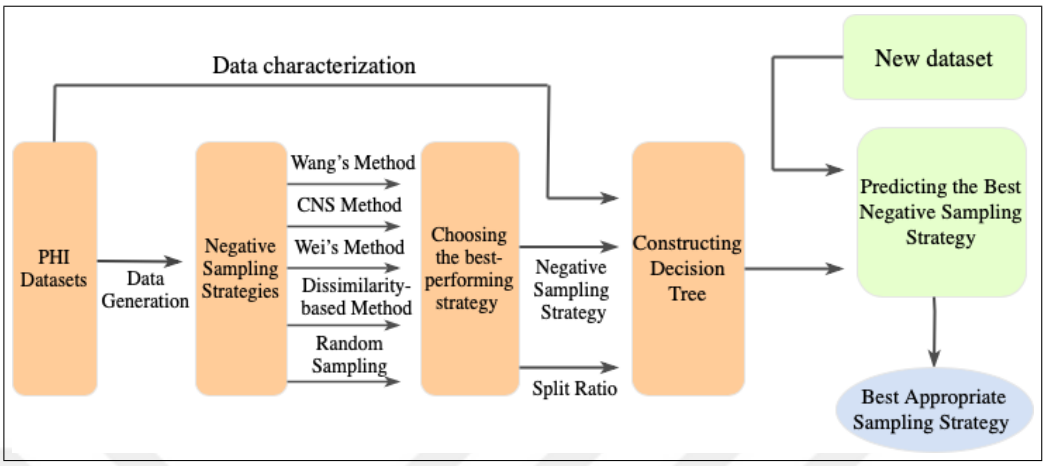
As illustrated in Figure 6.1, we formulated the problem of determining the most appropriate the negative sampling problem as a classification problem. The first step involves running experiments on different datasets using benchmark sampling strategies and recording their performances. The best negative sampling strategy for each setting is selected based on performance criteria (e.g., P.C. ratios in our study). The corresponding dataset and split ratio are mapped to the set of descriptive features for each setting. Afterward, a decision tree algorithm is trained with these dataset features and the corresponding best negative sampling strategy. The constructed decision tree can be utilized to predict the best negative sampling strategy for a given dataset.

In the following section, we extract descriptive features for each dataset and machine learning setting. Consequently, we construct a decision tree from our preprocessed experimental data (Section 4.2.1) and report its prediction performance (Section 5.2).

6.1 Characterization of Host-Pathogen Interaction Problems

We determine a total of 36 features to capture distinctive characteristics of pathogen-host protein-protein interaction datasets and corresponding machine learning settings. These features capture structural, compositional, and interaction-related properties of the

Figure 6.1: *An overview of the decision tree-based framework for selecting the optimal negative sampling strategy.*



protein sequences and are divided into six categories as follows:

- **Dataset-level summary ratios:** At the dataset level, two summary ratios are used to characterize key properties of the data. The *Labeled-to-Unlabeled Ratio* quantifies the proportion of known positive interactions relative to the total number of unlabeled pairs. This ratio indicates dataset sparsity, where a lower value reflects greater uncertainty and a higher degree of class imbalance. The *Unique Virus-to-Host Count Ratio* measures the total number of unique viral proteins to unique host proteins. This metric offers insight into the diversity and balance of interaction pairs between the two species; a high ratio may suggest a dominance of viral proteins in the interaction space.

- **Distance-based interaction feature:** It is used to assess the similarity between known and unlabeled protein-protein interaction pairs. In particular, the *mean cosine distance between embeddings* is employed to quantify this similarity. The cosine distance to all known positive interactions is computed for each unlabeled interaction based on concatenated vector embeddings of the viral and host proteins, which are generated using a sequence embedding model. The resulting average

distance reflects how closely the unlabeled interactions resemble known positives in the embedding space, offering insight into their potential likelihood of being true interactions.

- **Sequence similarity features:** These features include pairwise sequence similarity scores between host and viral protein sequences, calculated through global alignment using a scoring scheme that rewards matches ($+2$), penalizes mismatches (-1), and applies gap penalties of -0.5 for opening and -0.1 for extension. Each alignment score is normalized by the length of the longer sequence, resulting in a similarity ratio between 0 and 1. For each dataset, the average pairwise similarity scores between host and virus sequences are computed to yield a single representative value, referred to as the *average host-virus sequence similarity*. The *host-to-virus sequence length ratio* is also calculated to express the relative difference in protein sequence lengths between the host and virus.

- **Characteristics of the amino acid composition:** They are examined by analyzing host and viral protein sequences according to the chemical properties of their constituent amino acids. Amino acids are grouped into eight categories: aliphatic (Ala, Val, Leu, Ile, Gly), aromatic (Phe, Tyr, Trp), sulfur-containing (Cys, Met), acidic (Asp, Glu), basic (Lys, Arg, His), hydroxyl-containing (Ser, Thr), amide-containing (Asn, Gln), and cyclic (Pro). The average percentage of each group is calculated separately for host and viral proteins. This process involves three steps: (i) determining the relative abundance of each amino acid in individual protein sequences, (ii) summing the percentages of amino acids within each chemical group, and (iii) averaging these values across all proteins to obtain group-level percentages. The computations use the `ProteinAnalysis` module from Biopython. The resulting features—*aliphatic*, *aromatic*, *sulfur-containing*, *acidic*, *basic*,

hydroxyl-containing, *amide-containing*, and *cyclic residue*—capture proteins' key structural and functional attributes. For instance, aliphatic residues contribute to hydrophobicity, and aromatic residues enable stacking interactions, acidic and basic residues shape charge distribution, and sulfur-containing residues (such as cysteine) can form disulfide bonds that stabilize protein structure.

- **Physico-chemical properties:** It describe fundamental biochemical characteristics of host and viral proteins. The *average hydrophobicity*, measured using the GRAVY (Grand Average of Hydropathy) score, reflects the mean hydrophobicity index of all amino acids within a protein. A higher GRAVY score indicates greater hydrophobicity, which may influence a protein's folding, solubility, and interaction behavior. The *average molecular weight* is determined by summing the molecular weights of all amino acids in each sequence and computing the average across the dataset. The *average net charge* at pH 7.0 estimates the electrostatic charge of each protein under physiological conditions, which plays a critical role in interaction potential by affecting binding affinity and repulsion. These three metrics are calculated separately for host and viral proteins using the `ProteinAnalysis` module in Biopython. Prior to analysis, unusual amino acids such as X or U are excluded to ensure accurate computations.

- **Statistics of the dipeptide composition**: They are used to capture short-range sequence patterns in host and viral proteins. All 400 possible dipeptides, combinations of two consecutive amino acids, are counted for each protein. The relative frequency of each dipeptide is computed by dividing its count by the total number of dipeptides in the sequence. From this frequency distribution, four statistical features are extracted to summarize the composition: the *mean frequency* reflects the average occurrence of dipeptides across the dataset; the *median frequency* provides

a central tendency that is less affected by outliers; the *standard deviation* quantifies the variability in dipeptide usage among sequences; and the *Shannon entropy* measures the overall uncertainty or diversity in dipeptide usage, where higher values indicate a more uniform and complex pattern.

6.2 Our Decision Tree-based Selection Approach

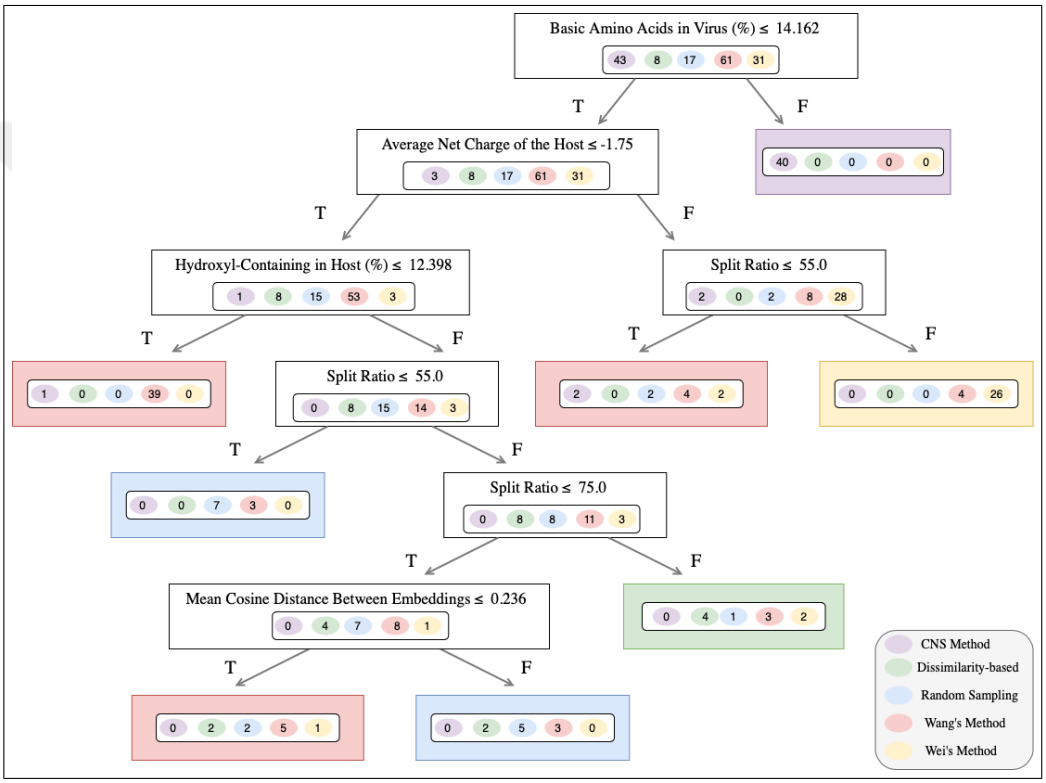
Following the aforementioned procedure in Figure 6.1, we consider 160 experimental results obtained when the Random Forest algorithm is used on four benchmark datasets for host-pathogen interactions. For each dataset, ten different random seeds are used to introduce variability, and four different split ratios ($50:25:25$, $60:20:20$, $70:15:15$, and $80:10:10$) are used to represent training, test, and simulated unknown positive sets. Five sampling strategies are considered: CNS, Wang's method, Wei's method, dissimilarity-based sampling, and random sampling. Based on the evaluation results previously obtained, each instance is labeled with the strategy that yielded the lowest number of mislabeled samples (P.C.), which served as the target variable for model training. The model is trained to associate dataset features (as described in Section 6) and split ratios with the most effective sampling method.

Figure 6.2 shows the structure of the resulting decision tree model that we obtained. Each node represents a decision point, while the leaf nodes indicate the recommended sampling method. The model shows a clear and interpretable decision path where features such as the percentage of basic amino acids in viral proteins, the average net charge of the host, hydroxyl-containing amino acids in host proteins, data split ratio, and the mean cosine distance between protein embeddings emerge as dominant split criteria.

The decision tree effectively captures the relationships between the dataset's features and the sampling method preferences. For example, instances with a high proportion of basic amino acids in the virus are consistently associated with the CNS method.

In contrast, instances with a low net host charge and low hydroxyl content often lead to the selection of the Wang method. While CNS and Wang dominate certain regions of the tree, Wei's method is chosen in a lower split ratio where higher net charges are observed. The repeated occurrence of split-ratio thresholds as internal decision points also suggests that sampling effectiveness is influenced by biological features and data splitting.

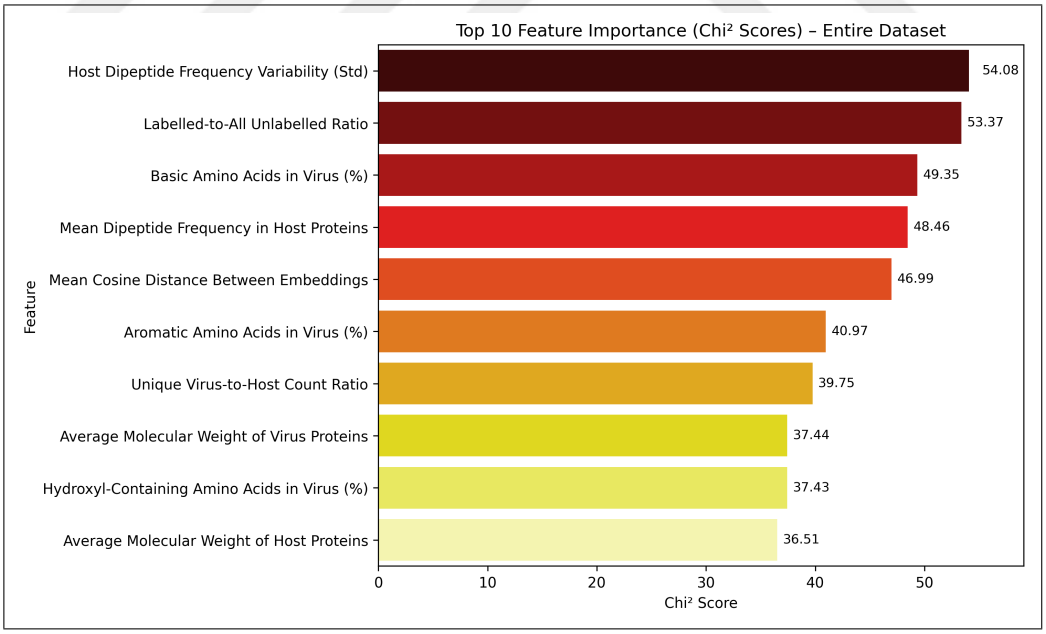
Figure 6.2: *Decision tree structure illustrating how dataset-level features are used to predict the optimal negative sampling strategy.*



Furthermore, using cosine distance between embeddings as a node at a deeper level suggests that sequence-based representations also make a meaningful contribution to decision boundaries. Notably, the grouping of sampling strategies in the leaf nodes shows that certain dataset profiles are consistently associated with specific sampling methods. This indicates that the model works logically and reliably in selecting the most appropriate strategy. It provides an interpretable and practical framework that translates complex dataset characteristics into clear and helpful sampling decisions.

To support the interpretability of the decision tree model, we performed a chi-square analysis (χ^2) of the importance of characteristics for protein-protein interaction datasets using the SelectKBest method (Figure 6.3). Of the 36 features at the dataset level used to train the model, the figure highlights the top 10 features with the highest values of χ^2. It can be seen that several features match the key decision points in the decision tree (Figure 6.2). In particular, the percentage of basic amino acids in virus proteins, which appears as the root node of the tree, is ranked as the third most important feature in the χ^2 results, confirming its strong discriminatory power. The mean cosine distance between embeddings, which is used to discriminate between strategies such as random sampling and Wang's method in the deeper branches, is ranked fifth, while hydroxyl-containing in host proteins, which helps to separate Wang-related pathways, appears as the ninth most important feature.

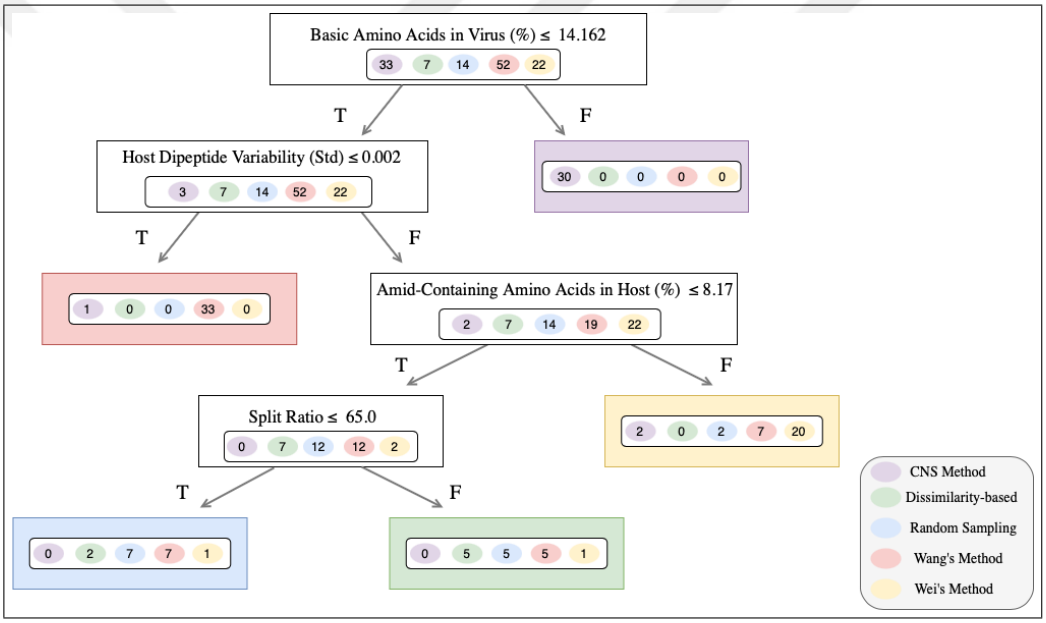
Figure 6.3: *Feature importance scores of the top* 10 *dataset-level features selected using the* `SelectKBest` *method with the* χ^2 *scoring function.*



To further evaluate the consistency and robustness of the decision process, we trained a decision tree model that uses only the 128 training instances from the whole

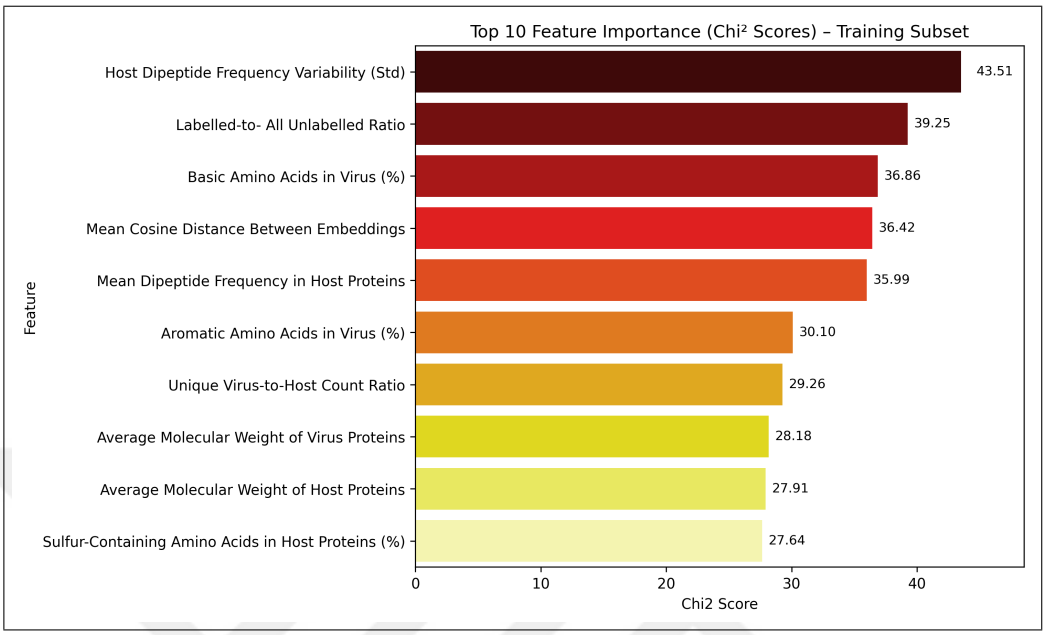
experimental setup and excludes the 32 test instances. The resulting structure is shown in Figure 6.4. Although the decision paths differ from those of the tree trained for all the instances (Figure 6.2), specific key patterns are preserved. In particular, the percentage of basic amino acids in viral proteins remains a common and dominant decision node in both models. In contrast, features such as host dipeptide variability, which did not appear in the tree trained on all data, emerge as informative split criteria in the training-specific model, suggesting that their discriminative power is more apparent in specific data partitions.

Figure 6.4: *Decision tree structure trained using only the training instances (128 samples).*



To support these results, we performed a feature importance analysis using the χ^2 scoring function for the training subset (Figure 6.5). The analysis shows that host dipeptide frequency variability is the highest-scoring feature, further confirming its role in the decision process of the training-specific model. Although the percentage of basic amino acids in viral proteins ranked third, this is the only feature consistently selected in both trees, highlighting its strong and generalizable predictive power. These results show that some features remain stable across different data configurations.

Figure 6.5: *Top 10 dataset-level feature importance scores based on χ^2 values from the training subset.*



Overall, the convergence between the splits of the decision tree and the top ranked χ^2 features emphasizes the model's ability to extract biologically and statistically meaningful patterns, supporting its reliability in predicting the most appropriate sampling strategy in different dataset scenarios.

6.3 Assessment of Selection Approach

To assess the effectiveness of the proposed selection model, we perform 10-fold cross-validation. Table 6.1 shows the prediction performance metrics such as precision, recall, and F1 on each fold and an average of those results. It can be observed that the model achieved an average F1 score of 0.718%, with a corresponding precision of 0.744% and recall of 0.738%. In addition to cross-validation, we also evaluated the performance of the model by splitting it into training and testing. The decision tree classifier was trained on the training set (128 instances) and achieved its best performance on the test set (32 instances) with an F1-score of 0.863%. Considering the small size of the training set for

the decision tree model, these results are reasonable for determining the most appropriate negative sampling strategy in protein-protein interaction datasets.

Table 6.1: *10-Fold Cross-Validation Results for Decision Tree Classifier.*

Fold	F1-score	Precision	Recall
1	0.718	0.812	0.687
2	0.797	0.950	0.750
3	0.742	0.774	0.750
4	0.820	0.777	0.875
5	0.732	0.671	0.812
6	0.791	0.821	0.812
7	0.623	0.596	0.687
8	0.771	0.862	0.750
9	0.576	0.580	0.625
10	0.607	0.593	0.625
Average	**0.718**	**0.744**	**0.737**

Overall, we provide a principled approach to assess the reliability of sampling strategies in the absence of known negatives. Integrating a decision tree model supports identifying key features of the dataset that influence sampling effectiveness, allowing for a more informed and interpretable method selection process.

7. CONCLUSION

In this thesis, we address the challenge of selecting negative samples from unlabeled data to predict protein-protein interactions (PPI) between pathogen and host. To increase the reliability of negative sample selection, we propose a novel clustering-based approach and emphasize its methodological differences from existing clustering-based techniques and the widely used random sampling strategy and dissimilarity-based method. In our experiments conducted on four virus-human PPI datasets, five sampling methods and four classifiers were evaluated. It has been shown that cluster-based approaches consistently outperform random and dissimilarity-based methods regarding prediction performance. To further investigate the behavior and effectiveness of these sampling strategies, we introduce a structured evaluation framework called R-SAFE, which simulates unknown positive interactions to quantify the risk of incorrectly labeling true positives as negatives. Using R-SAFE, we systematically evaluate five strategies, random, dissimilarity-based, Wei and Wang strategies, and our proposed CNS method, on four benchmark datasets with different data split ratios ($50:25:25$, $60:20:20$, $70:15:15$ $80:10:10$) and ten random seeds. The impact of each negative sampling strategy on prediction is evaluated using the rates of incorrectly selected unknown positive samples (P.C. scores) and the recall value achieved by the classifier. In addition, PCA visualizations are used to gain insight into the distribution of mislabeled negative samples in the feature space (i.e., to explain the relationship between sampling behavior and model performance). The results show that the number of incorrectly selected samples and the spatial distribution significantly influence the prediction performance. Although cluster-based strategies (CNS, Wei, Wang) performed well in specific contexts, no single method was universally superior. To complement the R-SAFE evaluation on benchmark datasets that contain only known positives, we incorporated the Negatome

database, which provides experimentally validated non-interacting protein pairs, to construct a more biologically grounded evaluation setting. By integrating Negatome-derived negatives into both training and testing phases, we were able to evaluate the robustness of negative sampling strategies when true negatives are explicitly available. The results demonstrated that clustering-based methods, particularly the CNS approach, maintained low mislabeling rates while achieving high recall and F1-score values. It is observed that the effectiveness of sampling strategies varies widely across datasets and depends on features such as amino acid composition, physicochemical properties, and sequence similarity. To better understand these differences and enable automated method selection, we trained a decision tree model that identifies the key features of the dataset that influence sampling strategy effectiveness and recommends the most appropriate strategy for a given dataset. The decision tree model successfully linked features of the dataset to the best negative sampling strategies and provided an interpretable and reliable framework for selecting biologically meaningful negatives. These results emphasize the importance of dataset-specific and biologically informed negative selection.

This thesis highlights the crucial role of negative sampling strategies not only in improving classification accuracy, but also in ensuring biological validity and reducing the risk of negative samples being incorrectly selected as positive, significantly impacting the overall reliability of prediction models. Future work may focus on refining feature representation, investigating the relationship between dataset characteristics and sampling effectiveness, and extending this framework to other domains with similar data problems. In addition, we believe that a more precise definition of protein feature representation will lead to more discriminative results in selecting negative samples.

REFERENCES

[1] J. Bekker and J. Davis, "Learning from positive and unlabeled data: a survey," *Mach. Learn.*, vol. 109, p. 719–760, Apr. 2020.

[2] M. Kshirsagar, J. Carbonell, and J. Klein-Seetharaman, "Multitask learning for host–pathogen protein interactions," *Bioinformatics*, vol. 29, p. i217–i226, Jun 2013.

[3] S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein–protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, pp. i802–i810, 09 2018.

[4] S. Sledzieski, R. Singh, L. Cowen, and B. Berger, "D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions," *Cell Systems*, vol. 12, no. 10, pp. 969–982.e6, 2021.

[5] W. Liu-Wei, S. Kafkas, and Chen, "Deepviral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes," *Bioinformatics*, vol. 37, p. 2722–2729, Mar 2021.

[6] T. N. Dong, G. Brogden, G. Gerold, and M. Khosla, "A multitask transfer learning framework for the prediction of virus-human protein–protein interactions," *BMC bioinformatics*, vol. 22, pp. 1–24, 2021.

[7] M. Khandelwal, R. K. Rout, and S. Umer, "Protein-protein interaction prediction from primary sequences using supervised machine learning algorithm," in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 268–272, 2022.

[8] M. B. Koca, E. Nourani, F. Abbasoğlu, İ. Karadeniz, and F. E. Sevilgen, "Graph convolutional network based virus-human protein-protein interaction prediction for novel viruses," *Computational Biology and Chemistry*, vol. 101, p. 107755, 2022.

[9] I. Ieremie, R. M. Ewing, and M. Niranjan, "Transformergo: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms," *Bioinformatics*, vol. 38, pp. 2269–2277, 02 2022.

[10] S. Madan, V. Demina, M. Stapf, O. Ernst, and H. Fröhlich, "Accurate prediction of virus-host protein-protein interactions via a siamese neural network using deep protein sequence embeddings," *Patterns*, vol. 3, no. 9, p. 100551, 2022.

[11] B. Song, X. Luo, X. Luo, Y. Liu, Z. Niu, and X. Zeng, "Learning spatial structures of proteins improves protein–protein interaction prediction," *Briefings in Bioinformatics*, vol. 23, p. bbab558, 01 2022.

[12] X. Li, P. Han, W. Chen, C. Gao, S. Wang, T. Song, M. Niu, and A. Rodriguez-Patón, "Marppi: boosting prediction of protein–protein interactions with multi-scale architecture residual network," *Briefings in Bioinformatics*, vol. 24, p. bbac524, 12 2022.

[13] Y. Liu, Y. Liu, and Z. Li, "Protein–protein interaction prediction via structure-based deep learning," *Proteins: Structure, Function, and Bioinformatics*, vol. 92, no. 11, pp. 1287–1296, 2024.

[14] Z. Wei, D. Yao, X. Zhan, and S. Zhang, "A clustering-based sampling method for mirna-disease association prediction," *Frontiers in Genetics*, vol. 13, Sep 2022.

[15] B. Wang, C. Mei, Y. Wang, Y. Zhou, and Cheng, "Imbalance data processing strategy for protein interaction sites prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, p. 985–994, May 2021.

[16] X. Yang, S. Yang, Q. Li, S. Wuchty, and Z. Zhang, "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 153–161, 2020.

[17] X. Yang, S. Yang, X. Lian, S. Wuchty, and Z. Zhang, "Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction," *Bioinformatics*, vol. 37, pp. 4771–4778, 07 2021.

[18] M. N. Asim, A. Fazeel, M. A. Ibrahim, A. Dengel, and S. Ahmed, "Mp-vhppi: Meta predictor for viral host protein-protein interaction prediction in multiple hosts and viruses," *Frontiers in Medicine*, vol. 9, 2022.

[19] S. Tsukiyama, M. M. Hasan, S. Fujii, and H. Kurata, "LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec," *Briefings in Bioinformatics*, vol. 22, p. bbab228, 06 2021.

[20] P. Xie, J. Zhuang, G. Tian, and J. Yang, "Emvirus: An embedding-based neural framework for human-virus protein-protein interactions prediction," *Biosafety and Health*, vol. 5, no. 3, pp. 152–158, 2023.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[22] D. Neumann, S. Roy, F. U. A. A. Minhas, and A. Ben-Hur, "On the choice of negative examples for prediction of host-pathogen protein interactions," *Frontiers in Bioinformatics*, vol. 2, Dec 2022.

[23] Z. Yang, M. Ding, T. Huang, Y. Cen, J. Song, B. Xu, Y. Dong, and J. Tang, "Does negative sampling matter? a review with insights into its theory and applications,"

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5692–5711, 2024.

[24] I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, pp. 303–305, 01 2002.

[25] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni, "Mint, the molecular interaction database: 2012 update," *Nucleic Acids Research*, vol. 40, pp. D857–D861, 11 2011.

[26] R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-aryamontri, K. Dolinski, and M. Tyers, "The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Science*, vol. 30, no. 1, pp. 187–200, 2021.

[27] N. del Toro, A. Shrivastava, E. Ragueneau, B. Meldal, C. Combe, E. Barrera, L. Perfetto, K. How, P. Ratan, G. Shirodkar, O. Lu, B. Mészáros, X. Watkins, S. Pundir, L. Licata, M. Iannuccelli, M. Pellegrini, M. J. Martin, S. Panni, M. Duesbury, S. Vallet, J. Rappsilber, S. Ricard-Blum, G. Cesareni, L. Salwinski, S. Orchard, P. Porras, K. Panneerselvam, and H. Hermjakob, "The intact database: efficient access to fine-grained molecular interaction data," *Nucleic Acids Research*, vol. 50, pp. D648–D653, 11 2021.

[28] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. Doncheva, S. Pyysalo, P. Bork, L. Jensen, and C. von Mering, "The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Research*, vol. 51, pp. D638–D646, 11 2022.

[29] H. Iuchi, J. Kawasaki, K. Kubo, T. Fukunaga, K. Hokao, G. Yokoyama, A. Ichinose, K. Suga, and M. Hamada, "Bioinformatics approaches for unveiling virus-host interactions," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 1774–1784, 2023.

[30] P. Blohm, G. Frishman, P. Smialowski, F. Goebels, B. Wachinger, A. Ruepp, and D. Frishman, "Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis," *Nucleic acids research*, vol. 42, no. D1, pp. D396–D400, 2014.

[31] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein–protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[32] X.-W. Chen and M. Liu, "Prediction of protein–protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, pp. 4394–4400, 10 2005.

[33] W. Liu-Wei, c. Kafkas, J. Chen, N. J. Dimonaco, J. Tegnér, and R. Hoehndorf, "DeepViral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes," *Bioinformatics*, vol. 37, pp. 2722–2729, 03 2021.

[34] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, pp. 3025–3030, 04 2008.

[35] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC Bioinformatics*, vol. 18, May 2017.

[36] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "Deepppi : Boosting prediction of protein-protein interactions with deep neural networks," *Journal of chemical information and modeling*, vol. 57, 05 2017.

[37] F.-E. Eid, M. ElHefnawi, and L. S. Heath, "Denovo: virus-host sequence-based protein–protein interaction prediction," *Bioinformatics*, vol. 32, p. 1144–1150, Dec 2015.

[38] X. Yang, S. Wuchty, Z. Liang, L. Ji, B. Wang, J. Zhu, Z. Zhang, and Y. Dong, "Multimodal features-based human-herpesvirus protein–protein interaction prediction by using lightgbm," *Briefings in Bioinformatics*, vol. 25, p. bbae005, 01 2024.

[39] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," 2017.

[40] F. Li, S. Dong, A. Leier, M. Han, X. Guo, J. Xu, X. Wang, S. Pan, C. Jia, Y. Zhang, G. I. Webb, L. J. M. Coin, C. Li, and J. Song, "Positive-unlabeled learning in bioinformatics and computational biology: a brief review," *Briefings in Bioinformatics*, vol. 23, p. bbab461, 11 2021.

[41] P. D. Sun, C. E. Foster, and J. C. Boyington, "Overview of protein structural and functional folds," *Current protocols in protein science*, vol. 35, no. 1, pp. 17–1, 2004.

[42] A. Casadevall and L. anne Pirofski, "Host-pathogen interactions: Basic concepts of microbial commensalism, colonization, infection, and disease," *Infection and Immunity*, vol. 68, no. 12, pp. 6511–6518, 2000.

[43] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, p. 5–32, Oct. 2001.

[44] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, (New York, NY, USA), p. 262–270, Association for Computing Machinery, 2012.

[45] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958.

[46] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, p. 273–297, Sept. 1995.

[47] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), p. 1177–1178, Association for Computing Machinery, 2010.

[48] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?," *Journal of Classification*, vol. 274–295, Oct 2014.

[49] P. Gage, "A new algorithm for data compression," *The C Users Journal archive*, vol. 12, pp. 23–38, 1994.

[50] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," 2018.

[51] B. Wang, C. Mei, Y. Wang, Y. Zhou, and Cheng, "Imbalance data processing strategy for protein interaction sites prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, p. 985–994, May 2021.

[52] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein–protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[53] S. Xu, Z. Li, S. Zhang, and J. Hu, "Primary structure similarity analysis of proteins sequences by a new graphical representation," *SAR and QSAR in Environmental Research*, vol. 25, p. 791–803, Sep 2014.

[54] A. Zielezinski and Girgis, "Benchmarking of alignment-free sequence comparison methods," *Genome Biology*, vol. 20, Jul 2019.

[55] S. Durmuş Tekir, T. Çakır, E. Ardıç, A. S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür, F. E. Sevilgen, and K. Ö. Ülgen, "PHISTO: pathogen–host interaction search tool," *Bioinformatics*, vol. 29, pp. 1357–1358, 03 2013.

[56] M. G. Ammari, C. R. Gresham, F. M. McCarthy, and B. Nanduri, "Hpidb 2.0: a curated database for host–pathogen interactions," *Database*, vol. 2016, p. baw103, 2016.

[57] X. Hu, C. Feng, Y. Zhou, A. Harrison, and M. Chen, "DeepTrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks," *Bioinformatics*, vol. 38, pp. 694–702, 10 2021.

APPENDICES

APPENDIX A. DETAILED EXPERIMENTAL RESULTS

Table A.1: *Performance comparison of different negative sampling strategies on the PHISTO dataset using the GA^{2}M model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	99.07	94.94	97.95	92.08	99.80	99.20	99.64
Closest & Farthest	Canberra	2	99.27	95.97	97.76	94.32	99.80	99.42	99.82
Farthest	Canberra	2	99.92	**99.60**	99.82	**99.42**	100.00	99.92	100.00
Uniform	Canberra	2	98.57	91.96	96.10	88.20	99.65	98.82	99.42
Closest	Cosine	2	99.72	**98.56**	99.40	**97.71**	99.90	99.80	99.94
Closest & Farthest	Cosine	2	98.72	92.76	95.41	90.26	99.58	99.04	99.58
Farthest	Cosine	2	98.47	91.46	95.59	87.70	99.60	98.78	99.40
Uniform	Cosine	2	99.58	97.58	98.74	96.48	99.90	99.64	99.90
Closest	Euclidean	2	99.26	95.78	98.53	93.22	99.90	99.30	99.72
Closest & Farthest	Euclidean	2	99.30	96.07	98.37	93.88	99.86	99.40	99.70
Farthest	Euclidean	2	99.60	**97.92**	99.36	**96.54**	99.94	99.63	99.90
Uniform	Euclidean	2	98.84	93.26	97.40	89.46	99.78	98.96	99.52
Wang's	-	2	99.26	95.87	98.53	93.36	99.90	99.34	99.64
Wei's	-	2	99.60	**97.73**	99.75	**95.81**	99.97	99.58	99.83
Random Sampling	-	-	95.49	73.32	79.66	67.94	98.26	96.86	96.90
Dissimilarity-based	-	-	95.25	71.78	78.03	66.45	98.12	96.69	96.74
Closest	Canberra	4	99.10	94.90	97.98	92.02	99.80	99.20	99.64
Closest & Farthest	Canberra	4	99.30	95.96	97.77	94.19	99.80	99.42	99.82
Farthest	Canberra	4	99.90	**99.62**	99.80	**99.45**	100.00	99.92	100.00
Uniform	Canberra	4	98.57	91.88	96.00	88.12	99.62	98.82	99.34
Closest	Cosine	4	99.72	**98.43**	99.32	**97.60**	99.90	99.76	99.90
Closest & Farthest	Cosine	4	98.68	92.46	95.28	89.81	99.54	98.98	99.58
Farthest	Cosine	4	98.53	91.64	95.44	88.12	99.58	98.84	99.44
Uniform	Cosine	4	99.52	97.43	98.60	96.28	99.87	99.62	99.90
Closest	Euclidean	4	99.24	95.74	98.53	93.12	99.86	99.30	99.70
Closest & Farthest	Euclidean	4	99.32	96.16	98.36	94.00	99.82	99.40	99.70
Farthest	Euclidean	4	99.70	**98.32**	99.55	**97.08**	99.98	99.70	99.90
Uniform	Euclidean	4	98.88	93.46	97.47	89.74	99.78	99.00	99.52
Wang's	-	4	99.32	96.18	98.56	93.95	99.87	99.37	99.68
Wei's	-	4	99.61	**97.84**	99.71	**96.08**	99.97	99.60	99.84
Random Sampling	-	-	95.49	73.32	79.66	67.94	98.26	96.86	96.90
Dissimilarity-based	-	-	95.25	71.78	78.03	66.45	98.12	96.69	96.74
Closest	Canberra	8	99.00	94.31	97.68	91.20	99.80	99.10	99.55
Closest & Farthest	Canberra	8	99.20	95.61	97.70	93.64	99.78	99.36	99.80
Farthest	Canberra	8	99.92	**99.64**	99.76	**99.52**	100.00	99.96	100.00
Uniform	Canberra	8	98.54	91.56	95.58	87.85	99.60	98.79	99.32
Closest	Cosine	8	99.78	**98.68**	99.48	**97.86**	99.97	99.80	99.94
Closest & Farthest	Cosine	8	98.60	92.04	94.69	89.54	99.49	98.98	99.58
Farthest	Cosine	8	98.41	90.99	95.10	87.30	99.54	98.75	99.37
Uniform	Cosine	8	99.62	97.92	98.98	96.92	99.90	99.70	99.90
Closest	Euclidean	8	99.16	95.26	98.18	92.52	99.80	99.27	99.70
Closest & Farthest	Euclidean	8	99.34	96.24	98.10	94.43	99.80	99.44	99.76
Farthest	Euclidean	8	99.72	**98.56**	99.62	**97.50**	100.00	99.76	99.87
Uniform	Euclidean	8	98.76	92.82	97.00	89.00	99.72	98.90	99.49
Wang's	-	8	99.00	94.28	97.15	91.58	99.72	99.17	99.54
Wei's	-	8	99.86	**99.22**	99.60	**98.90**	99.96	99.88	99.99
Random Sampling	-	-	95.49	73.32	79.66	67.94	98.26	96.86	96.90
Dissimilarity-based	-	-	95.25	71.78	78.03	66.45	98.12	96.69	96.74

Table A.2: *Performance comparison of different negative sampling strategies on the PHISTO dataset using the LR model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	99.10	94.90	98.64	91.46	99.90	99.14	99.65
Closest & Farthest	Canberra	2	99.34	96.34	98.75	94.05	99.90	99.40	99.82
Farthest	Canberra	2	99.90	**99.62**	99.87	**99.30**	100.00	99.90	100.00
Uniform	Canberra	2	98.65	92.36	97.46	87.72	99.80	98.78	99.48
Closest	Cosine	2	99.80	**98.82**	99.78	**97.84**	100.00	99.80	99.86
Closest & Farthest	Cosine	2	98.85	93.44	97.19	89.92	99.74	99.00	99.68
Farthest	Cosine	2	98.68	92.26	97.15	87.85	99.74	98.78	99.49
Uniform	Cosine	2	99.65	98.02	99.42	96.67	99.92	99.68	99.90
Closest	Euclidean	2	99.30	96.16	99.20	93.28	99.90	99.34	99.69
Closest & Farthest	Euclidean	2	99.38	96.44	99.10	93.88	99.90	99.40	99.68
Farthest	Euclidean	2	99.78	**98.60**	99.82	**97.46**	100.00	99.74	99.74
Uniform	Euclidean	2	99.02	94.33	98.46	90.56	99.87	99.05	99.54
Wang's	-	2	99.30	95.96	99.11	93.01	99.90	99.32	99.62
Wei's	-	2	99.49	**97.14**	99.65	**94.71**	99.96	99.47	99.79
Random Sampling	-	-	96.54	79.39	86.30	73.52	98.83	97.38	98.14
Dissimilarity-based	-	-	96.45	78.81	86.33	72.50	98.85	97.29	98.23
Closest	Canberra	4	99.10	94.88	98.69	91.34	99.90	99.14	99.65
Closest & Farthest	Canberra	4	99.30	96.21	98.58	93.95	99.90	99.40	99.82
Farthest	Canberra	4	99.90	**99.62**	99.90	**99.40**	100.00	99.90	100.00
Uniform	Canberra	4	98.65	92.21	97.56	87.44	99.80	98.75	99.44
Closest	Cosine	4	99.80	**98.84**	99.80	**97.85**	100.00	99.80	99.90
Closest & Farthest	Cosine	4	98.85	93.44	97.29	89.87	99.76	99.00	99.68
Farthest	Cosine	4	98.68	92.41	97.19	88.08	99.76	98.82	99.49
Uniform	Cosine	4	99.63	98.02	99.38	96.72	99.94	99.70	99.90
Closest	Euclidean	4	99.32	96.17	99.32	93.28	99.92	99.32	99.68
Closest & Farthest	Euclidean	4	99.40	96.50	99.17	93.98	99.90	99.40	99.72
Farthest	Euclidean	4	99.76	**98.63**	99.74	**97.51**	100.00	99.78	99.72
Uniform	Euclidean	4	99.04	94.48	98.52	90.76	99.87	99.10	99.56
Wang's	-	4	99.36	96.44	99.12	93.91	99.90	99.40	99.70
Wei's	-	4	99.53	**97.39**	99.57	**95.36**	99.95	99.53	99.79
Random Sampling	-	-	96.54	79.39	86.30	73.52	98.83	97.38	98.14
Dissimilarity-based	-	-	96.45	78.81	86.33	72.50	98.85	97.29	98.23
Closest	Canberra	8	99.00	94.38	98.78	90.38	99.90	99.04	99.64
Closest & Farthest	Canberra	8	99.30	95.91	98.67	93.30	99.87	99.32	99.82
Farthest	Canberra	8	99.90	**99.64**	99.85	**99.44**	100.00	99.92	100.00
Uniform	Canberra	8	98.60	91.96	97.72	86.82	99.80	98.69	99.42
Closest	Cosine	8	99.80	**98.88**	99.80	**97.98**	100.00	99.80	99.92
Closest & Farthest	Cosine	8	98.80	93.10	96.86	89.62	99.70	99.00	99.65
Farthest	Cosine	8	98.62	92.02	97.25	87.28	99.78	98.75	99.45
Uniform	Cosine	8	99.70	98.26	99.54	96.97	99.98	99.70	99.92
Closest	Euclidean	8	99.24	95.71	99.04	92.64	99.90	99.28	99.65
Closest & Farthest	Euclidean	8	99.40	96.60	99.07	94.23	99.90	99.42	99.74
Farthest	Euclidean	8	99.80	**98.78**	99.82	**97.76**	100.00	99.80	99.76
Uniform	Euclidean	8	98.96	94.04	98.15	90.24	99.82	99.04	99.54
Wang's	-	8	99.06	94.53	98.50	90.94	99.90	99.10	99.62
Wei's	-	8	99.88	**99.36**	99.70	**99.10**	99.96	99.90	99.96
Random Sampling	-	-	96.54	79.39	86.30	73.52	98.83	97.38	98.14
Dissimilarity-based	-	-	96.45	78.81	86.33	72.50	98.85	97.29	98.23

Table A.3: *Performance comparison of different negative sampling strategies on the PHISTO dataset using the RF model.*

Table A.4: *Performance comparison of different negative sampling strategies on the PHISTO dataset using the LR model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	98.72	92.76	96.52	89.30	99.68	98.94	99.34
Closest & Farthest	Canberra	2	98.82	93.42	95.71	91.18	99.60	99.14	99.58
Farthest	Canberra	2	99.90	**99.44**	99.78	**99.14**	100.00	99.90	100.00
Uniform	Canberra	2	98.05	88.60	94.16	83.67	99.48	98.38	98.88
Closest	Cosine	2	99.52	**97.47**	98.74	**96.28**	99.90	99.62	99.87
Closest & Farthest	Cosine	2	97.95	88.34	92.18	84.82	99.27	98.47	98.98
Farthest	Cosine	2	98.05	88.82	94.04	84.20	99.48	98.43	99.00
Uniform	Cosine	2	99.24	95.70	97.28	94.20	99.74	99.42	99.72
Closest	Euclidean	2	99.02	94.35	97.98	90.99	99.80	99.10	99.58
Closest & Farthest	Euclidean	2	98.95	94.16	97.34	91.16	99.76	99.11	99.34
Farthest	Euclidean	2	99.58	**97.50**	99.58	**95.48**	100.00	99.55	99.80
Uniform	Euclidean	2	98.50	91.29	96.77	86.39	99.70	98.64	99.24
Wang's	-	2	98.89	**93.84**	97.52	**90.46**	99.78	99.02	99.44
Wei's	-	2	98.62	91.99	96.87	87.60	99.70	98.78	99.34
Random Sampling	-	-	94.81	62.34	72.92	54.44	97.98	95.51	94.32
Dissimilarity-based	-	-	93.09	54.78	67.71	46.00	97.80	94.76	93.40
Closest	Canberra	4	98.78	92.90	96.63	89.46	99.68	98.96	99.34
Closest & Farthest	Canberra	4	98.73	92.92	95.38	90.60	99.54	99.06	99.55
Farthest	Canberra	4	99.90	**99.42**	99.72	**99.12**	100.00	99.90	100.00
Uniform	Canberra	4	98.05	88.55	94.28	83.54	99.48	98.38	98.85
Closest	Cosine	4	99.52	**97.40**	98.60	**96.21**	99.86	99.62	99.87
Closest & Farthest	Cosine	4	97.84	87.70	91.79	84.00	99.25	98.42	98.95
Farthest	Cosine	4	98.08	88.86	94.01	84.20	99.48	98.43	99.02
Uniform	Cosine	4	99.26	95.72	97.22	94.27	99.70	99.42	99.72
Closest	Euclidean	4	99.00	94.36	97.85	91.12	99.80	99.10	99.60
Closest & Farthest	Euclidean	4	99.00	94.30	97.46	91.36	99.78	99.14	99.32
Farthest	Euclidean	4	99.58	**97.50**	99.72	**95.41**	100.00	99.55	99.80
Uniform	Euclidean	4	98.56	91.48	96.88	86.64	99.72	98.68	99.26
Wang's	-	4	98.94	94.00	97.48	90.76	99.78	99.07	99.49
Wei's	-	4	99.82	**99.10**	99.94	**98.20**	100.00	99.80	99.90
Random Sampling	-	-	94.81	62.34	72.92	54.44	97.98	95.51	94.32
Dissimilarity-based	-	-	93.09	54.78	67.71	46.00	97.80	94.76	93.40
Closest	Canberra	8	98.57	91.96	96.38	87.95	99.65	98.79	99.20
Closest & Farthest	Canberra	8	98.68	92.68	95.34	90.14	99.55	99.02	99.54
Farthest	Canberra	8	99.90	**99.49**	99.76	**99.24**	100.00	99.90	100.00
Uniform	Canberra	8	97.89	87.78	93.80	82.48	99.45	98.27	98.64
Closest	Cosine	8	99.58	**97.70**	98.75	**96.67**	99.90	99.65	99.90
Closest & Farthest	Cosine	8	97.70	86.80	90.82	83.12	99.16	98.33	98.85
Farthest	Cosine	8	97.85	87.48	93.00	82.60	99.37	98.27	98.88
Uniform	Cosine	8	99.34	96.30	97.82	94.85	99.80	99.48	99.80
Closest	Euclidean	8	98.82	93.26	97.09	89.70	99.72	98.98	99.48
Closest & Farthest	Euclidean	8	98.90	93.85	96.56	91.30	99.70	99.11	99.38
Farthest	Euclidean	8	99.65	**98.14**	99.48	**96.82**	99.94	99.68	99.80
Uniform	Euclidean	8	98.36	90.38	96.26	85.18	99.68	98.53	99.12
Wang's	-	8	98.46	91.24	95.20	87.60	99.55	98.78	99.20
Wei's	-	8	99.79	**98.88**	99.45	**98.37**	99.94	99.83	99.99
Random Sampling	-	-	94.81	62.34	72.92	54.44	97.98	95.51	94.32
Dissimilarity-based	-	-	93.09	54.78	67.71	46.00	97.80	94.76	93.40

Table A.5: *Performance comparison of different negative sampling strategies on the PHISTO dataset using the SVM model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	98.98	94.33	97.57	91.28	99.78	99.12	99.42
Closest & Farthest	Canberra	2	99.27	95.97	97.85	94.20	99.80	99.40	99.72
Farthest	Canberra	2	99.90	**99.54**	99.78	**99.31**	100.00	99.90	100.00
Uniform	Canberra	2	98.50	91.30	95.93	87.05	99.62	98.72	99.11
Closest	Cosine	2	99.70	**98.32**	99.38	**97.32**	99.90	99.72	99.90
Closest & Farthest	Cosine	2	98.60	92.06	95.61	88.72	99.60	98.88	99.38
Farthest	Cosine	2	98.40	90.74	95.68	86.28	99.58	98.65	99.07
Uniform	Cosine	2	99.49	97.19	98.50	95.86	99.86	99.60	99.90
Closest	Euclidean	2	99.17	95.33	98.40	92.48	99.86	99.26	99.42
Closest & Farthest	Euclidean	2	99.26	95.90	98.50	93.44	99.87	99.32	99.45
Farthest	Euclidean	2	99.58	**97.57**	99.70	**95.56**	100.00	99.58	99.68
Uniform	Euclidean	2	98.73	92.78	97.46	88.50	99.80	98.88	99.18
Wang's	-	2	99.11	**95.03**	98.36	**91.94**	99.84	99.20	99.06
Wei's	-	2	99.10	95.00	98.41	91.78	99.86	99.20	98.88
Random Sampling	-	-	95.26	70.97	81.15	63.06	98.65	96.21	95.61
Dissimilarity-based	-	-	95.08	69.07	80.74	60.36	98.55	96.13	95.93
Closest	Canberra	4	99.00	94.37	97.66	91.34	99.80	99.16	99.34
Closest & Farthest	Canberra	4	99.26	95.81	97.68	94.05	99.80	99.40	99.74
Farthest	Canberra	4	99.90	**99.55**	99.75	**99.31**	100.00	99.92	100.00
Uniform	Canberra	4	98.43	90.96	95.76	86.64	99.62	98.68	99.05
Closest	Cosine	4	99.70	**98.36**	99.31	**97.42**	99.90	99.72	99.90
Closest & Farthest	Cosine	4	98.56	91.78	95.44	88.42	99.55	98.88	99.36
Farthest	Cosine	4	98.43	90.88	95.81	86.44	99.62	98.65	99.10
Uniform	Cosine	4	99.49	97.18	98.47	95.94	99.85	99.60	99.87
Closest	Euclidean	4	99.20	95.40	98.44	92.52	99.86	99.26	99.37
Closest & Farthest	Euclidean	4	99.32	96.11	98.58	93.72	99.85	99.38	99.54
Farthest	Euclidean	4	99.58	**97.60**	99.70	**95.56**	100.00	99.58	99.68
Uniform	Euclidean	4	98.73	92.76	97.43	88.50	99.78	98.85	99.22
Wang's	-	4	99.17	95.45	98.10	92.96	99.80	99.30	99.37
Wei's	-	4	99.82	**99.10**	99.94	**98.27**	100.00	99.80	99.80
Random Sampling	-	-	95.26	70.97	81.15	63.06	98.65	96.21	95.61
Dissimilarity-based	-	-	95.08	69.07	80.74	60.36	98.55	96.13	95.93
Closest	Canberra	8	98.98	94.20	97.51	91.06	99.76	99.10	99.44
Closest & Farthest	Canberra	8	99.17	95.44	97.60	93.36	99.80	99.32	99.78
Farthest	Canberra	8	99.90	**99.58**	99.76	**99.40**	100.00	99.92	100.00
Uniform	Canberra	8	98.46	90.96	95.79	86.62	99.62	98.68	99.14
Closest	Cosine	8	99.74	**98.46**	99.32	**97.60**	99.91	99.76	99.92
Closest & Farthest	Cosine	8	98.46	91.28	94.91	87.92	99.52	98.80	99.40
Farthest	Cosine	8	98.30	90.24	95.48	85.55	99.60	98.57	99.10
Uniform	Cosine	8	99.58	97.60	98.82	96.42	99.90	99.62	99.90
Closest	Euclidean	8	99.11	94.96	98.05	92.08	99.80	99.20	99.52
Closest & Farthest	Euclidean	8	99.28	95.95	98.50	93.57	99.87	99.36	99.37
Farthest	Euclidean	8	99.72	**98.58**	99.75	**97.35**	100.00	99.74	99.55
Uniform	Euclidean	8	98.68	92.40	97.12	88.14	99.72	98.82	99.22
Wang's	-	8	98.84	93.42	97.15	89.97	99.72	99.00	99.27
Wei's	-	8	99.82	**98.99**	99.65	**98.36**	99.96	99.83	99.98
Random Sampling	-	-	95.26	70.97	81.15	63.06	98.65	96.21	95.61
Dissimilarity-based	-	-	95.08	69.07	80.74	60.36	98.55	96.13	95.93

Table A.6: *Performance comparison of different negative sampling strategies on the Human-Virus dataset using the GA^{2}M model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	98.22	89.49	96.70	83.30	99.72	98.33	98.76
Closest & Farthest	Canberra	2	97.62	85.97	93.56	79.52	99.45	97.98	98.14
Farthest	Canberra	2	98.52	**91.56**	96.34	**87.25**	99.66	98.72	98.57
Uniform	Canberra	2	97.52	85.20	93.89	77.96	99.49	97.83	98.46
Closest	Cosine	2	98.96	**94.08**	97.50	**90.86**	99.76	99.10	99.40
Closest & Farthest	Cosine	2	97.74	86.58	93.43	80.70	99.44	98.09	98.63
Farthest	Cosine	2	97.92	87.58	95.13	81.09	99.55	98.12	98.70
Uniform	Cosine	2	97.72	86.78	92.02	82.10	99.25	98.22	98.20
Closest	Euclidean	2	98.62	92.08	98.02	86.80	99.82	98.69	99.20
Closest & Farthest	Euclidean	2	98.57	91.94	97.04	87.34	99.74	98.75	99.16
Farthest	Euclidean	2	99.68	**98.30**	99.36	97.25	99.91	99.72	99.87
Uniform	Euclidean	2	97.88	87.35	94.63	81.14	99.54	98.12	98.82
Wang's	-	2	98.75	92.88	97.58	88.62	99.78	98.85	99.32
Wei's	-	2	99.49	**97.19**	99.44	**95.05**	99.92	99.49	99.76
Random Sampling	-	-	94.72	67.32	76.82	59.98	98.18	96.07	95.54
Dissimilarity-based	-	-	93.67	59.15	71.65	50.37	98.00	95.18	94.05
Closest	Canberra	4	98.08	88.69	95.74	82.60	99.62	98.27	98.64
Closest & Farthest	Canberra	4	97.43	84.82	92.38	78.40	99.34	97.88	98.02
Farthest	Canberra	4	98.36	**90.50**	95.28	**86.18**	99.58	98.64	98.38
Uniform	Canberra	4	97.34	83.85	92.78	76.50	99.40	97.72	98.23
Closest	Cosine	4	98.65	**92.22**	96.13	**88.70**	99.62	98.88	99.20
Closest & Farthest	Cosine	4	97.42	84.65	92.20	78.30	99.34	97.84	98.20
Farthest	Cosine	4	97.67	86.14	93.91	79.58	99.49	97.98	98.47
Uniform	Cosine	4	97.54	85.76	90.67	81.34	99.17	98.16	98.36
Closest	Euclidean	4	98.52	91.26	97.38	85.90	99.78	98.60	99.10
Closest & Farthest	Euclidean	4	98.64	92.16	96.42	88.26	99.65	98.85	99.34
Farthest	Euclidean	4	99.54	**97.39**	98.90	**96.00**	99.90	99.58	99.84
Uniform	Euclidean	4	97.85	87.30	94.77	80.96	99.56	98.12	98.62
Wang's	-	4	98.23	89.68	95.21	84.75	99.58	98.48	98.94
Wei's	-	4	99.58	**97.66**	99.58	**95.80**	99.98	99.55	99.85
Random Sampling	-	-	94.72	67.32	76.82	59.98	98.18	96.07	95.54
Dissimilarity-based	-	-	93.67	59.15	71.65	50.37	98.00	95.18	94.05
Closest	Canberra	8	98.06	88.54	95.26	82.73	99.58	98.30	98.64
Closest & Farthest	Canberra	8	97.39	84.64	91.72	78.58	99.28	97.89	97.94
Farthest	Canberra	8	98.60	**91.82**	96.12	**87.95**	99.68	98.82	98.85
Uniform	Canberra	8	97.12	82.68	91.78	75.25	99.32	97.57	97.89
Closest	Cosine	8	97.62	**86.12**	92.42	**80.61**	99.36	98.08	98.52
Closest & Farthest	Cosine	8	96.50	79.10	87.02	72.43	98.92	97.28	97.53
Farthest	Cosine	8	97.56	85.54	92.50	79.56	99.34	98.00	98.46
Uniform	Cosine	8	96.35	78.06	86.80	70.96	98.92	97.15	97.04
Closest	Euclidean	8	98.47	91.09	96.40	86.32	99.68	98.64	98.98
Closest & Farthest	Euclidean	8	97.56	85.62	93.10	79.26	99.42	97.95	98.14
Farthest	Euclidean	8	98.76	**92.86**	97.34	**88.78**	99.75	98.88	99.06
Uniform	Euclidean	8	97.43	84.84	91.72	78.92	99.27	97.91	98.32
Wang's	-	8	98.19	89.42	95.43	84.12	99.58	98.43	98.78
Wei's	-	8	99.58	**97.66**	99.58	**95.80**	99.98	99.55	99.85
Random Sampling	-	-	94.72	67.32	76.82	59.98	98.18	96.07	95.54
Dissimilarity-based	-	-	93.67	59.15	71.65	50.37	98.00	95.18	94.05

Table A.7: *Performance comparison of different negative sampling strategies on the Human-Virus dataset using the LR model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	97.67	85.93	95.96	77.78	99.68	97.82	97.76
Closest & Farthest	Canberra	2	96.24	76.38	89.2	66.79	99.2	96.75	94.86
Farthest	Canberra	2	98.18	**89.10**	95.92	**83.18**	99.62	98.38	97.3
Uniform	Canberra	2	96.92	80.50	94.33	70.20	99.6	97.08	97.04
Closest	Cosine	2	98.27	**90.02**	96.21	**84.54**	99.65	98.46	97.89
Closest & Farthest	Cosine	2	96.18	75.84	89.22	65.96	99.2	96.7	95.36
Farthest	Cosine	2	97.42	84.46	93.94	76.76	99.49	97.72	97.88
Uniform	Cosine	2	96.72	80.30	88.74	73.27	99.06	97.38	95.83
Closest	Euclidean	2	98.33	90.07	97.72	83.54	99.8	98.38	98.63
Closest & Farthest	Euclidean	2	97.82	87.04	94.94	80.36	99.58	98.08	97.92
Farthest	Euclidean	2	99.54	**97.28**	99.06	**95.58**	99.9	99.55	99.82
Uniform	Euclidean	2	97.42	84.36	94.63	76.12	99.55	97.66	97.66
Wang's	-	2	98.18	89.07	97.03	82.3	99.74	98.27	98.6
Wei's	-	2	99.42	**96.80**	99.47	**94.28**	99.94	99.42	99.62
Random Sampling	-	-	92.76	48.59	68.78	37.6	98.3	94.04	90.02
Dissimilarity-based	-	-	91.95	38.30	63.32	27.46	98.40	93.13	89.12
Closest	Canberra	4	97.35	83.92	94.33	75.56	99.54	97.62	97.52
Closest & Farthest	Canberra	4	95.91	74.30	87.25	64.68	99.06	96.58	94.48
Farthest	Canberra	4	98.02	**88.35**	94.43	**83.00**	99.49	98.33	97.29
Uniform	Canberra	4	96.58	78.36	92.86	67.74	99.45	96.86	96.38
Closest	Cosine	4	97.76	**86.81**	92.96	**81.41**	99.38	98.16	97.66
Closest & Farthest	Cosine	4	95.86	73.05	89.3	61.8	99.27	96.28	94.36
Farthest	Cosine	4	97.08	81.97	93.2	73.16	99.46	97.38	97.09
Uniform	Cosine	4	96.07	76.28	84.33	69.56	98.69	96.98	95.54
Closest	Euclidean	4	98.03	88.36	97.05	81.06	99.74	98.14	98.12
Closest & Farthest	Euclidean	4	97.8	87.00	94.37	80.64	99.52	98.08	98.22
Farthest	Euclidean	4	99.24	**95.90**	97.63	**94.23**	99.78	99.44	99.78
Uniform	Euclidean	4	97.22	83.01	94.26	74.20	99.55	97.45	97.25
Wang's	-	4	97.38	84.21	92.9	77.02	99.4	97.74	98.03
Wei's	-	4	99.37	**96.52**	99.16	**94.04**	99.9	99.4	99.8
Random Sampling	-	-	92.76	48.59	68.78	37.6	98.3	94.04	90.02
Dissimilarity-based	-	-	91.95	38.30	63.32	27.46	98.40	93.13	89.12
Closest	Canberra	8	96.86	80.32	94.23	70.0	99.58	97.06	96.29
Closest & Farthest	Canberra	8	95.9	73.98	87.05	64.3	99.04	96.5	94.28
Farthest	Canberra	8	97.98	**88.00**	95.38	**81.66**	99.58	98.22	97.32
Uniform	Canberra	8	96.03	74.05	92.14	61.9	99.49	96.32	95.49
Closest	Cosine	8	96.82	**80.68**	90.19	**72.98**	99.22	97.34	96.03
Closest & Farthest	Cosine	8	94.51	63.76	80.06	53.0	98.68	95.45	92.59
Farthest	Cosine	8	96.62	78.66	91.66	68.94	99.36	96.98	96.5
Uniform	Cosine	8	95.45	71.60	82.9	63.0	98.7	96.38	93.91
Closest	Euclidean	8	97.42	84.36	93.42	76.96	99.46	97.72	97.58
Closest & Farthest	Euclidean	8	95.41	70.16	85.94	59.25	99.02	96.06	94.42
Farthest	Euclidean	8	98.05	**88.60**	95.27	**82.83**	99.58	98.31	97.96
Uniform	Euclidean	8	96	74.54	88.17	64.58	99.14	96.54	95.92
Wang's	-	8	97.02	81.84	91.46	74.04	99.32	97.47	97.5
Wei's	-	8	99.37	**96.52**	99.16	**94.04**	99.9	99.94	99.8
Random Sampling	-	-	92.76	48.59	68.78	37.6	98.3	94.04	90.02
Dissimilarity-based	-	-	91.95	38.30	63.32	27.46	98.40	93.13	89.12

Table A.8: *Performance comparison of different negative sampling strategies on the Human-Virus dataset using the RF model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	98.2	89.22	98.18	81.76	99.85	98.18	98.56
Closest & Farthest	Canberra	2	97.63	85.55	96.67	76.7	99.74	97.72	98.05
Farthest	Canberra	2	98.7	**92.34**	98.36	**87.04**	99.85	98.72	98.73
Uniform	Canberra	2	97.65	85.59	97.06	76.56	99.78	97.7	98.35
Closest	Cosine	2	98.96	**94.20**	98.72	**90.02**	99.87	99.02	99.3
Closest & Farthest	Cosine	2	97.94	87.72	96.29	80.52	99.72	98.09	98.78
Farthest	Cosine	2	97.92	87.70	96.48	80.39	99.70	98.05	98.65
Uniform	Cosine	2	98.18	89.36	95.91	83.62	99.63	98.4	98.65
Closest	Euclidean	2	98.63	92	98.8	86.08	99.9	98.62	99
Closest & Farthest	Euclidean	2	98.6	91.92	98.36	86.28	99.84	98.64	99.22
Farthest	Euclidean	2	99.69	**98.36**	99.65	**97.12**	99.98	99.72	99.66
Uniform	Euclidean	2	97.95	87.78	96.77	80.38	99.72	98.03	98.65
Wang's	-	2	98.65	92.12	98.44	86.52	99.86	98.65	99.24
Wei's	-	2	99.52	**97.38**	99.62	**95.23**	99.98	99.54	99.45
Random Sampling	-	-	95.51	72.32	82.86	64.16	98.65	96.5	96.38
Dissimilarity-based	-	-	94.66	65.47	79.60	55.60	98.57	95.69	95.34
Closest	Canberra	4	98.1	88.46	97.92	80.7	99.8	98.08	98.42
Closest & Farthest	Canberra	4	97.54	84.92	95.59	76.42	99.66	97.7	98.2
Farthest	Canberra	4	98.64	**92.06**	97.88	**86.88**	99.82	98.7	98.72
Uniform	Canberra	4	97.4	83.85	96.2	74.34	99.7	97.48	98.02
Closest	Cosine	4	98.78	**92.98**	98.08	**88.38**	99.82	98.85	99.16
Closest & Farthest	Cosine	4	97.58	85.13	95.4	76.86	99.62	97.73	98.5
Farthest	Cosine	4	97.94	87.57	96.2	80.34	99.69	98.05	98.68
Uniform	Cosine	4	97.89	87.68	94.3	81.96	99.49	98.22	98.38
Closest	Euclidean	4	98.62	91.70	98.68	85.58	99.9	98.57	98.98
Closest & Farthest	Euclidean	4	98.68	92.30	97.97	87.25	99.82	98.75	99.3
Farthest	Euclidean	4	99.6	**97.85**	99.4	**96.33**	99.94	99.62	99.65
Uniform	Euclidean	4	98.02	88.04	97.35	80.32	99.80	98.08	98.65
Wang's	-	4	98.23	89.76	96.94	83.53	99.72	98.38	98.84
Wei's	-	4	99.62	**97.80**	99.68	**96.02**	100.00	99.62	99.62
Random Sampling	-	-	95.51	72.32	82.86	64.16	98.65	96.5	96.38
Dissimilarity-based	-	-	94.66	65.47	79.60	55.60	98.57	95.69	95.34
Closest	Canberra	8	97.95	87.74	97.56	79.72	99.82	97.99	98.4
Closest & Farthest	Canberra	8	97.47	84.62	95.93	75.69	99.66	97.62	98.2
Farthest	Canberra	8	98.67	**92.31**	98.18	**87.12**	99.84	98.72	98.72
Uniform	Canberra	8	97.25	82.93	95.87	73.08	99.68	97.36	97.89
Closest	Cosine	8	97.84	**87.05**	95.33	**80.08**	99.62	98.04	98.4
Closest & Farthest	Cosine	8	96.77	80.09	92.00	70.92	99.38	97.13	97.85
Farthest	Cosine	8	97.76	86.32	95.58	78.74	99.63	97.89	98.38
Uniform	Cosine	8	97.05	81.98	91.72	74.16	99.31	97.46	97.76
Closest	Euclidean	8	98.50	91.18	98.19	85.13	99.82	98.53	98.82
Closest & Farthest	Euclidean	8	97.34	83.68	95.32	74.54	99.60	97.50	98.56
Farthest	Euclidean	8	98.78	**92.92**	98.74	**87.78**	99.9	98.78	99.14
Uniform	Euclidean	8	97.57	85.20	95.30	77.08	99.64	97.73	98.38
Wang's	-	8	98.14	89.00	96.63	82.48	99.74	98.26	98.85
Wei's	-	8	99.62	**97.80**	99.68	**96.02**	100.00	99.62	99.62
Random Sampling	-	-	95.51	72.32	82.86	64.16	98.65	96.5	96.38
Dissimilarity-based	-	-	94.66	65.47	79.60	55.60	98.57	95.69	95.34

Table A.9: *Performance comparison of different negative sampling strategies on the Human-Virus dataset using the SVM model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	98.08	88.52	96.34	81.91	99.7	98.2	97.74
Closest & Farthest	Canberra	2	97.28	83.31	93.56	75.16	99.49	97.56	97.28
Farthest	Canberra	2	98.46	**90.80**	97.42	**84.99**	99.78	98.52	97.47
Uniform	Canberra	2	97.25	83.18	94.5	74.28	99.58	97.47	97.35
Closest	Cosine	2	98.92	**93.76**	97.82	**90.06**	99.78	99.02	99.18
Closest & Farthest	Cosine	2	97.76	86.58	93.88	80.36	99.48	98.05	98.56
Farthest	Cosine	2	97.51	84.97	95.28	76.68	99.62	97.72	97.82
Uniform	Cosine	2	97.82	87.19	94.04	81.3	99.45	98.14	97.98
Closest	Euclidean	2	98.43	90.80	97.34	85.10	99.75	98.53	98.02
Closest & Farthest	Euclidean	2	98.3	90.07	96.21	84.72	99.65	98.50	98.53
Farthest	Euclidean	2	99.62	**97.88**	99.48	**96.29**	99.96	99.64	99.62
Uniform	Euclidean	2	97.56	85.20	95.17	77.08	99.60	97.76	97.71
Wang's	-	2	98.5	91.14	97.3	85.72	99.78	98.56	98.5
Wei's	-	2	99.45	**97.00**	99.45	**94.68**	99.94	99.48	99.02
Random Sampling	-	-	93.76	56.74	76.74	45.02	98.64	94.74	93.32
Dissimilarity-based	-	-	92.66	45.38	70.20	33.53	98.57	93.68	90.81
Closest	Canberra	4	97.88	87.30	95.30	80.54	99.59	98.08	97.62
Closest & Farthest	Canberra	4	97.09	82.39	91.88	74.76	99.34	97.51	97.34
Farthest	Canberra	4	98.26	**89.74**	96.63	**83.75**	99.72	98.40	97.15
Uniform	Canberra	4	97.00	81.16	93.49	71.7	99.49	97.23	97.18
Closest	Cosine	4	98.53	**91.54**	96.80	**86.80**	99.72	98.70	98.50
Closest & Farthest	Cosine	4	97.22	83.24	92.48	75.7	99.37	97.62	97.92
Farthest	Cosine	4	97.43	84.42	94.56	76.28	99.55	97.68	98.02
Uniform	Cosine	4	97.45	84.88	92.18	78.62	99.34	97.9	97.63
Closest	Euclidean	4	98.31	90.04	96.93	84.04	99.72	98.4	98.06
Closest & Farthest	Euclidean	4	98.3	90.14	95.43	85.38	99.6	98.53	98.82
Farthest	Euclidean	4	99.45	**97.00**	98.73	**95.33**	99.90	99.52	99.65
Uniform	Euclidean	4	97.57	85.39	95.10	77.50	99.60	97.82	97.70
Wang's	-	4	97.7	86.17	94.60	79.16	99.54	97.94	97.60
Wei's	-	4	99.49	**97.18**	99.26	**95.20**	99.90	99.54	99.55
Random Sampling	-	-	93.76	56.74	76.74	45.02	98.64	94.74	93.32
Dissimilarity-based	-	-	92.66	45.38	70.20	33.53	98.57	93.68	90.81
Closest	Canberra	8	97.72	86.46	94.26	79.88	99.52	98.02	97.83
Closest & Farthest	Canberra	8	97.12	82.57	91.32	75.34	99.3	97.57	97.7
Farthest	Canberra	8	98.40	**90.76**	96.58	**85.58**	99.7	98.57	98.08
Uniform	Canberra	8	96.87	80.84	92.04	72.05	99.37	97.25	97.02
Closest	Cosine	8	97.45	**84.63**	93.24	**77.50**	99.42	97.78	97.68
Closest & Farthest	Cosine	8	95.88	74.00	86.56	64.62	99.00	96.58	96.32
Farthest	Cosine	8	97.24	83.30	93.26	75.27	99.44	97.57	97.30
Uniform	Cosine	8	96.30	77.00	88.74	67.99	99.11	96.87	96.35
Closest	Euclidean	8	98.32	90.06	95.90	84.9	99.65	98.48	98.32
Closest & Farthest	Euclidean	8	96.44	77.76	89.86	68.54	99.22	96.94	96.67
Farthest	Euclidean	8	98.60	**91.92**	97.05	**87.28**	99.74	98.75	98.68
Uniform	Euclidean	8	96.96	81.70	91.02	73.98	99.27	97.43	97.34
Wang's	-	8	97.45	84.90	93.30	77.84	99.44	97.82	97.50
Wei's	-	8	99.49	**97.18**	99.26	**95.20**	99.90	99.54	99.55
Random Sampling	-	-	93.76	56.74	76.74	45.02	98.64	94.74	93.32
Dissimilarity-based	-	-	92.66	45.38	70.20	33.53	98.57	93.68	90.81

Table A.10: *Performance comparison of different negative sampling strategies on the Tsukiyama's dataset using the GA^{2}M model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	2	Canberra	99.3	96.00	98.85	93.30	99.90	99.34	99.68
Closest & Farthest	2	Canberra	97.24	74.70	98.58	71.18	99.84	97.28	89.34
Farthest	2	Canberra	99.3	**95.97**	99.45	**92.76**	99.96	99.3	99.2
Uniform	2	Canberra	98.85	93.49	97.6	89.69	99.8	98.96	99.4
Closest	2	Cosine	98.9	**93.68**	98.80	**89.06**	99.90	98.90	98.87
Closest & Farthest	2	Cosine	92.25	79.46	80.99	88.05	92.68	98.84	98.5
Farthest	2	Cosine	98.79	93.04	97.15	89.22	99.74	98.94	99.44
Uniform	2	Cosine	98.23	89.6	97.10	83.18	99.75	98.36	98.46
Closest	2	Euclidean	99.40	96.72	99.21	94.33	99.9	99.44	99.76
Closest & Farthest	2	Euclidean	99.30	96.12	98.9	93.53	99.9	99.36	99.69
Farthest	2	Euclidean	99.58	**97.68**	99.63	**95.80**	99.98	99.58	99.82
Uniform	2	Euclidean	99.02	94.27	98.26	90.61	99.82	99.07	99.52
Wang's	2	-	99.37	96.48	99.34	93.78	99.90	99.40	99.74
Wei's	2	-	99.42	**96.63**	99.62	**93.85**	100.00	99.40	99.68
Random Sampling	-	-	94.18	63.62	73.51	56.06	97.98	95.69	94.84
Dissimilarity-based	-	-	93.90	61.19	72.64	52.86	98.00	95.41	94.52
Closest	4	Canberra	99.12	95.07	98.85	91.56	99.9	99.16	99.48
Closest & Farthest	4	Canberra	84.75	60.02	51.05	72.82	85.98	97.46	88.4
Farthest	4	Canberra	99.24	**95.71**	99.16	**92.48**	99.9	99.24	99.22
Uniform	4	Canberra	98.69	92.44	97.71	87.75	99.8	98.79	99.2
Closest	4	Cosine	99.00	**94.18**	99.10	**89.76**	99.90	99.00	99.05
Closest & Farthest	4	Cosine	98.36	90.46	95.43	85.91	99.58	98.62	99.04
Farthest	4	Cosine	98.63	92.14	97.05	87.74	99.72	98.75	99.28
Uniform	4	Cosine	98.27	89.94	97.19	83.7	99.78	98.38	98.47
Closest	4	Euclidean	99.32	96.08	99.32	93	99.92	99.3	99.62
Closest & Farthest	4	Euclidean	98.98	94.11	97.84	90.67	99.8	99.1	99.37
Farthest	4	Euclidean	99.30	**96.16**	99.32	**93.16**	99.92	99.3	99.64
Uniform	4	Euclidean	98.92	93.81	98.6	89.46	99.87	98.98	99.37
Wang's	4	-	99.49	**97.09**	99.66	**94.68**	100.00	99.45	99.68
Wei's	4	-	99.37	96.56	99.62	93.68	99.98	99.37	99.64
Random Sampling	-	-	94.18	63.62	73.51	56.06	97.98	95.69	94.84
Dissimilarity-based	-	-	93.90	61.19	72.64	52.86	98.00	95.41	94.52
Closest	8	Canberra	98.78	**92.86**	98.14	**88.12**	99.8	98.82	99.1
Closest & Farthest	8	Canberra	98.36	90.60	96.60	85.28	99.7	98.54	98.88
Farthest	8	Canberra	79.92	61.02	61.34	75.94	80.36	97.64	88.86
Uniform	8	Canberra	98.06	88.55	95.97	82.19	99.65	98.23	98.64
Closest	8	Cosine	63.74	63.04	62.52	93.91	60.73	99.28	88.28
Closest & Farthest	8	Cosine	97.78	87.00	93.72	81.14	99.48	98.14	98.5
Farthest	8	Cosine	98.18	**89.30**	96.10	**83.41**	99.68	98.33	98.82
Uniform	8	Cosine	89.38	77.04	80.23	86.00	89.72	98.52	96.48
Closest	8	Euclidean	98.94	93.91	98.5	89.76	99.87	98.96	99.4
Closest & Farthest	8	Euclidean	98.61	92.02	97.09	87.44	99.74	98.75	99.06
Farthest	8	Euclidean	99.38	**96.58**	99.36	**93.91**	99.94	99.4	99.36
Uniform	8	Euclidean	98.36	90.23	96.49	84.78	99.7	98.5	98.9
Wang's	8	-	99.00	94.32	98.46	90.52	99.87	99.06	99.4
Wei's	8	-	99.64	**97.94**	99.86	**96.10**	100.00	99.6	99.84
Random Sampling	-	-	94.18	63.62	73.51	56.06	97.98	95.69	94.84
Dissimilarity-based	-	-	93.90	61.19	72.64	52.86	98.00	95.41	94.52

Table A.11: *Performance comparison of different negative sampling strategies on the Tsukiyama's dataset using the LR model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	2	Canberra	98.92	93.84	97.6	90.33	99.8	99.04	99.36
Closest & Farthest	2	Canberra	97.83	87.14	94.14	81.12	99.49	98.14	97.63
Farthest	2	Canberra	99.1	**94.85**	99.17	**90.92**	99.9	99.1	98.85
Uniform	2	Canberra	98.4	90.7	96.16	85.77	99.64	98.6	98.92
Closest	2	Cosine	98.68	**92.2**	98.52	**86.64**	99.87	98.68	98.41
Closest & Farthest	2	Cosine	96.96	81.66	90.8	74.17	99.24	97.46	96.88
Farthest	2	Cosine	98.33	90.3	96.56	84.75	99.7	98.5	98.98
Uniform	2	Cosine	97.83	86.92	96.02	79.4	99.7	97.95	97.05
Closest	2	Euclidean	99.26	95.76	99.14	92.58	99.9	99.26	99.55
Closest & Farthest	2	Euclidean	98.94	93.94	97.94	90.24	99.82	99.04	98.84
Farthest	2	Euclidean	99.48	**97.00**	99.2	**94.88**	99.9	99.48	99.65
Uniform	2	Euclidean	98.84	93.3	98.33	88.78	99.84	98.88	99.26
Wang's	2	-	98.85	**93.47**	98.27	**89.12**	99.84	98.94	99.34
Wei's	2	-	98.32	89.94	97.52	83.45	99.8	98.36	98.56
Random Sampling	-	-	92.46	44.78	66.96	33.64	98.32	93.68	89.82
Dissimilarity-based	-	-	91.92	39.08	62.18	28.49	98.26	93.21	89.40
Closest	4	Canberra	98.84	93.28	97.78	89.2	99.8	98.94	99.1
Closest & Farthest	4	Canberra	97.66	86.04	94.36	79.12	99.52	97.95	97.35
Farthest	4	Canberra	99.02	**94.46**	98.94	**90.36**	99.9	99.04	98.85
Uniform	4	Canberra	98.22	89.66	95.81	84.22	99.63	98.41	98.56
Closest	4	Cosine	98.72	**92.46**	98.64	**86.98**	99.9	98.72	98.31
Closest & Farthest	4	Cosine	96.86	80.93	90.4	73.27	99.22	97.38	96.71
Farthest	4	Cosine	98.2	89.52	95.87	83.98	99.63	98.4	98.57
Uniform	4	Cosine	97.85	87.24	95.87	80.02	99.63	98.04	97.24
Closest	4	Euclidean	99.11	**94.94**	98.65	**91.50**	99.87	99.16	99.34
Closest & Farthest	4	Euclidean	97.95	87.84	94.9	81.72	99.55	98.18	98.26
Farthest	4	Euclidean	99.05	94.5	98.31	90.94	99.84	99.1	99.37
Uniform	4	Euclidean	98.64	92.18	97.62	87.32	99.8	98.73	98.94
Wang's	4	-	98.98	**94.00**	98.20	**90.18**	99.82	99.02	99.36
Wei's	4	-	98.14	88.84	96.52	82.31	99.7	98.23	98.52
Random Sampling	-	-	92.46	44.78	66.96	33.64	98.32	93.68	89.82
Dissimilarity-based	-	-	91.92	39.08	62.18	28.49	98.26	93.21	89.40
Closest	8	Canberra	98.12	88.82	96.07	82.6	99.7	98.3	98.08
Closest & Farthest	8	Canberra	97.28	83.46	92.59	75.92	99.37	97.62	96.65
Farthest	8	Canberra	99.02	**94.32**	98.52	**90.48**	99.87	99.04	98.88
Uniform	8	Canberra	97.29	83.54	93.43	75.54	99.48	97.62	97.12
Closest	8	Cosine	98.56	**91.66**	96.94	**86.92**	99.72	98.68	98.38
Closest & Farthest	8	Cosine	96.06	75.3	87.44	66.2	99.05	96.71	95.55
Farthest	8	Cosine	97.19	82.94	92.67	75.06	99.42	95.57	97.4
Uniform	8	Cosine	97.82	86.99	94.81	80.42	99.58	98.05	97.28
Closest	8	Euclidean	98.31	90.24	97.15	84.3	99.76	98.43	98.4
Closest & Farthest	8	Euclidean	96.88	81.12	89.92	73.85	99.16	97.43	96.87
Farthest	8	Euclidean	99.02	**94.46**	97.99	**91.13**	99.8	99.1	99.1
Uniform	8	Euclidean	97.74	86.35	94.94	79.2	99.58	97.95	97.78
Wang's	8	-	98	88.28	95.4	82.19	99.6	98.22	98.6
Wei's	8	-	99.48	**96.93**	99.68	**94.33**	100.00	99.44	99.66
Random Sampling	-	-	92.46	44.78	66.96	33.64	98.32	93.68	89.82
Dissimilarity-based	-	-	91.92	39.08	62.18	28.49	98.26	93.21	89.40

Table A.12: *Performance comparison of different negative sampling strategies on the Tsukiyama's dataset using the RF model.*

Selection Method	Distance Metric	Cluster Number	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	2	Canberra	99.24	95.64	99.28	92.32	99.91	99.26	99.48
Closest & Farthest	2	Canberra	98.9	93.63	98.84	88.94	99.9	98.9	99.32
Farthest	2	Canberra	99.3	**96.02**	99.58	**92.68**	99.98	99.28	99.2
Uniform	2	Canberra	98.82	93.24	98.43	88.55	99.86	98.88	99.22
Closest	2	Cosine	99.1	**94.73**	99.14	**90.66**	99.9	99.1	99.3
Closest & Farthest	2	Cosine	98.43	90.86	97.38	85.15	99.78	98.53	99.17
Farthest	2	Cosine	98.88	93.54	98.03	89.36	99.8	98.92	99.28
Uniform	2	Cosine	98.5	91.2	98.16	85.13	99.82	98.53	98.88
Closest	2	Euclidean	99.42	96.73	99.49	94.14	99.96	99.42	99.52
Closest & Farthest	2	Euclidean	99.36	96.42	99.34	93.7	99.9	99.37	99.58
Farthest	2	Euclidean	99.7	**98.23**	99.62	**96.83**	99.97	99.68	99.74
Uniform	2	Euclidean	99.07	94.65	98.9	90.74	99.9	99.07	99.3
Wang's	2	-	99.32	96.17	99.54	93.06	100.00	99.32	99.54
Wei's	2	-	99.40	**96.62**	99.65	**93.78**	99.96	99.38	99.69
Random Sampling	-	-	95.07	68.84	81.18	59.8	98.6	96.06	96.03
Dissimilarity-based	-	-	94.79	66.44	80.36	56.63	98.61	95.78	95.66
Closest	4	Canberra	99.14	95.12	99.1	91.46	99.9	99.16	99.42
Closest & Farthest	4	Canberra	98.78	92.9	98.64	87.78	99.9	98.79	99.3
Farthest	4	Canberra	99.27	**95.85**	99.3	**92.62**	99.91	99.27	99.22
Uniform	4	Canberra	98.75	92.76	98.31	87.82	99.84	98.79	99.1
Closest	4	Cosine	99.07	**94.61**	99.11	**90.52**	99.9	99.05	99.24
Closest & Farthest	4	Cosine	98.43	90.67	97.42	84.82	99.8	98.52	99.05
Farthest	4	Cosine	98.79	93.06	97.92	86.4	99.8	98.85	99.26
Uniform	4	Cosine	98.5	91.26	98.15	85.24	99.84	98.53	98.76
Closest	4	Euclidean	99.32	96.2	99.4	93.24	99.92	99.32	99.44
Closest & Farthest	4	Euclidean	99.1	94.8	98.96	90.96	99.9	99.1	99.48
Farthest	4	Euclidean	99.4	**96.49**	99.28	**93.85**	99.92	99.4	99.52
Uniform	4	Euclidean	99.02	94.27	98.74	90.19	99.9	99.02	99.24
Wang's	4	-	99.4	96.42	99.58	93.44	100.00	99.34	99.48
Wei's	4	-	98.39	**96.50**	99.82	**93.39**	99.98	99.34	99.74
Random Sampling	-	-	95.07	68.84	81.18	59.8	98.6	96.06	96.03
Dissimilarity-based	-	-	94.79	66.44	80.36	56.63	98.61	95.78	95.66
Closest	8	Canberra	98.74	92.46	98.79	86.9	99.9	98.69	99.86
Closest & Farthest	8	Canberra	98.4	90.57	98.41	83.9	99.87	98.4	99.14
Farthest	8	Canberra	99.3	**96.06**	99.42	**92.92**	99.92	99.3	99.36
Uniform	8	Canberra	98.09	88.65	97.43	81.31	99.8	98.16	98.69
Closest	8	Cosine	99.02	**94.33**	98.79	**90.24**	99.9	99.02	99.12
Closest & Farthest	8	Cosine	97.84	86.92	96.45	79.14	99.7	97.93	98.72
Farthest	8	Cosine	98.22	89.48	97.34	82.73	99.78	98.31	98.8
Uniform	8	Cosine	98.36	90.32	97.46	84.18	99.8	98.46	98.74
Closest	8	Euclidean	98.92	93.76	99.17	88.86	99.92	98.92	99.26
Closest & Farthest	8	Euclidean	98.46	90.9	98.4	84.48	99.9	98.47	99.24
Farthest	8	Euclidean	99.4	**96.62**	99.68	**93.72**	100.00	99.4	99.52
Uniform	8	Euclidean	98.41	90.74	98.18	84.36	99.86	98.46	98.87
Wang's	8	-	98.92	93.78	98.8	89.24	99.9	98.94	99.22
Wei's	8	-	99.58	**97.6**	99.78	**95.54**	100.00	99.58	99.72
Random Sampling	-	-	95.07	68.84	81.18	59.8	98.6	96.06	96.03
Dissimilarity-based	-	-	94.79	66.44	80.36	56.63	98.61	95.78	95.66

Table A.13: *Performance comparison of different negative sampling strategies on the Tsukiyama's dataset using the SVM model.*

Selection	Distance	Cluster	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	99.14	95.22	98.42	92.2	99.86	99.22	99.12
Closest&Farthest	Canberra	2	98.86	93.52	97.14	90.18	99.74	99.02	99.20
Farthest	Canberra	2	99.20	**95.38**	98.98	**92.06**	99.90	99.22	98.48
Uniform	Canberra	2	98.74	92.68	97.36	88.42	99.76	98.84	98.80
Closest	Cosine	2	98.92	**93.78**	98.44	**89.54**	99.86	98.96	98.38
Closest&Farthest	Cosine	2	98.24	89.68	94.96	84.94	99.56	98.52	98.88
Farthest	Cosine	2	98.58	91.80	96.88	87.22	99.70	98.74	98.86
Uniform	Cosine	2	98.14	89.00	96.36	82.64	99.70	98.28	97.36
Closest	Euclidean	2	99.34	96.24	99.08	93.56	99.90	99.36	99.28
Closest&Farthest	Euclidean	2	99.24	95.72	98.66	92.96	99.90	99.32	99.34
Farthest	Euclidean	2	99.54	**97.48**	99.48	**95.58**	99.96	99.54	99.60
Uniform	Euclidean	2	98.96	93.96	98.44	89.84	99.88	99.00	98.78
Wang's	-	2	99.16	**95.20**	98.74	**91.94**	99.90	99.20	99.26
Wei's	-	2	98.62	91.88	97.72	86.74	99.80	98.70	98.66
Random Sampling	-	-	93.22	51.68	73.24	39.90	98.56	94.24	91.74
Dissimilarity-based	-	-	93.33	53.83	72.61	42.77	98.38	94.50	92.29
Closest	Canberra	4	99.06	94.64	98.26	91.26	99.84	99.12	99.04
Closest&Farthest	Canberra	4	98.74	92.84	96.76	89.26	99.72	98.92	99.20
Farthest	Canberra	4	99.14	**95.14**	98.80	**91.72**	99.90	99.18	98.60
Uniform	Canberra	4	98.62	91.88	97.28	87.08	99.76	98.72	98.48
Closest	Cosine	4	98.92	**93.70**	98.48	**89.36**	99.88	98.96	98.14
Closest&Farthest	Cosine	4	98.16	89.18	94.68	84.28	99.54	98.44	98.70
Farthest	Cosine	4	98.50	91.38	96.76	86.60	99.70	98.66	98.74
Uniform	Cosine	4	98.28	89.92	96.76	83.96	99.72	98.44	97.52
Closest	Euclidean	4	99.22	**95.70**	98.86	**92.76**	99.90	99.30	98.96
Closest&Farthest	Euclidean	4	98.88	93.48	97.18	90.10	99.72	99.00	99.26
Farthest	Euclidean	4	99.18	95.36	98.62	92.32	99.88	99.22	99.44
Uniform	Euclidean	4	98.90	93.62	98.30	89.40	99.82	98.96	98.56
Wang's	-	4	99.14	**95.04**	98.48	**91.88**	99.88	99.18	99.20
Wei's	-	4	98.30	89.80	97.34	83.34	99.78	98.36	97.94
Random Sampling	-	-	93.22	51.68	73.24	39.90	98.56	94.24	91.74
Dissimilarity-based	-	-	93.33	53.83	72.61	42.77	98.38	94.50	92.29
Closest	Canberra	8	98.58	91.76	97.28	86.80	99.78	98.70	98.52
Closest&Farthest	Canberra	8	98.30	90.18	95.72	85.20	99.62	98.56	98.62
Farthest	Canberra	8	99.16	**95.38**	98.58	**92.32**	99.88	99.22	99.10
Uniform	Canberra	8	97.90	87.46	95.42	80.76	99.62	98.10	97.60
Closest	Cosine	8	98.82	**93.20**	97.46	**89.30**	99.80	98.96	98.82
Closest&Farthest	Cosine	8	97.60	85.66	92.78	79.58	99.38	97.98	98.26
Farthest	Cosine	8	97.94	87.88	95.14	81.66	99.60	98.20	98.36
Uniform	Cosine	8	98.02	88.34	96.08	81.72	99.68	98.18	97.80
Closest	Euclidean	8	98.82	92.98	98.02	88.44	99.80	98.82	98.88
Closest&Farthest	Euclidean	8	98.28	90.12	95.36	85.44	99.58	98.56	98.90
Farthest	Euclidean	8	99.30	**96.14**	98.78	**93.62**	99.90	99.38	99.36
Uniform	Euclidean	8	98.32	89.96	96.86	84.02	99.72	98.42	98.26
Wang's	-	8	98.58	91.78	96.74	87.30	99.70	98.76	98.64
Wei's	-	8	99.48	**97.04**	99.56	**94.70**	100.00	99.48	99.28
Random Sampling	-	-	93.22	51.68	73.24	39.90	98.56	94.24	91.74
Dissimilarity-based	-	-	93.33	53.83	72.61	42.77	98.38	94.50	92.29

Table A.14: *Performance comparison of different negative sampling strategies on the Herpes dataset using the RF model.*

Selection	Distance	Cluster	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	99.09	**94.77**	99.41	**90.54**	99.95	99.06	99.27
Closest & Farthest	Canberra	2	98.26	89.62	98.12	82.48	99.84	98.27	98.96
Farthest	Canberra	2	99.05	94.57	98.6	90.86	99.87	99.09	99.26
Uniform	Canberra	2	98.48	91.05	98.44	84.68	99.87	98.49	98.69
Closest	Cosine	2	99.07	**94.69**	98.71	**90.99**	99.88	99.11	99.34
Closest & Farthest	Cosine	2	98.19	89.2	97.47	82.23	99.79	98.25	98.8
Farthest	Cosine	2	98.85	93.32	98.83	88.39	99.89	98.85	99.07
Uniform	Cosine	2	98.52	91.28	98.05	85.39	99.83	98.56	99
Closest	Euclidean	2	99.22	95.52	99.44	91.9	99.95	99.2	99.36
Closest & Farthest	Euclidean	2	99.06	94.57	99.09	90.45	99.92	99.05	99.32
Farthest	Euclidean	2	99.56	**97.52**	99.65	**95.47**	99.97	99.55	99.57
Uniform	Euclidean	2	98.65	92.09	98.37	86.57	99.86	98.67	99.01
Wang's	-	-	99.15	95.11	99.24	91.32	99.93	99.14	99.26
Wei's	-	-	99.36	**96.39**	99.59	**93.39**	99.96	99.34	99.66
Random Sampling	-	-	95.09	66.7	86.95	54.13	99.19	95.58	95.33
Dissimilarity-based	-	-	95.44	70.57	85.32	60.19	98.96	96.13	96.53
Closest	Canberra	4	99.04	**94.49**	99.43	**90.02**	99.95	99.01	99.31
Closest & Farthest	Canberra	4	98.22	89.31	98.35	81.8	99.86	98.21	98.91
Farthest	Canberra	4	99.07	94.67	99.14	90.59	99.92	99.07	99.19
Uniform	Canberra	4	98.49	91.05	98.41	84.73	99.86	98.49	98.81
Closest	Cosine	4	98.97	**94.06**	98.65	**89.88**	99.88	99	99.26
Closest & Farthest	Cosine	4	97.96	87.66	97.08	79.91	99.76	98.03	98.64
Farthest	Cosine	4	98.75	92.73	98.6	87.52	99.88	98.76	99.09
Uniform	Cosine	4	98.39	90.48	97.96	84.06	99.83	98.43	98.92
Closest	Euclidean	4	99.17	95.25	99.27	91.53	99.93	99.16	99.25
Closest & Farthest	Euclidean	4	98.9	93.64	98.77	89.01	99.89	98.91	99.27
Farthest	Euclidean	4	99.58	**97.65**	99.81	**95.58**	99.98	99.56	99.6
Uniform	Euclidean	4	98.57	91.65	98	86.08	99.82	98.62	99.03
Wang's	-	-	99.11	**94.9**	99.28	**90.9**	99.93	99.1	99.26
Wei's	-	-	99.07	94.67	99.19	90.54	99.93	99.06	99.56
Random Sampling	-	-	95.09	66.7	86.95	54.13	99.19	95.58	95.33
Dissimilarity-based	-	-	95.44	70.57	85.32	60.19	98.96	96.13	96.53
Closest	Canberra	8	98.98	94.08	99.29	89.4	99.94	98.95	99.26
Closest & Farthest	Canberra	8	98.2	89.25	97.68	82.16	99.8	98.24	98.8
Farthest	Canberra	8	99.04	**94.52**	98.77	**90.62**	99.89	99.07	99.22
Uniform	Canberra	8	98.39	90.44	98.37	83.69	99.86	98.39	98.71
Closest	Cosine	8	98.77	92.85	98.45	87.86	99.86	98.79	
Closest & Farthest	Cosine	8	97.77	86.5	95.96	78.74	99.67	97.91	98.46
Farthest	Cosine	8	98.68	**92.3**	98.56	**86.79**	99.87	98.69	98.98
Uniform	Cosine	8	98.22	89.39	97.61	82.45	99.8	98.27	98.68
Closest	Euclidean	8	99.16	95.23	99.13	91.62	99.92	99.17	99.33
Closest & Farthest	Euclidean	8	98.92	93.78	98.79	89.25	99.89	98.94	99.3
Farthest	Euclidean	8	99.54	**97.43**	99.71	**95.25**	99.97	99.53	99.59
Uniform	Euclidean	8	98.61	91.83	98.39	86.1	99.86	98.63	98.92
Wang's	-	-	98.69	92.33	98.6	86.82	99.88	98.7	98.99
Wei's	-	-	99.64	**97.97**	99.57	**96.41**	99.96	99.64	99.86
Random Sampling	-	-	95.09	66.7	86.95	54.13	99.19	95.58	95.33
Dissimilarity-based	-	-	95.44	70.57	85.32	60.19	98.96	96.13	96.53

Table A.15: *Performance comparison of different negative sampling strategies on the Herpes dataset using the GA2M model.*

Selection	Distance	Cluster	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	99.22	95.55	99.25	92.12	99.93	99.22	99.42
Closest & Farthest	Canberra	2	98.31	90.07	97.02	84.06	99.74	98.43	98.4
Farthest	Canberra	2	96.74	84.6	85.95	86.2	97.78	98.63	94.48
Uniform	Canberra	2	98.62	92	97.33	87.22	99.76	98.73	99.08
Closest	Cosine	2	81.37	68.53	66.39	91.72	80.32	98.98	88.43
Closest & Farthest	Cosine	2	98.18	89.23	96.44	83.03	99.69	98.33	98.45
Farthest	Cosine	2	98.84	93.35	97.81	89.29	99.8	98.94	99.3
Uniform	Cosine	2	89.81	70.62	72.75	73.02	91.49	96.73	86.16
Closest	Euclidean	2	99.24	95.66	98.85	92.68	99.89	99.27	99.59
Closest & Farthest	Euclidean	2	89.87	66.46	65.52	77.64	91.12	97.72	88.51
Farthest	Euclidean	2	82.54	53.72	54.99	75.66	83.31	97.42	79.9
Uniform	Euclidean	2	98.65	92.18	97.28	87.59	99.76	98.77	99.21
Wei's	-	2	99.42	96.74	99.63	94.01	99.96	99.4	99.57
Wang's	-	2	99.17	95.24	99.13	91.65	99.92	99.17	99.41
Dissimilarity-based	-	-	94.64	66.95	76.11	59.78	98.12	96.06	95.63
Random Sampling	-	-	94.56	65.61	77.08	57.13	98.3	95.82	94.86
Closest	Canberra	4	99.11	94.88	99.34	90.8	99.94	99.09	99.23
Closest & Farthest	Canberra	4	88.8	61.65	65.03	69.85	90.65	97.05	87.74
Farthest	Canberra	4	84.51	39.63	38.78	51.08	87.8	94.85	69.44
Uniform	Canberra	4	98.37	90.41	97.07	84.61	99.74	98.48	98.54
Closest	Cosine	4	87.03	53.21	48.12	68.95	88.87	96.64	79.74
Closest & Farthest	Cosine	4	78.61	55.03	-	67.56	79.74	95.16	78.68
Farthest	Cosine	4	98.69	92.39	97.88	87.49	99.81	98.76	99.04
Uniform	Cosine	4	90.91	77.42	80.91	83.96	91.63	98.23	95.42
Closest	Euclidean	4	99.19	95.42	98.91	92.16	99.9	99.22	99.52
Closest & Farthest	Euclidean	4	98.78	92.96	98.11	88.32	99.83	98.84	98.48
Farthest	Euclidean	4	61.97	44.03	41.27	80.15	60.14	97.62	70.15
Wei's	-	4	99.35	96.3	99.49	93.3	99.95	99.33	99.62
Wang's	-	4	99.13	95.04	99.35	91.08	99.94	99.12	99.27
Dissimilarity-based	-	-	94.64	66.95	76.11	59.78	98.12	96.06	95.63
Random Sampling	-	-	94.56	65.61	77.08	57.13	98.3	95.82	94.86
Closest	Canberra	8	99.03	94.37	99.35	89.87	99.94	99	99.03
Closest & Farthest	Canberra	8	96.74	71.22	77.48	65.9	99.78	96.83	88.57
Farthest	Canberra	8	65.25	63.62	63.08	94.1	62.35	99.31	80.71
Uniform	Canberra	8	84.01	67.28	66.01	89.26	83.46	98.81	96.51
Closest	Cosine	8	80.83	41.25	32.8	61.36	82.74	95.43	72.05
Closest & Farthest	Cosine	8	97.68	86.14	94.38	79.22	99.53	97.95	98
Farthest	Cosine	8	98.63	92.01	97.76	86.9	99.8	98.7	98.86
Uniform	Cosine	8	93.75	77.27	81.88	78.43	95.29	97.73	93.37
Closest	Euclidean	8	99.21	95.49	98.84	92.36	99.89	99.24	99.51
Closest & Farthest	Euclidean	8	95.38	62.93	-	59.71	98.97	96.17	81.77
Farthest	Euclidean	8	85.68	46.25	44.76	64.75	87.77	96.3	76.26
Wang's	-	8	98.73	92.6	98.03	87.75	99.82	98.79	98.96
Wei's	-	8	99.66	98.09	99.35	96.87	99.94	99.69	99.94
Dissimilarity-based	-	-	94.64	66.95	76.11	59.78	98.12	96.06	95.63
Random Sampling	-	-	94.56	65.61	77.08	57.13	98.3	95.82	94.86

Table A.16: *Performance comparison of different negative sampling strategies on the Herpes dataset using the LR model.*

Selection	Distance	Cluster	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	98.6	91.91	96.94	87.37	99.73	98.75	98.94
Closest & Farthest	Canberra	2	97.28	83.57	92.81	76	99.41	97.64	96.58
Farthest	Canberra	2	98.85	93.33	98.89	88.37	99.90	98.84	98.83
Uniform	Canberra	2	97.8	86.86	95.01	79.99	99.58	98.03	97.89
Closest	Cosine	2	98.75	92.76	97.8	88.22	99.8	98.83	98.88
Closest & Farthest	Cosine	2	97.25	83.35	92.85	75.62	99.42	97.61	97.13
Farthest	Cosine	2	98.43	90.79	97.06	85.28	99.74	98.54	98.72
Uniform	Cosine	2	97.9	87.55	94.95	81.22	99.57	98.15	97.71
Closest	Euclidean	2	98.74	92.74	97.25	88.64	99.75	98.87	99.27
Closest & Farthest	Euclidean	2	98.52	91.39	97.09	86.32	99.74	98.65	98.28
Farthest	Euclidean	2	99.39	96.56	99.71	93.6	99.97	99.36	99.27
Uniform	Euclidean	2	98.21	89.4	96.86	83.02	99.73	98.33	98.54
Wang's	-	2	98.75	92.8	97.53	88.51	99.78	98.86	99.13
Wei's	-	2	97.36	83.82	94.64	75.24	99.58	97.57	98.31
Dissimilarity-based	-	-	91.24	32.18	54.34	22.87	98.08	92.71	89.96
Random Sampling	-	-	91.54	29.66	60.64	19.64	98.73	92.47	86.59
Closest	Canberra	4	98.54	91.49	97.02	86.55	99.73	98.67	98.76
Closest & Farthest	Canberra	4	97.19	83	92.38	75.36	99.38	97.58	96.44
Farthest	Canberra	4	98.85	93.36	98.68	88.58	99.88	98.87	98.79
Uniform	Canberra	4	97.75	86.48	95.44	79.07	99.62	97.94	97.56
Closest	Cosine	4	98.68	92.37	97.48	87.78	99.77	98.79	98.67
Closest & Farthest	Cosine	4	96.75	80.14	90.19	72.11	99.22	97.27	96.36
Farthest	Cosine	4	98.18	89.27	96.35	83.16	99.69	98.34	98.45
Uniform	Cosine	4	97.74	86.51	94.38	79.85	99.53	98.02	97.47
Closest	Euclidean	4	98.72	92.61	97.35	88.32	99.76	98.84	99.12
Closest & Farthest	Euclidean	4	98.19	89.39	95.42	84.08	99.6	98.43	97.95
Farthest	Euclidean	4	99.42	96.76	99.22	94.41	99.93	99.44	99.3
Uniform	Euclidean	4	98.05	88.41	96.32	81.7	99.69	98.2	98.35
Wang's	-	4	98.54	91.53	97.05	86.61	99.74	98.67	98.97
Wei's	-	4	98.06	88.45	96.35	81.76	99.69	98.20	98.81
Dissimilarity-based	-	-	91.24	32.18	54.34	22.87	98.08	92.71	89.96
Random Sampling	-	-	91.54	29.66	60.64	19.64	98.73	92.47	86.59
Closest	Canberra	8	98.48	91.13	96.98	85.95	99.73	98.61	98.57
Closest & Farthest	Canberra	8	97.1	82.46	91.57	75	99.31	97.54	96.84
Farthest	Canberra	8	98.83	93.23	97.99	88.9	99.82	98.9	98.88
Uniform	Canberra	8	97.65	85.9	94.54	78.71	99.54	97.9	97.43
Closest	Cosine	8	98.36	90.44	96.26	85.29	99.67	98.55	98.27
Closest & Farthest	Cosine	8	96.53	78.58	89.49	70.04	99.18	97.07	96.07
Farthest	Cosine	8	98.19	89.34	96.36	83.27	99.69	98.35	98.29
Uniform	Cosine	8	97.39	84.23	93.53	76.63	99.47	97.71	96.88
Closest	Euclidean	8	98.72	92.61	97.33	88.33	99.76	98.84	99.16
Closest & Farthest	Euclidean	8	98.17	89.3	95.36	83.97	99.59	98.41	97.89
Farthest	Euclidean	8	99.4	96.63	99.19	94.2	99.92	99.42	99.2
Uniform	Euclidean	8	98.12	88.83	96.77	82.1	99.73	98.24	98.42
Wang's	-	8	97.85	87.26	94.48	81.08	99.53	98.13	98.32
Wei's	-	8	99.1	94.85	98.85	91.17	99.89	99.12	99.73
Dissimilarity-based	-	-	91.24	32.18	54.34	22.87	98.08	92.71	89.96
Random Sampling	-	-	91.54	29.66	60.64	19.64	98.73	92.47	86.59

Table A.17: *Performance comparison of different negative sampling strategies on the Herpes dataset using the SVM model.*

Selection	Distance	Cluster	ACC	F1	PPV	Recall	SP	NPV	AUC
Closest	Canberra	2	99.01	94.33	97.83	91.07	99.8	99.11	99.2
Closest & Farthest	Canberra	2	98.28	89.99	95.89	84.77	99.64	98.49	98.76
Farthest	Canberra	2	99.04	94.49	98.63	90.69	99.87	99.08	98.91
Uniform	Canberra	2	98.31	90.14	96.24	84.76	99.67	98.49	98.41
Closest	Cosine	2	99.08	94.75	98.62	91.17	99.87	99.12	99.05
Closest & Farthest	Cosine	2	97.83	87.13	94.63	80.74	99.54	98.1	98.11
Farthest	Cosine	2	98.69	92.43	97.25	88.06	99.75	98.82	98.75
Uniform	Cosine	2	98.49	91.17	97.07	85.95	99.74	98.61	98.41
Closest	Euclidean	2	99.08	94.79	98.1	91.69	99.82	99.17	99.23
Closest & Farthest	Euclidean	2	98.96	94.04	97.98	90.4	99.81	99.05	99.14
Farthest	Euclidean	2	99.44	96.83	99.55	94.25	99.96	99.43	99.36
Uniform	Euclidean	2	98.5	91.3	96.84	86.36	99.72	98.65	98.26
Wang's	-	2	99.01	94.33	98.14	90.8	99.83	99.09	98.69
Wei's	-	2	97.99	88.22	94.37	82.82	99.51	98.3	98.83
Dissimilarity-based	-	-	93.57	56.01	73.97	45.08	98.42	94.72	93.64
Random Sampling	-	-	93.15	50.25	73.8	38.12	98.65	94.1	93.05
Closest	Canberra	4	99.03	94.48	97.9	91.29	99.8	99.13	99.25
Closest & Farthest	Canberra	4	98.38	90.57	95.97	85.74	99.64	98.59	98.88
Farthest	Canberra	4	99.01	94.29	98.74	90.22	99.88	99.03	98.73
Uniform	Canberra	4	98.27	89.92	95.67	84.82	99.62	98.5	98.43
Closest	Cosine	4	99	94.26	98.3	90.54	99.84	99.06	98.99
Closest & Farthest	Cosine	4	97.81	87	94.21	80.82	99.5	98.11	98.21
Farthest	Cosine	4	98.54	91.49	96.84	86.71	99.72	98.68	98.47
Uniform	Cosine	4	98.37	90.44	96.8	84.86	99.72	98.5	98.33
Closest	Euclidean	4	99.01	94.34	98.15	90.81	99.83	99.09	98.54
Closest & Farthest	Euclidean	4	98.8	93.09	97.44	89.12	99.77	98.92	98.97
Farthest	Euclidean	4	99.49	97.15	99.39	95.01	99.94	99.5	99.45
Uniform	Euclidean	4	98.35	90.36	96.67	84.82	99.71	98.5	98.08
Wang's	-	4	98.92	93.82	97.72	90.24	99.79	99.03	99.04
Wei's	-	4	98.58	91.81	96.24	87.78	99.66	98.79	98.87
Dissimilarity-based	-	-	93.57	56.01	73.97	45.08	98.42	94.72	93.64
Random Sampling	-	-	93.15	50.25	73.8	38.12	98.65	94.1	93.05
Closest	Canberra	8	98.96	94.09	97.64	90.81	99.78	99.09	99.18
Closest & Farthest	Canberra	8	98.21	89.57	95.38	84.43	99.59	98.46	98.58
Farthest	Canberra	8	98.97	94.06	98.57	89.95	99.87	99	98.46
Uniform	Canberra	8	98.25	89.79	95.55	84.69	99.61	98.49	98.58
Closest	Cosine	8	98.7	92.48	97.4	88.04	99.76	98.82	98.64
Closest & Farthest	Cosine	8	97.54	85.41	92.87	79.06	99.39	97.94	98.12
Farthest	Cosine	8	98.51	91.31	96.91	86.32	99.72	98.65	98.29
Uniform	Cosine	8	98.05	88.41	96.06	81.9	99.66	98.22	98.23
Closest	Euclidean	8	99.06	94.64	98.22	91.31	99.83	99.14	98.62
Closest & Farthest	Euclidean	8	98.85	93.39	97.36	89.74	99.76	98.98	98.97
Farthest	Euclidean	8	99.48	97.05	99.3	94.91	99.93	99.49	99.46
Uniform	Euclidean	8	98.46	91	96.9	85.78	99.73	98.59	98.24
Wang's	-	8	98.33	90.28	96.03	85.2	99.65	98.54	98.39
Wei's	-	8	99.22	95.56	98.93	92.41	99.9	99.25	99.79
Dissimilarity-based	-	-	93.57	56.01	73.97	45.08	98.42	94.72	93.64
Random Sampling	-	-	93.15	50.25	73.8	38.12	98.65	94.1	93.05