ThpPred: An ML based tool for predicting therapeutic proteins/peptides

by Srijanee Gupta

Under the Supervision of Dr. G.P.S. Raghava

Indraprastha Institute of Information Technology Delhi
Month, Year

ThpPred: An ML based tool for predicting therapeutic proteins/peptides

by Srijanee Gupta

Submitted in partial fulfillment of the requirements for the degree of Master of Technology

to

Indraprastha Institute of Information Technology Delhi Month, Year

Certificate

This is to certify that the thesis titled "ThpPred: An ML based tool for predicting therapeutic proteins/peptides" being submitted by Srijanee Gupta to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

July, 2023

G.P.S. Raghava
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

First and foremost, I want to express my gratitude to Dr. G. P. S. Raghava, Head of the Computational Biology Department at the Indraprastha Institute of Information Technology (IIIT) in New Delhi, for his support, encouragement, inspiration, and helpful suggestions. I want to thank him from the bottom of my heart for allowing me to work with him to finish my thesis, which was named "ThpPred: An ML-based tool for predicting therapeutic Proteins/peptides".

I also want to extend my deep appreciation to Ms. Shipra Jain for all of her help with the project, including advice, encouragement, insightful remarks, and vast expertise.

I also want to express my gratitude to the non-teaching staff members for their wonderful assistance in whatever manner they could.

Last but not the least, this work would have never been a success without the constant inspiration of my parents, family members and my friends.

Srijanee Gupta

Roll No.: MT21231

Abstracts

ThpPred is a web-based tool, developed for predicting druggable proteins/peptides. The main dataset used in this study contained 356 therapeutic proteins/peptides and 356 random proteins/peptides, curated from DrugBank, Uniprot and other sources. In order to provide a fair assessment, we did internal validation on 80% of the data and external validation on the remaining 20%. In this study, we have implemented the following methods for predicting druggability of proteins/peptides; i) machine learning models on features chosen using SVC-L1, Variance Threshold, and correlation coefficient; ii) machine learning models on single feature (AAC, DPC & TPC); and iii) MERCI-based motif search. The goal was to construct the best model and install it on a web server by training it on protein sequences of already existing medications. When compared to other models, the XGB-based model performed the best on AAC features and obtained maximum AUCs of 0.91 and 0.91 on the training and validation datasets, respectively for the alternate dataset consisting of 356 positive sequences and 3560 negative sequences. On the other hand, the RF-based model performed admirably on DPC features and obtained maximum AUCs of 0.91 and 0.89 on the training and validation datasets for the main dataset. The AUC score and accuracy for both datasets improved when motif labels were added to ML predicted labels. ThpPred was created to determine if a protein is therapeutic or not by combining motif search with RF and XGB models. The platform helps the scientific community create more effective protein-based medicines by providing a free web server and a standalone package. Overall, the results of the study indicate that ThpPred has the potential to improve the development of pharmaceuticals and protein-based treatments for the treatment of numerous diseases.

List of Figures

List of Tables

Contents

Chapter 1: INTRODUCTION

1.1 Background

The body's tissues contain virtually all of the twenty natural amino acids that make up proteins/peptides, which are essential biological components. The fact that proteins/peptides perform the majority of biological processes within organisms explains why they are indispensable. Since Jöns Jakob Berzelius first discovered them in 1838, the scientific world has been investigating their medicinal potential. The development of serum treatment in the 1880s marked an important turning point for protein therapies. The first Nobel Prize in Medicine was awarded to researcher von Behring in 1901 in recognition of this outstanding accomplishment. With the development of the hybridoma technology in the 1970s, the area of protein therapies saw a significant uptick. This innovative method made it possible to create an endless supply of unique monoclonal antibodies [1]. Fully human antibodies were created thanks to later technological advances. Apart from antibodies, the first protein treatment was insulin, which was made from animal pancreas. Since then, incremental advancements in effectiveness, safety, quality, and cost, as well as the discovery of novel targets, have been the main concerns in the development of therapeutic proteins/peptides[1]. Recombinant antibodies, hormones, cytokines, interferons, and enzymes of human origin are now widely accessible on the market as a result of this advancement[2]. A research projects that by 2026, protein/peptide therapies would earn an estimated 50.6 billion USD in worldwide revenue [3]. The tremendous developments and possibilities of the field are reflected in this prediction. The creation of new therapeutic proteins/peptides holds enormous promise for solving unmet medical needs and enhancing patient outcomes as research advances. Scientists and medical experts hope to revolutionise the treatment of several illnesses and ailments by harnessing the power of proteins/peptides, ushering in a new age of personalised and focused medicine [4].

Proteins/peptides and peptides in the body are essential for carrying out several processes and activities required for sustaining cellular systems [1]. They have notably been linked to the onset of a number of disorders, including diabetes, cancer, and neurological conditions [2]. As a result, using proteins/peptides and peptides as therapeutic agents has become a viable way to treat a variety of diseases. Due to their remarkable tissue penetration, strong biological activity, and affordability, they have recently shown their potential to revolutionise medical treatment [3]. The U.S. Food and

Drug Administration (FDA) has authorised the clinical use of various proteins/peptides and peptides as medical therapies due to its recognition of their importance [4, 5]. Understanding the numerous therapeutic protein types, which may be further categorised based on their therapeutic function and common properties, is essential to understanding the link between proteins/peptides and their therapeutic potential [6].

Virtual screening is a method frequently used to speed up the discovery of therapeutic proteins/peptides. It helps prediction models based on machine learning. The type of data utilised and how it is represented determines how well these models are able to predict outcomes. Although the amino acid sequence is useful, it may be missing important aspects including physicochemical, conformational, topological, and geometrical characteristics. The protein's physicochemical and compositional representations provide important insights even if they are unable to distinguish between different sequence permutations [7].

The ability of machine learning-based predictors to analyse enormous volumes of data has been demonstrated, giving them a quick and affordable technique for discovering therapeutic proteins/peptides [8]. Due to their proficiency in vast data processing, machine learning-based predictors have demonstrated their ability in identifying therapeutic proteins/peptides as a cheap and quick method[5].

1.2 Motivation of Work

Antibiotic resistance in bacteria is currently thought to be one of the most crucial issues in the treatment of health and illness. It is connected all over the world and poses a threat to the population's health. Unfortunately, it does not get criticism from all facets of society. As a result, the identification and management of microbe illnesses are becoming more challenging and dangerous due to microbe medication resistance. Specific illness manifestations are connected to microbe growth and multiplication within the host, and their survival tactics are linked to antibiotic resistance. We know very little about the nature of drug resistance despite thorough research of clinical and experimental drug resistance. As a result, it is a problem that has to be addressed on a worldwide scale with adaptable drug resistance efforts and organisations that share scientific knowledge. Microbe infections are often quite patchy, which has a big impact on the selection of resistance. The majority of microorganisms' medication resistance-causing genes are still mostly unknown. Even though certain genes linked to medication resistance have been found, it might be

challenging to comprehend the intricate interactions between vectors and microorganisms as well as their complicated life cycles. Thus, the course of drug resistance evolution is difficult to predict[6].

However, due to the lack of appropriate vaccines, drugs remain the pillar for defense against microbes that are responsible for a wide range of human diseases. Drug-resistance mechanisms in microbes are quite inconsistent with respect to the form and life cycle of microbes[7]. The majority of medication resistance mechanisms in pathogenic bacteria include increased drug efflux, decreased drug absorption, mutations in targeted enzymes, metabolic upregulation, target site deficiency for antiparasitic medicines, etc. Numerous genetic mechanisms, including as gene deletion, gene mutations, and, most significantly, gene chimerization, also play a significant part in the development of medication resistance in connection to these issues[6]. That is why, rapid development of drugs is the need of the hour.

Although several computational methods to predict various therapeutic protein types have been presented for the purpose of increasing the selection of therapies and drugs, the majority of these models have concentrated on proteins/peptides with therapeutic qualities rather than proteins/peptides that have just been employed to make drugs molecules. It is highly probable that a protein molecule that has therapeutic properties may not be eligible to become a drug molecule. Thus, developing a model with those proteins/peptides which have already been used to make drugs can give better results in selecting therapeutic proteins/peptides for future drug discoveries.

2.1 Objective

The major goal of this research was to create a machine learning (ML) model that would predict therapeutic proteins/peptides that might be employed in the development of new therapeutics by using protein sequences from currently approved medications as training data.

2.1.1 Overall Objective

The final goal was to train the models on main and alternate datasets, choose the best model and deploy it on the webserver.

2.2 Scope

Due to their distinctive biochemical properties and therapeutic potential, proteins/peptides have recently emerged as a separate class of therapeutic agents. Although proteins/peptides perform better than small compounds and big biologics in certain ways, they frequently have poor in vivo

stability and membrane impermeability because of the inherent constraints of amino acids. In order to address these issues, much research has been done on the development, manufacture, and optimisation of protein medicines. The rapid generation of efficient and selective lead proteins/peptides is made possible by the combination of conventional lead protein discovery techniques with cutting-edge technologies like rational design and phage display[8].

In order to reduce the cost of development of drugs, prediction models are highly imperative. Models that not only predict therapeutic proteins/peptides but whose predictions are based on protein sequences already used for drug development are need of the hour. We believe that this model will be highly useful in this domain, leading to more research and development of better models in this area.

Chapter 2: LITERATURE REVIEW

3.1 Information

In recent years, interest in proteins/peptides as medicines has grown. The safe and powerful mode of action of proteins/peptides is a significant factor in this achievement. Future protein drug research is projected to build on the benefits of proteins/peptides occurring naturally while addressing their drawbacks, such as their chemical, physical properties, through traditional rational design. The therapeutic uses of proteins/peptides are also anticipated to increase with the development of novel protein technologies, including, protein drug conjugates, multifunctional proteins/peptides and cell-penetrating proteins/peptides [9].

Numerous proteins/peptides have been thought of as possible therapeutic targets. These targets are of relevance to many fields of biological and pharmaceutical research, including the creation and assessment of bioinformatics, molecular modelling, computer-aided drug design, and analytical techniques. The relevant groups would consequently benefit from a publicly available databases that offer thorough information about these targets [10]. Prediction of therapeutic peptides is crucial for medication development and treatment. This crucial duty has been investigated by researchers, who have created a number of computer techniques to recognize various medicinal peptide kinds[11].

In comparison to the conventional drug development method, the manufacture of peptide-based biologic medicines is less expensive and difficult. In order to create novel and potent therapeutic medications, it is crucial to identify peptides with restorative properties, which expedites their usage in clinical therapy. As a result, there is enormous potential for the development of universal, original, and extendable experimentally methods for the accurate prediction of therapeutic peptides[12]. To create prediction models, a variety of peptide sequence amino acid compositions (AACs) have been used. These include Chou's pseudo amino acid composition (PseAAC) [13], AAC combinations, average chemical shifts (acACS), and reduced AAC (RAAC) [14], pseudo g-Gap DPC, amphiphilic PseAAC, and reduced amino acid alphabet (RAAAC) [15].

Machine learning methods were also used to increase model efficacy. Several models have combined the quantitative outcomes of individual classifiers (RF, K-nearest neighbour, SVM, generalised neural network, and probabilistic neural network) as well as support vector machine (SVM) and random forest (RF) machine learning techniques. Additionally, a collection of SVM-based models trained using sequence-based features has also been used[16].

Several models have been developed to predict therapeutic nature of proteins/peptides and peptides. The stacking framework, together with the KNN, SVM, ET, RF, and XGB, are used to construct TP-MV. To extract the discriminative feature for various peptides, TP-MV then builds a multi-view learning model as meta-classifiers[11]. A prediction approach based on Random Forest was developed called Physicochemical Property-based Therapeutic Peptide Predictor (PPTPP), to solve this problem. For the purpose of producing and ranking features linked to physicochemical properties, a unique feature encoding and learning strategy was started. The proposed technique is able to detect the informative IPP of peptides in addition to being able to predict several therapeutic peptides with excellent comparability to known predictors[5].

In another research, the bioinformatics tool PEPred-Suite was created for the general prediction of medicinal peptides. For learning the optimum representative features for different peptide types in PEPred-Suite, they provide an adaptive feature representation approach. They train a variety of sequence-based feature descriptors, incorporate the learned class information into the features, and use a two-step feature optimisation technique based on the area under the receiver operating characteristic curve to be more precise, to extract the most discriminative features. Using the discovered representative properties, they trained eight random forest models for eight distinct types of functional peptides. Benchmarking findings shown that PEPred-Suite outperforms other predictors for a variety of peptides with better and more reliable results[17]. A creative stacking system PreTP-Stack has also been suggested as a method of medicinal peptide prediction. Ten distinct features and four predictors—Random Forest, Linear Discriminant Analysis, XGBoost, and Support Vector Machine—make up the PreTP-Stack[18]. A predictor known as PreTP-EL has been suggested in a different study by using the ensemble learning method to combine the various attributes and machine learning techniques in order to capture the varied properties of distinct therapeutic peptides[19]. However, none of these models use sequences of proteins/peptides that have already been used for the purposes of drug development.

3.2 Summary

In recent years, interest in proteins/peptides as medicines has grown. The safe and powerful mode of action of proteins/peptides is a significant factor in this achievement. Future protein drug research is projected to build on the benefits of proteins/peptides occurring naturally while

addressing their drawbacks, such as their chemical, physical properties, through traditional rational design. The therapeutic uses of proteins/peptides are also anticipated to increase with the development of novel protein technologies, including, protein drug conjugates, multifunctional proteins/peptides and cell-penetrating proteins/peptides. Numerous proteins/peptides have been thought of as possible therapeutic targets. Prediction of therapeutic peptides is crucial for medication development and treatment. This crucial duty has been investigated by researchers, who have created a number of computer techniques to recognize various medicinal peptide kinds. In comparison to the conventional drug development method, the manufacture of peptide-based biologic medicines is less expensive and difficult. In order to create novel and potent therapeutic medications, it is crucial to identify peptides with restorative properties, which expedites their usage in clinical therapy. As a result, there is enormous potential for the development of universal, original, and extendable experimentally methods for the accurate prediction of therapeutic peptides. To create prediction models, a variety of peptide sequence amino acid compositions have been used. Several models have been developed to predict therapeutic nature of proteins/peptides and peptides. To extract the discriminative feature for various peptides, TP-MV then builds a multi-view learning model as meta-classifiers. A prediction approach based on Random Forest was developed called Physicochemical Property-based Therapeutic Peptide Predictor, to solve this problem. The proposed technique is able to detect the informative IPP of peptides in addition to being able to predict several therapeutic peptides with excellent comparability to known predictors. In another research, the bioinformatics tool PEPred-Suite was created for the general prediction of medicinal peptides. They provide an adaptive feature representation technique for learning the best representative features for various peptide types in PEPred-Suite. Benchmarking findings shown that PEPred-Suite outperforms other predictors for a variety of peptides with better and more reliable results. A creative stacking system PreTP-Stack has also been suggested as a method of medicinal peptide prediction. Since, none of these models have used protein sequences of already established drug molecules, it makes our study different as for the positive dataset (Therapeutic proteins/peptides), we have only considered sequences proteins/peptides that have been used to make drug molecules from DrugBank [20].

Chapter 3: METHODOLOGY

3.1 Creation and compilation of datasets

A total of 831 drugs were extracted and compiled into ThpDB2. These drugs were considered for out positive data. Drugs which had sequence information were filtered from the 831 drugs list with a total of 387 drugs. Of these 387 proteins/peptides, sequences of length 30-1500 were filtered, leaving the final total sequences to be 356 which were referred to as a positive dataset.

For our negative dataset, we downloaded random protein sequences from Uniprot using keywords such as non + anti-antiogenic, anti-bacterial, anti-cancer, anti-inflammatory, anti-viral, cell-penetrating, polystyrene surface binding, and quorum sensing. After removing duplicate entries, we were left with 83655 sequences. Of these 83655 proteins/peptides, sequences of length 30-1500 were filtered, leaving the total sequences to be 83313. After that, CD-HIT software[21] was applied to this dataset at 40% sequence identity. 83210 sequences remained. Out of 83210, 356 sequences were selected randomly. Thus, our main dataset comprised of 356 negative sequences along with the 356 positive sequences (Figure 1).

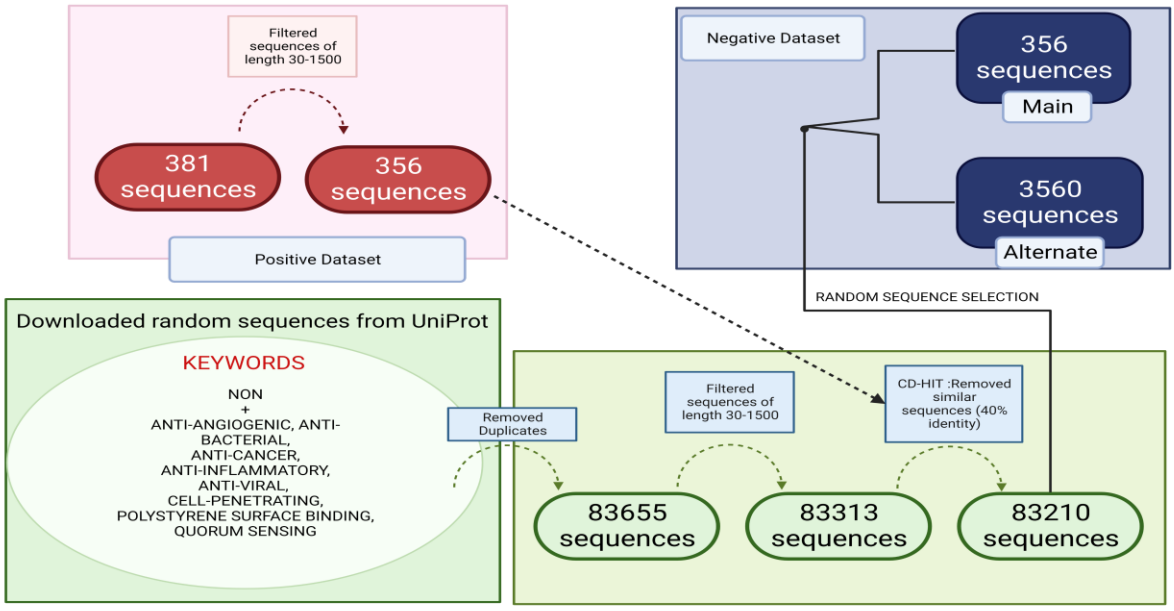

Figure 1: Workflow Showing Generation of Positive & negative Datasets

For our alternate dataset, we selected 3560 sequences randomly from the 83210 and compiled them with the 356 positive sequences. According to the quantity of therapeutic and non-therapeutic protein sequences, we produced two datasets, which are detailed below.

(a) The main dataset consists of 356 therapeutic protein sequences (acquired after pre-processing positive data) and 356 non-therapeutic protein sequences (randomly picked from 83210 negative data obtained after CD-HIT).

(b) Alternative Dataset (Ten Times Negative Dataset): This dataset includes 3560 non-therapeutic protein sequences randomly chosen from 83210 sequences acquired after CD-HIT, together with 356 therapeutic protein sequences (obtained after pre-processing of positive data).

3.2 Feature generation

To create any prediction model, a collection of pertinent attributes must be extracted for each protein or peptide sequence. For this, a standalone tool called Pfeature was utilised. It can produce a variety of composition features, including features based on composition and evolutionary information[22]. Using a composition-based feature module of Pfeature [22], a vector of 9189 features was calculated against each sequence for all three datasets.

3.3 Feature selection

Previous research has demonstrated that not all traits are crucial. Selecting the relevant characteristics from a bigger number of features is therefore quite difficult. To choose the important features from the high-dimensional feature set for this investigation, we employed various feature selection methods:

1. *The SVC-L1-based feature selection approach (Scikit-learn package).*

The support vector classifier (SVC) with linear kernel and L1 regularization is the foundation of this approach([23].

From the pool of 9189 features, we have selected the most significant characteristics for each of the two datasets using this procedure. 94 for the alternate dataset, and 62 for the main dataset were

chosen from this group. 48 features were common between the main and alternate dataset (Figure 2).

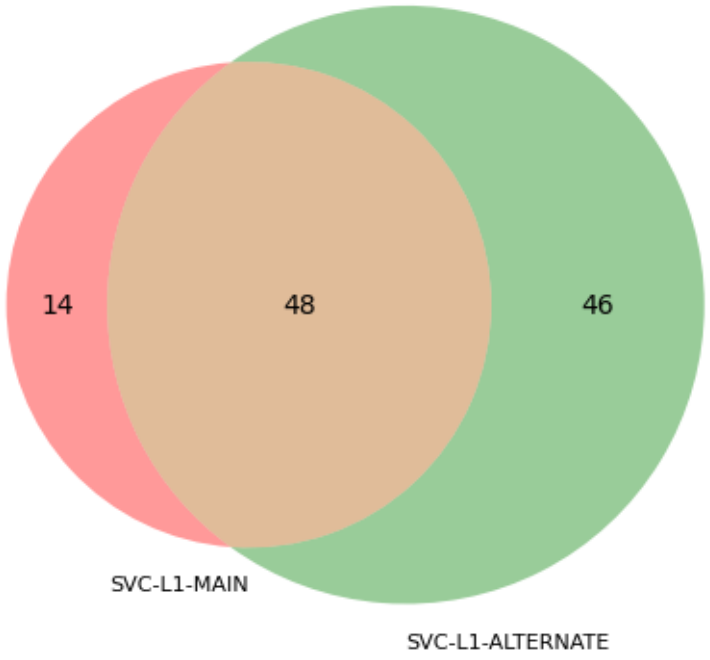

Figure 2: Venn Diagram showing common SVC-L1 features between main and alternate dataset

2. *Correlation Coefficient*

As correlated features might cause the model to become redundant and perform worse, removing them is a crucial stage in the feature selection process for machine learning. When two or more characteristics have a high degree of correlation, they provide the model comparable information, which can lead to overfitting or instability[24].

From the pool of 9189 features, we have selected the most significant characteristics for each of the two datasets using this procedure. 47 for the alternate dataset, and 283 for the main dataset were chosen from this group. 42 features were common between the main and alternate dataset (Figure 3).

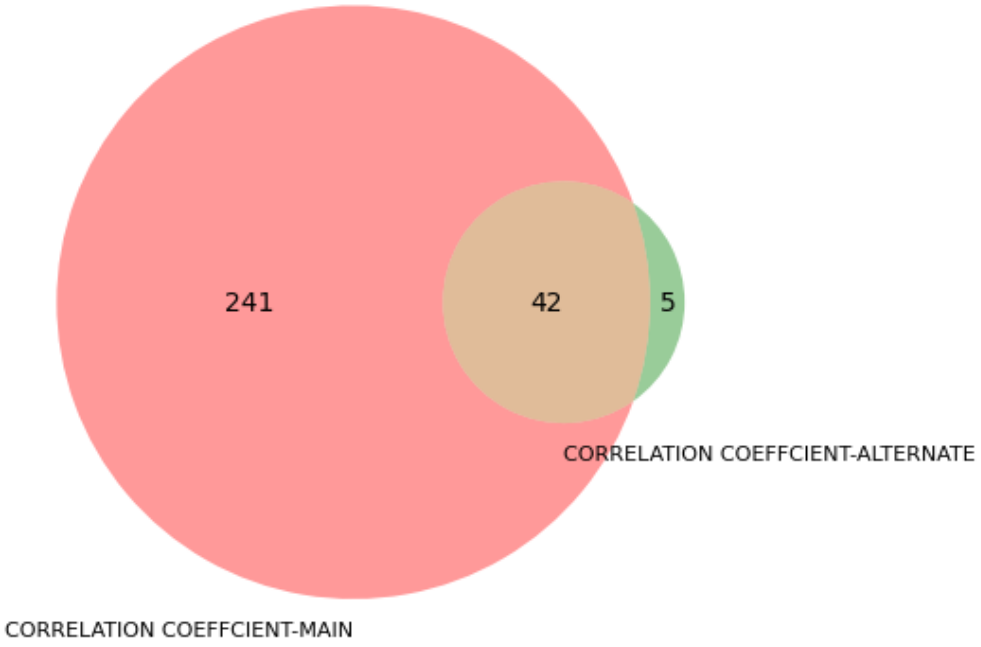

Figure 3: Venn Diagram showing common features after removing correlated features between main and alternate dataset

3. *Variance Threshold*

A feature's variance is a measurement of how far it deviates from its mean value. A feature with low variance is one whose values are closely grouped around the mean, indicating that the characteristic does not fluctuate significantly throughout the dataset. In the variance threshold procedure, the variance of the features is given a threshold value, and any features with variance below that value are removed[25].

From the pool of 9189 features, we have selected the most significant characteristics for each of the two datasets using this procedure. 76 for the alternate dataset, and 102 for the main dataset were chosen from this group (Table 1). All the features of the alternate dataset were covered in the main dataset features (Figure 4).

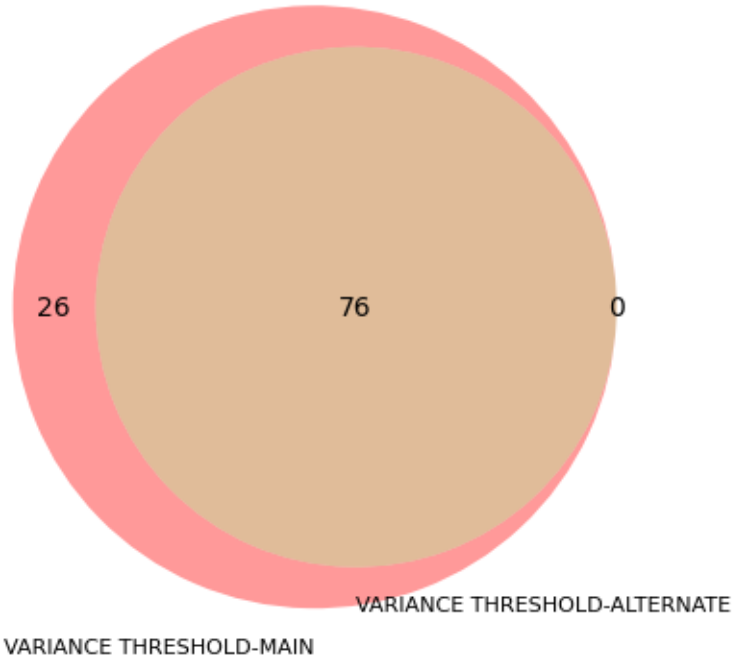

Figure 4: Venn Diagram showing common variance threshold features between main and alternate dataset

4. PCA

Finding the correlation between each feature and the target variable, then choosing the features with the greatest absolute correlation coefficients, is one method of employing correlation coefficient for feature selection. The traits that are most closely connected to the target variable may be found using this method, making them more likely to be relevant for generating precise predictions. The correlation coefficient, which may not be able to capture more intricate or non-linear interactions, solely assesses linear correlations between variables. In order to find the most useful features for a specific machine learning issue, various feature selection techniques, such as mutual information, principal component analysis (PCA), or decision trees, may also be employed in conjunction with correlation coefficient[26].

In order to keep the 10 principal components that account for the majority of variance in the dataset, we set n_components to 10.

5. *Selecting Single feature*

From the pool of 9189 features, we also extract single features such as all amino acid composition features, all dipeptide composition features and all tripeptide composition features to apply different ML models on each of them separately (Figure 5,6).

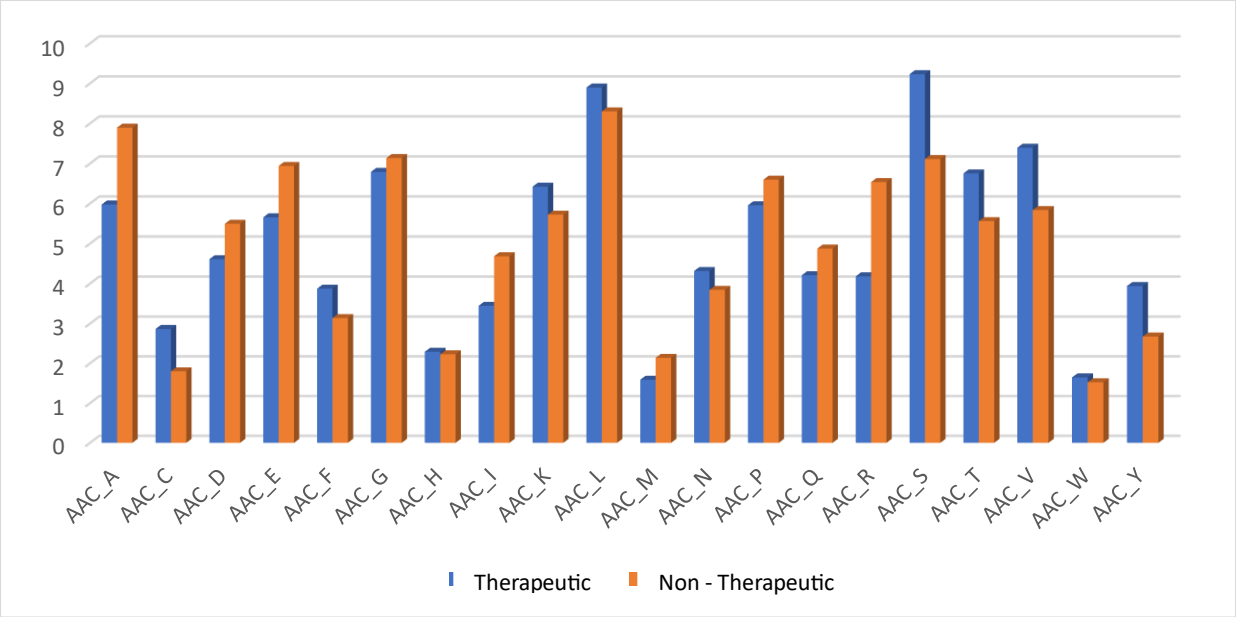

Figure 5A: Amino acid composition in the protein sequences of the main dataset

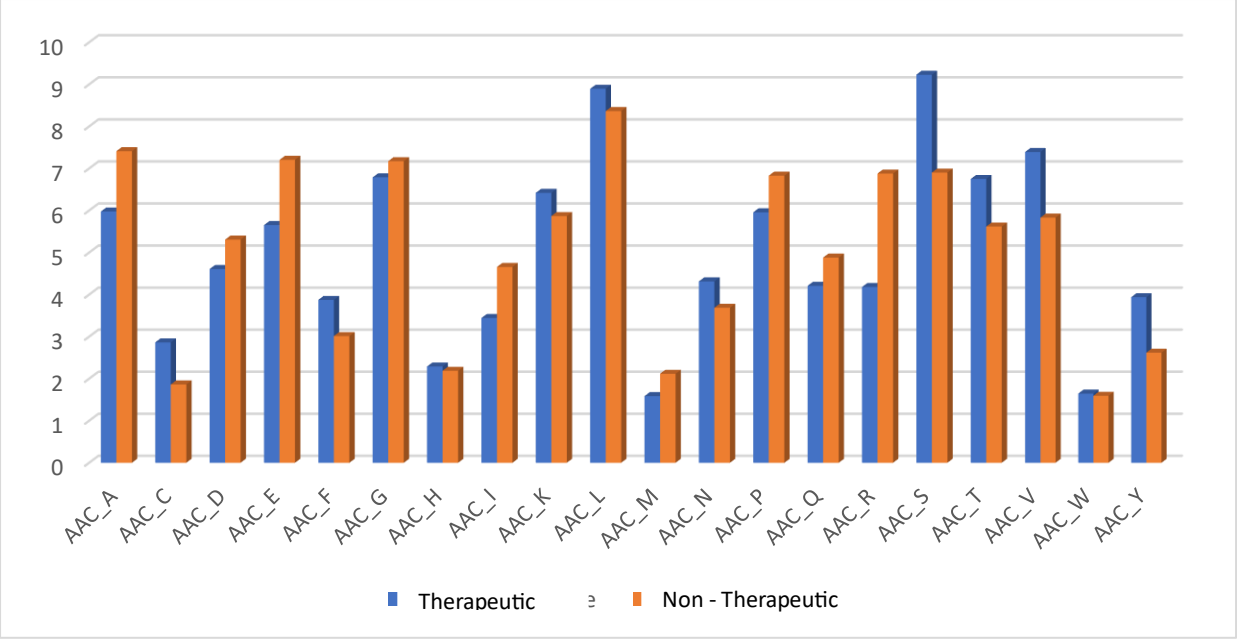

Figure 5B: Amino acid composition in the protein sequences of the alternate dataset

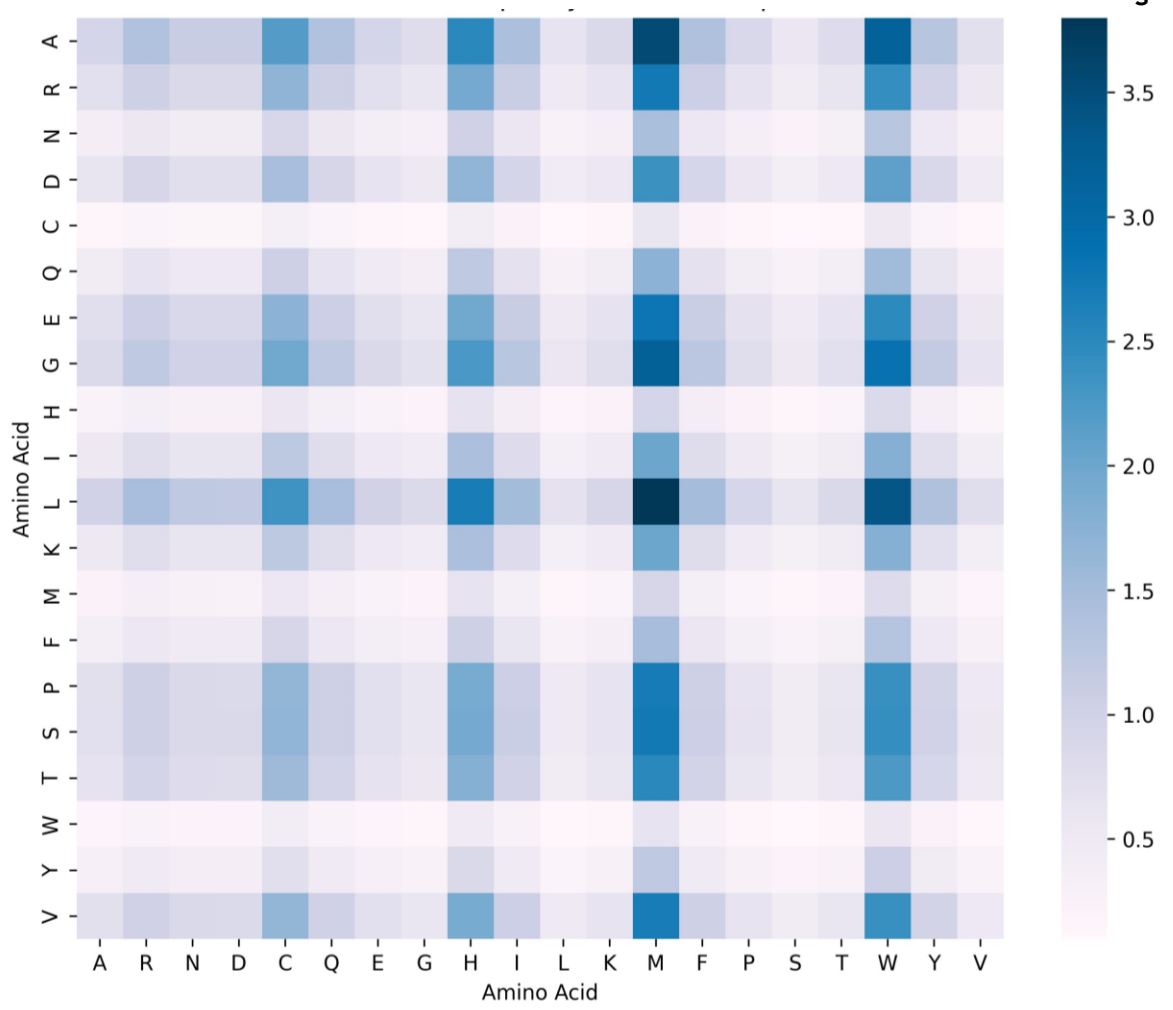

Figure 6: Heatmap depicting frequency ratio of amino acids between sequences of main dataset and alternate dataset

3.4 Motif Analysis

The Programme, Motif-EmeRging and the Classes- Identification (MERCI) tool, which can find motifs in any set of sequences, were used to look for motifs in the hazardous proteins/peptides [31]. The investigation of patterns found in the hazardous sequences is known as motif analysis. This Programme locates patterns in the files using Perl script [27]. Motifs in the main dataset and alternate dataset were listed using this Programme. Motifs were filtered and 45 motifs were selected for the main dataset out of which 25 occurred only in positive sequences and 20 occurred only in negative sequences. Here, listed are the top 15 motifs on the basis of their frequencies in sequences (Table 1).

S.No.	Positive Motifs	Frequency	Negative Motif	Frequency
1	sulfurbasicgapSneutral	0.379	PVchargedP	0.379
2	YaromaticC	0.374	PVchargedPsurface	0.371
3	neutralKNneutral	0.374	PVchargedPpolar	0.340
4	WgapQG	0.371	PVacidicP	0.303
5	Qaliphaticaromaticneutralaliphatic	0.371	PVchargedPneutral	0.298
6	basicVneutralN	0.371	TPaliphaticPaliphatic	0.295
7	CbasicgapSneutral	0.371	aliphaticPaliphaticbasicP	0.292
8	aromaticgapCchargedmedium	0.368	LacidicLaromaticaliphatic	0.281
9	VSW	0.362	LacidicLaromatic	0.281
10	PSVF	0.362	ILDL	0.258
11	CSV	0.354	GFPV	0.258
12	NWY	0.348	LDLW	0.256
13	KGQP	0.340	FPDW	0.256
14	neutralVpolarWsurface	0.323	PMgapT	0.039
15	SgapWneutralneutral	0.264	PVchargedgapP	0.034

Table 1: List of positive and negative motifs from sequences of main datasets and their frequency

Similarly, 72 motifs were selected for alternate dataset out of which 27 occurred only in positive sequences and 45 occurred only in negative sequences. Here, listed are the top 15 motifs on the basis of their frequencies in sequences (Table 2).

S.No.	Positive Motifs	Frequency	Negative Motif	Frequency
1	aliphaticKbasicLN	0.014	PVchargedP	0.390
2	FKRaliphaticN	0.014	PVchargedPsurface	0.387
3	aromaticKRaliphaticN	0.014	PVchargedPpolar	0.363
4	largehydrophobicWIS	0.014	PVacidicP	0.309
5	FKRhydrophobicN	0.014	aliphaticPaliphaticbasicP	0.304
6	FKRlargeN	0.014	TPaliphaticPaliphatic	0.300
7	aromaticKRhydrophobicN	0.014	aliphaticPaliphaticRP	0.300
8	aliphaticMHNL	0.011	TPaliphaticPG	0.296
9	aliphaticMHneutralL	0.011	PneutralVP	0.288
10	aliphaticsulfurHNL	0.011	neutralaliphaticPaliphaticR	0.284
11	MHNL	0.011	PVbasicP	0.282
12	neutralIWValiphatic	0.011	LacidicLaromatic	0.281
13	DTPEE	0.008	VPaliphaticR	0.279
14	DVHNF	0.008	VPgapRP	0.272

15	DYYM	0.008	QVPgapR	0.271

Table 2: List of positive and negative motifs from sequences of alternate datasets

3.5 Machine learning based classifiers

Differentiating between therapeutic and non-therapeutic proteins/peptides has been accomplished using a variety of machine learning approaches. The classification models were created using Random Forest (RF)[28], Decision Tree (DT)[29], k-nearest neighbors (KNNs) [30], XGBoost (XGB) [31], and Support Vector Classifier (SVC) [32] The best outcomes from these classifiers' optimisation utilizing multiple hyperparameters were included. Figure 7 shows the ThpPred workflow in its entirety.

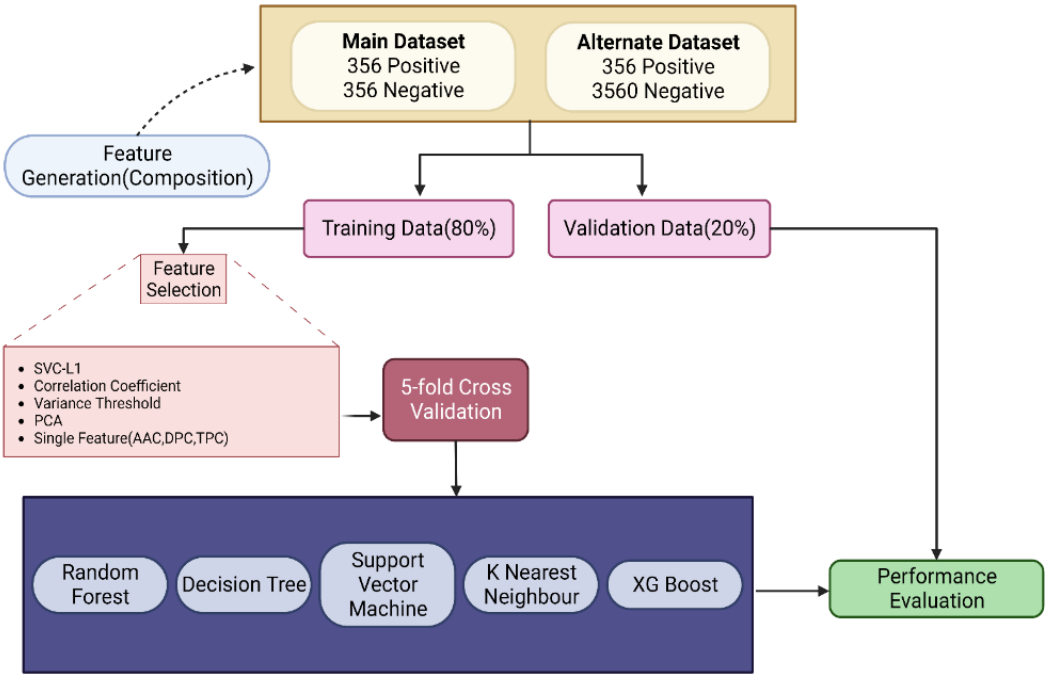

Figure 7: Complete Workflow of ThpPred

3.6 Cross-validation and performance metrics

The two datasets in this study were divided in an 80:20 ratio, with 80% serving as training datasets and 20% as validation datasets. In this study, 80% of the training data was utilised to evaluate the machine learning models using five-fold cross-validation (CV)[33, 34]. The training data is divided

into five folds for internal validation, with four folds utilized for training and the last fold for testing. Five rounds of the same process are iterated in order to test each of the five folds at least once. The average of the common evaluation metrics, comprising threshold dependent and independent parameters, was used to evaluate the performance of multiple machine learning models. The area under the receiver operating characteristic curve (AUC) is a parameter that is independent of threshold, whereas sensitivity, specificity, accuracy, and Matthews Correlation Coefficient (MCC) are metrics that depend on threshold. The earlier research [35, 36] include thorough annotations for those metrics.

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \qquad (1)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \qquad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \times 100 \qquad (3)$$

$$MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (4)$$

where FP, FN, TP, and TN are false positive, false negative, true positive, and true negative, respectively.

3.7 Selection of Best Model & Improving the Score

After applying different machine learning techniques to discriminate therapeutic from non-therapeutic proteins/peptides, the best models were selected (one for main dataset and one for alternate dataset). The predicted labels were corrected based on the occurrence of positive and negative motifs in the sequences. For instance, if a sequence had a positive motif, it was labeled positive, if it contained negative motif, it was labelled negative and if no motif was found, the predicted label was kept as it is.

3.8 Development of Motif Scan Model

A small program was created which took sequences in fasta format as input, and returned list of therapeutic motifs found in those sequences if any.

3.9 Development of Design Module

The "Design Module" was created to create every protein mutation or alternative. One residue at a time is changed to create the new protein. In order to find the therapeutic proteins or peptides from the altered sequences, it can also offer a prediction score. The prediction score may also be sorted by the user to reveal the best/highest scoring mutant protein. Additionally, the user can produce a number of mutant proteins and peptides from the provided sequence. The user may create the best mutant with the highest score with this tool.

3.10 Webserver Development

The webserver, ThpPred was developed to use motif search in conjunction with RF and XGB models to assess if a protein is therapeutic or not. Four models have been incorporated in the prediction module (AAC based XGB, AAC based XGB+Motif, DPC based RF, AND DPC based RF+Motif). The other modules in the webserver are the motif scan module and the design module. The motif scan module, predicts therapeutic and non-therapeutic proteins/peptides solely on the basis of presence or absence of therapeutic motifs. The design module, on the other hand performs single point mutation in the non-therapeutic query sequence and gives those mutated sequences which are now therapeutic.

Chapter 4: RESULTS

Selected features

A total of 9189 features were evaluated for both datasets, as was described above in the "Feature Generation" and "Feature Selection" sections. These characteristics were then condensed using the SVC-L1 approach to 62 (main dataset) and 94 (alternative dataset), the variance threshold method to 102 (main dataset) and 76 (alternate dataset), and the correlation-coefficient method to 283 (main dataset) and 47 (alternate dataset). Different categorization models were developed using these condensed characteristics.

With the SVC-L1 features, the RF model was able to obtain a maximum AUC of 0.89 on the training dataset and 0.91 on the validation dataset, while the XGB model was able to get a maximum AUC of 0.94 on the training dataset and 0.93 on the validation dataset in the alternative dataset (Table 3).

MAIN DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.91	0.88	0.12	0.90	0.89	0.89	0.89	0.79	0.78
	VALIDATION	0.95	0.87	0.13	0.90	0.86	0.91	0.91	0.82	0.82
	HYBRID	0.95	0.90	0.10	0.92	0.88	0.92	0.93	0.85	0.85
DT	TRAIN	0.86	0.80	0.20	0.84	0.82	0.83	0.83	0.66	0.66
	VALIDATION	0.91	0.85	0.15	0.87	0.83	0.87	0.88	0.75	0.75
	HYBRID	0.94	0.86	0.14	0.89	0.85	0.90	0.90	0.79	0.79
KNN	TRAIN	0.76	0.85	0.15	0.80	0.85	0.80	0.81	0.61	0.61
	VALIDATION	0.80	0.82	0.18	0.79	0.78	0.81	0.81	0.62	0.62
	HYBRID	0.89	0.84	0.16	0.85	0.81	0.86	0.86	0.72	0.72
SVM	TRAIN	0.87	0.83	0.17	0.86	0.84	0.85	0.85	0.70	0.70
	VALIDATION	0.83	0.82	0.18	0.81	0.79	0.83	0.83	0.65	0.65
	HYBRID	0.86	0.86	0.14	0.85	0.83	0.86	0.86	0.72	0.72
XGB	TRAIN	0.91	0.88	0.12	0.90	0.89	0.89	0.89	0.79	0.78
	VALIDATION	0.94	0.86	0.14	0.89	0.85	0.90	0.90	0.79	0.79
	HYBRID	0.95	0.89	0.11	0.91	0.87	0.92	0.92	0.83	0.83
ALTERNATE DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.90	0.92	0.08	0.67	0.53	0.92	0.91	0.65	0.62

	VALIDATION	0.87	0.89	0.11	0.59	0.45	0.89	0.88	0.58	0.53
	HYBRID	0.90	0.90	0.10	0.62	0.48	0.90	0.90	0.61	0.57
DT	TRAIN	0.69	0.96	0.04	0.68	0.67	0.94	0.83	0.65	0.65
	VALIDATION	0.66	0.95	0.05	0.62	0.59	0.93	0.81	0.58	0.58
	HYBRID	0.77	0.95	0.05	0.68	0.61	0.93	0.86	0.65	0.65
KNN	TRAIN	0.58	0.96	0.04	0.59	0.61	0.93	0.77	0.55	0.55
	VALIDATION	0.62	0.95	0.05	0.59	0.56	0.92	0.79	0.55	0.55
	HYBRID	0.68	0.95	0.05	0.61	0.56	0.92	0.81	0.57	0.57
SVM	TRAIN	0.88	0.91	0.09	0.65	0.51	0.91	0.90	0.63	0.60
	VALIDATION	0.80	0.89	0.11	0.55	0.41	0.88	0.84	0.52	0.48
	HYBRID	0.87	0.89	0.11	0.59	0.45	0.89	0.88	0.58	0.54
XGB	TRAIN	0.92	0.94	0.06	0.74	0.62	0.94	0.93	0.73	0.71
	VALIDATION	0.90	0.93	0.07	0.70	0.57	0.93	0.92	0.68	0.66
	HYBRID	0.93	0.94	0.06	0.73	0.59	0.94	0.93	0.71	0.69

Table 3: Performance of ML models developed with features selected using SVC-L1

In the main dataset with the XGB model, a maximum AUC of 0.87 on training and 0.88 on the validation dataset was attained, whereas in the alternate dataset with the XGB model, a maximum AUC of 0.89 on training and 0.88 on the validation dataset was attained with variance threshold features (Table 4).

MAIN DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.87	0.86	0.14	0.87	0.87	0.87	0.87	0.74	0.73
	VALIDATION	0.91	0.89	0.11	0.89	0.87	0.90	0.90	0.79	0.79
	HYBRID	0.92	0.90	0.10	0.90	0.88	0.91	0.91	0.82	0.82
DT	TRAIN	0.85	0.81	0.19	0.84	0.83	0.83	0.83	0.66	0.66
	VALIDATION	0.77	0.81	0.19	0.77	0.77	0.79	0.79	0.58	0.58
	HYBRID	0.88	0.84	0.16	0.84	0.81	0.85	0.86	0.71	0.71
KNN	TRAIN	0.76	0.85	0.15	0.80	0.84	0.80	0.81	0.61	0.61
	VALIDATION	0.83	0.84	0.16	0.82	0.80	0.83	0.83	0.66	0.66
	HYBRID	0.91	0.85	0.15	0.87	0.83	0.87	0.88	0.75	0.75
SVM	TRAIN	0.84	0.85	0.15	0.85	0.85	0.84	0.84	0.68	0.68
	VALIDATION	0.75	0.86	0.14	0.78	0.81	0.81	0.81	0.62	0.62
	HYBRID	0.84	0.89	0.11	0.85	0.86	0.87	0.86	0.73	0.73
XGB	TRAIN	0.88	0.87	0.13	0.88	0.88	0.87	0.87	0.75	0.74
	VALIDATION	0.89	0.86	0.14	0.86	0.84	0.87	0.88	0.75	0.75
	HYBRID	0.91	0.89	0.11	0.89	0.87	0.90	0.90	0.79	0.79
ALTERNATE DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.85	0.87	0.13	0.53	0.39	0.86	0.86	0.52	0.47

	VALIDATION	0.80	0.86	0.14	0.51	0.37	0.86	0.83	0.48	0.44
	HYBRID	0.83	0.88	0.12	0.55	0.41	0.88	0.86	0.53	0.49
DT	**TRAIN**	0.60	0.97	0.03	0.62	0.65	0.93	0.78	0.59	0.59
	VALIDATION	0.62	0.97	0.03	0.64	0.67	0.94	0.79	0.61	0.61
	HYBRID	0.70	0.96	0.04	0.68	0.67	0.94	0.83	0.65	0.65
KNN	**TRAIN**	0.57	0.96	0.04	0.58	0.59	0.92	0.76	0.54	0.53
	VALIDATION	0.61	0.95	0.05	0.58	0.55	0.92	0.78	0.53	0.53
	HYBRID	0.66	0.95	0.05	0.60	0.55	0.92	0.80	0.56	0.56
SVM	**TRAIN**	0.87	0.85	0.15	0.51	0.36	0.85	0.85	0.49	0.44
	VALIDATION	0.89	0.85	0.15	0.52	0.36	0.85	0.87	0.51	0.45
	HYBRID	0.90	0.87	0.13	0.56	0.41	0.87	0.89	0.55	0.50
XGB	**TRAIN**	0.87	0.92	0.08	0.64	0.51	0.91	0.89	0.62	0.59
	VALIDATION	0.86	0.91	0.09	0.62	0.48	0.90	0.88	0.60	0.57
	HYBRID	0.90	0.91	0.09	0.65	0.51	0.91	0.91	0.64	0.61

Table 4: Performance of ML models developed with features selected using variance threshold

ML models on features extracted by removing correlated features gave poor results. A maximum AUC of 0.68 on training and 0.61 on the validation dataset was achieved in the main dataset with XGB model whereas a maximum AUC of 0.66 on training and 0.66 on the validation dataset was achieved in the alternate dataset with the RF model (Table 5).

MAIN DATASET										
ML MODEL	**Dataset**	**Sensitivity**	**Specificity**	**FPR**	**F1**	**Precision**	**Accuracy**	**AUC**	**MCC**	**kappa**
RF	**TRAIN**	0.69	0.65	0.35	0.68	0.67	0.67	0.67	0.34	0.34
	VALIDATION	0.61	0.54	0.46	0.56	0.52	0.57	0.58	0.15	0.15
	HYBRID	0.84	0.73	0.27	0.78	0.72	0.78	0.79	0.58	0.57
DT	**TRAIN**	0.65	0.68	0.32	0.66	0.68	0.66	0.66	0.33	0.32
	VALIDATION	0.56	0.61	0.39	0.55	0.54	0.59	0.59	0.17	0.17
	HYBRID	0.83	0.77	0.23	0.79	0.75	0.80	0.80	0.60	0.59
KNN	**TRAIN**	0.65	0.62	0.38	0.65	0.65	0.64	0.63	0.28	0.28
	VALIDATION	0.64	0.56	0.44	0.59	0.54	0.59	0.60	0.20	0.19
	HYBRID	0.81	0.73	0.27	0.76	0.71	0.77	0.77	0.54	0.54
SVM	**TRAIN**	0.86	0.30	0.70	0.68	0.56	0.59	0.58	0.19	0.16
	VALIDATION	0.94	0.23	0.77	0.65	0.50	0.55	0.58	0.23	0.15
	HYBRID	0.97	0.61	0.39	0.79	0.67	0.77	0.79	0.60	0.55
XGB	**TRAIN**	0.68	0.67	0.33	0.68	0.69	0.68	0.68	0.35	0.35
	VALIDATION	0.64	0.58	0.42	0.59	0.55	0.61	0.61	0.22	0.22
	HYBRID	0.83	0.73	0.27	0.77	0.72	0.78	0.78	0.56	0.55

ALTERNATE DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.60	0.72	0.28	0.27	0.18	0.71	0.66	0.20	0.16
	VALIDATION	0.61	0.72	0.28	0.27	0.18	0.71	0.66	0.20	0.16
	HYBRID	0.70	0.81	0.19	0.40	0.27	0.80	0.76	0.35	0.30
DT	TRAIN	0.17	0.97	0.03	0.23	0.35	0.90	0.57	0.19	0.18
	VALIDATION	0.25	0.98	0.02	0.36	0.60	0.92	0.62	0.35	0.32
	HYBRID	0.42	0.98	0.02	0.52	0.68	0.93	0.70	0.50	0.49
KNN	TRAIN	0.30	0.94	0.06	0.31	0.32	0.88	0.62	0.25	0.24
	VALIDATION	0.30	0.95	0.05	0.34	0.39	0.89	0.62	0.28	0.28
	HYBRID	0.46	0.96	0.04	0.50	0.55	0.92	0.71	0.46	0.46
SVM	TRAIN	0.86	0.25	0.75	0.18	0.10	0.31	0.56	0.08	0.03
	VALIDATION	0.83	0.27	0.73	0.18	0.10	0.32	0.55	0.06	0.02
	HYBRID	0.85	0.68	0.32	0.33	0.21	0.69	0.76	0.31	0.22
XGB	TRAIN	0.95	0.24	0.76	0.20	0.11	0.31	0.60	0.13	0.04
	VALIDATION	0.89	0.24	0.76	0.19	0.10	0.30	0.57	0.09	0.03
	HYBRID	0.90	0.63	0.37	0.32	0.19	0.65	0.76	0.31	0.20

Table 5: Performance of ML models developed with features selected by removing correlated features

Of the three feature selection methods, removal of correlation co-efficient gave the poorest results. It was observed that the none of the features that remained after removing correlated features were common with the features selected using SVC-L1 or variance threshold (Figure 8,9).

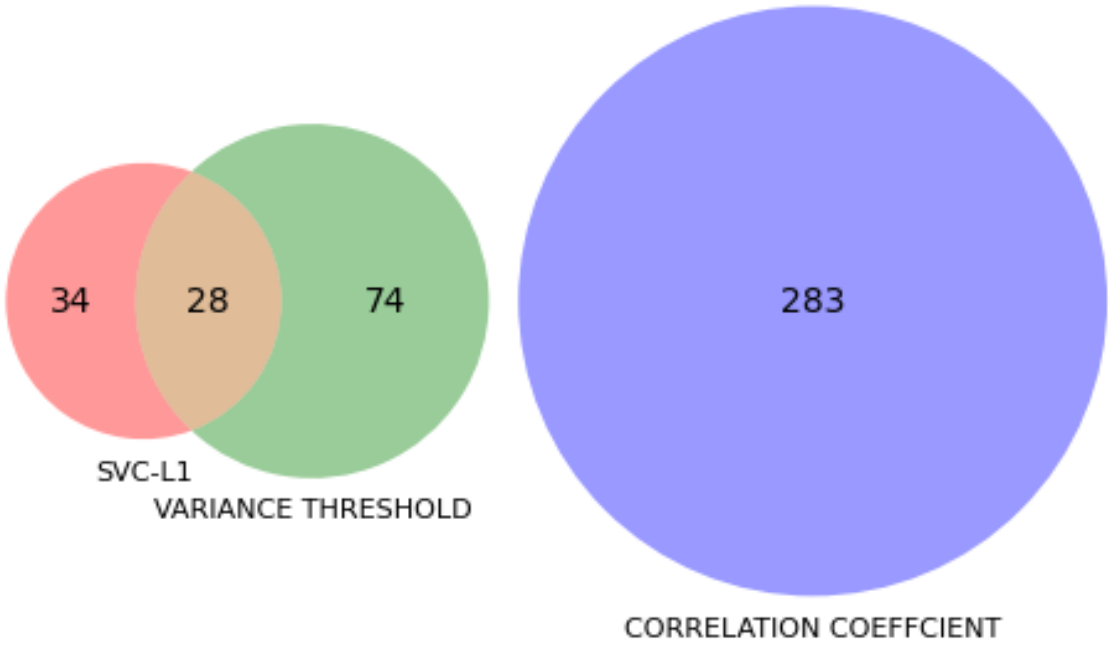

Figure 8: Venn Diagram depicting common features among three sets of features in main dataset

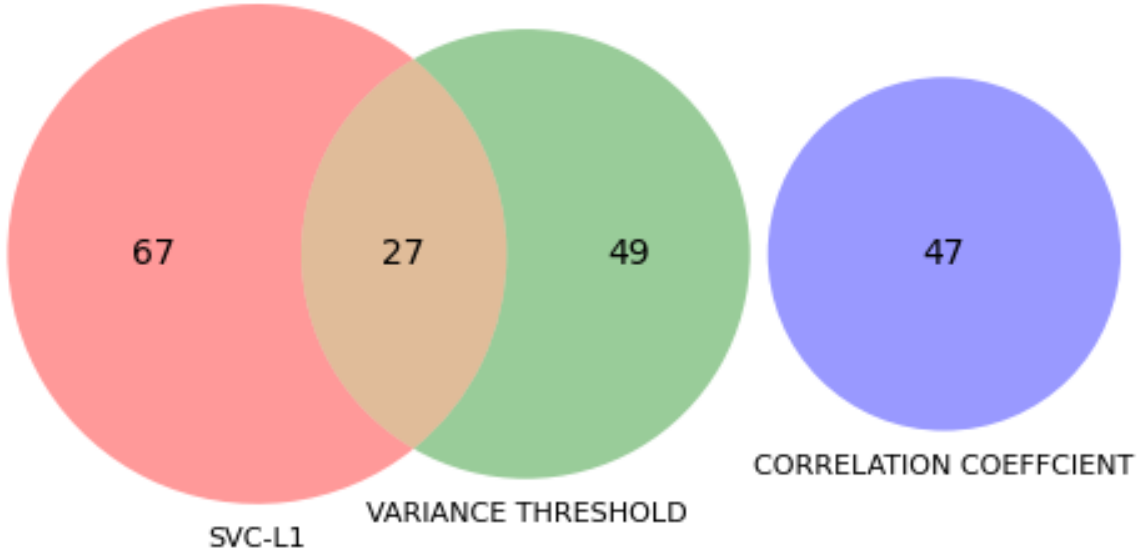

Figure 9: Venn Diagram depicting common features among three sets of features in alternate dataset

By using the PCA method, a maximum AUC of 0.89 on training and 0.92 on the validation dataset was achieved in the main dataset with RF model whereas a maximum AUC of 0.92 on training and 0.88 on the validation dataset was achieved in the alternate dataset with the XGB model (Table 6).

MAIN DATASET										
ML MODEL	**Dataset**	**Sensitivity**	**Specificity**	**FPR**	**F1**	**Precision**	**Accuracy**	**AUC**	**MCC**	**kappa**
RF	**TRAIN**	0.92	0.87	0.128	0.9	0.89	0.89	0.894	0.796	0.788
	VALIDATION	0.95	0.89	0.11	0.91	0.87	0.92	0.92	0.83	0.83
	HYBRID	0.97	0.89	0.11	0.92	0.87	0.92	0.93	0.85	0.85
DT	**TRAIN**	0.85	0.84	0.164	0.848	0.85	0.84	0.844	0.686	0.686
	VALIDATION	0.84	0.87	0.13	0.84	0.84	0.86	0.86	0.72	0.72
	HYBRID	0.86	0.89	0.11	0.86	0.86	0.87	0.87	0.75	0.75
KNN	**TRAIN**	0.87	0.84	0.158	0.86	0.86	0.86	0.856	0.716	0.71
	VALIDATION	0.94	0.84	0.16	0.88	0.82	0.88	0.89	0.77	0.76
	HYBRID	0.95	0.86	0.14	0.9	0.85	0.9	0.91	0.81	0.8
SVM	**TRAIN**	0.884	0.84	0.16	0.868	0.854	0.864	0.862	0.724	0.722
	VALIDATION	0.88	0.84	0.16	0.84	0.81	0.85	0.86	0.71	0.71
	HYBRID	0.91	0.86	0.14	0.87	0.84	0.88	0.88	0.76	0.76
XGB	**TRAIN**	0.896	0.864	0.136	0.886	0.878	0.88	0.88	0.764	0.758
	VALIDATION	0.94	0.9	0.1	0.91	0.88	0.92	0.92	0.83	0.83
	HYBRID	0.97	0.91	0.09	0.93	0.9	0.94	0.94	0.88	0.87
ALTERNATE DATASET										
ML MODEL	**Dataset**	**Sensitivity**	**Specificity**	**FPR**	**F1**	**Precision**	**Accuracy**	**AUC**	**MCC**	**kappa**
RF	**TRAIN**	0.88	0.91	0.092	0.628	0.49	0.91	0.894	0.612	0.58
	VALIDATION	0.87	0.89	0.11	0.59	0.44	0.89	0.88	0.57	0.53
	HYBRID	0.87	0.90	0.1	0.61	0.47	0.90	0.89	0.6	0.56
DT	**TRAIN**	0.66	0.97	0.026	0.682	0.71	0.95	0.814	0.654	0.654
	VALIDATION	0.63	0.97	0.03	0.64	0.65	0.94	0.8	0.61	0.61
	HYBRID	0.72	0.96	0.04	0.69	0.67	0.94	0.84	0.66	0.66
KNN	**TRAIN**	0.734	0.98	0.02	0.76	0.78	0.96	0.856	0.734	0.734
	VALIDATION	0.68	0.98	0.02	0.71	0.75	0.95	0.83	0.69	0.68
	HYBRID	0.77	0.97	0.03	0.76	0.74	0.96	0.87	0.73	0.73
SVM	**TRAIN**	0.866	0.86	0.14	0.53	0.384	0.862	0.862	0.516	0.462
	VALIDATION	0.92	0.86	0.14	0.54	0.39	0.86	0.89	0.54	0.48
	HYBRID	0.92	0.88	0.12	0.58	0.42	0.88	0.9	0.57	0.52
XGB	**TRAIN**	0.922	0.92	0.08	0.674	0.534	0.918	0.92	0.666	0.632
	VALIDATION	0.89	0.88	0.12	0.57	0.42	0.88	0.88	0.56	0.51
	HYBRID	0.89	0.89	0.11	0.6	0.45	0.89	0.89	0.59	0.55

Table 6: Performance of ML models developed with features selected using PCA

Composition based features

A number of machine learning models were developed using the characteristics, such as the amino acid composition (AAC), dipeptide composition (DPC), and tripeptide composition (TPC) of therapeutic and non-therapeutic proteins/peptides (Table 7, 8, 9). When compared to other models for the alternative dataset, it was found that the XGB-based model performed the best on AAC characteristics and had a maximum AUC of 0.91 and 0.91 on the training and validation datasets, respectively (Table 10). On the training and validation datasets for the main dataset's DPC features, the RF-based model achieved AUC values of 0.91 and 0.89, respectively (Table 10).

Motif approach

In both models i.e., RF-based AAC model and XGB-based DPC model, the predicted labels were replaced by the motif labels except for those sequences in which no motif obtained from the training sequences were found. For those test sequences, the predicted label was kept as it is. We observed that doing so improved the prediction score for both main and alternate datasets.

In the main dataset, the AUC score for the validation dataset increased from 0.89 to 0.92, and in the alternate dataset, the AUC score for the validation dataset increased from 0.91 to 0.94 (Table 10). This implicated that presence or absence of these motifs play a role in the classification of sequences into therapeutic and non-therapeutic proteins/peptides.

MAIN DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.91	0.90	0.10	0.91	0.91	0.90	0.91	0.81	0.81
	VALIDATION	0.91	0.87	0.13	0.88	0.85	0.89	0.89	0.78	0.78
	HYBRID	0.94	0.89	0.11	0.90	0.87	0.91	0.91	0.82	0.82
DT	TRAIN	0.82	0.76	0.24	0.80	0.78	0.79	0.79	0.58	0.58
	VALIDATION	0.80	0.77	0.23	0.77	0.74	0.78	0.78	0.57	0.56
	HYBRID	0.88	0.84	0.16	0.84	0.81	0.85	0.86	0.71	0.71
KNN	TRAIN	0.91	0.84	0.16	0.88	0.85	0.87	0.87	0.75	0.75
	VALIDATION	0.92	0.81	0.19	0.86	0.80	0.86	0.87	0.73	0.72
	HYBRID	0.97	0.84	0.16	0.89	0.83	0.90	0.90	0.80	0.79
SVM	TRAIN	0.85	0.86	0.14	0.86	0.87	0.86	0.86	0.71	0.71
	VALIDATION	0.89	0.84	0.16	0.85	0.81	0.86	0.86	0.72	0.72
	HYBRID	0.91	0.87	0.13	0.88	0.85	0.89	0.89	0.78	0.78
XGB	TRAIN	0.89	0.89	0.11	0.89	0.90	0.89	0.89	0.78	0.78
	VALIDATION	0.91	0.89	0.11	0.89	0.87	0.90	0.90	0.79	0.79
	HYBRID	0.92	0.91	0.09	0.91	0.89	0.92	0.92	0.83	0.83

ALTERNATE DATASET										
ML MODEL	**Dataset**	**Sensitivity**	**Specificity**	**FPR**	**F1**	**Precision**	**Accuracy**	**AUC**	**MCC**	**kappa**
RF	**TRAIN**	0.82	0.94	0.06	0.69	0.60	0.93	0.88	0.67	0.65
	VALIDATION	0.83	0.93	0.07	0.65	0.54	0.92	0.88	0.63	0.61
	HYBRID	0.90	0.93	0.07	0.69	0.56	0.93	0.91	0.67	0.65
DT	**TRAIN**	0.63	0.97	0.03	0.64	0.65	0.93	0.80	0.60	0.60
	VALIDATION	0.73	0.94	0.06	0.64	0.57	0.92	0.84	0.60	0.60
	HYBRID	0.82	0.95	0.05	0.69	0.60	0.93	0.88	0.66	0.65
KNN	**TRAIN**	0.78	0.98	0.02	0.80	0.83	0.97	0.88	0.78	0.78
	VALIDATION	0.70	0.98	0.02	0.74	0.78	0.96	0.84	0.72	0.72
	HYBRID	0.79	0.97	0.03	0.76	0.74	0.96	0.88	0.74	0.74
SVM	**TRAIN**	0.87	0.89	0.11	0.59	0.44	0.89	0.88	0.57	0.53
	VALIDATION	0.86	0.87	0.13	0.55	0.40	0.87	0.87	0.53	0.48
	HYBRID	0.93	0.89	0.11	0.61	0.45	0.89	0.91	0.60	0.55
XGB	**TRAIN**	0.89	0.94	0.06	0.71	0.59	0.93	0.91	0.69	0.67
	VALIDATION	0.90	0.92	0.08	0.67	0.54	0.92	0.91	0.66	0.63
	HYBRID	0.94	0.93	0.07	0.71	0.56	0.93	0.94	0.70	0.67

Table 7: Performance of ML models developed with amino acid composition features

MAIN DATASET										
ML MODEL	**Dataset**	**Sensitivity**	**Specificity**	**FPR**	**F1**	**Precision**	**Accuracy**	**AUC**	**MCC**	**kappa**
RF	**TRAIN**	0.91	0.92	0.08	0.92	0.92	0.91	0.91	0.83	0.83
	VALIDATION	0.89	0.90	0.10	0.88	0.88	0.90	0.89	0.79	0.79
	HYBRID	0.92	0.92	0.08	0.91	0.91	0.92	0.92	0.84	0.84
DT	**TRAIN**	0.84	0.79	0.21	0.82	0.81	0.81	0.81	0.63	0.62
	VALIDATION	0.81	0.84	0.16	0.81	0.80	0.83	0.82	0.65	0.65
	HYBRID	0.89	0.86	0.14	0.86	0.84	0.87	0.88	0.75	0.75
KNN	**TRAIN**	0.91	0.73	0.27	0.84	0.78	0.82	0.82	0.66	0.64
	VALIDATION	0.89	0.70	0.30	0.79	0.70	0.78	0.79	0.59	0.57
	HYBRID	0.92	0.72	0.28	0.81	0.73	0.81	0.82	0.65	0.63
SVM	**TRAIN**	0.87	0.87	0.13	0.87	0.88	0.87	0.87	0.74	0.74
	VALIDATION	0.88	0.86	0.14	0.85	0.84	0.87	0.87	0.73	0.73
	HYBRID	0.92	0.89	0.11	0.89	0.87	0.90	0.90	0.80	0.80
XGB	**TRAIN**	0.88	0.88	0.12	0.88	0.89	0.88	0.88	0.76	0.76
	VALIDATION	0.88	0.86	0.14	0.85	0.84	0.87	0.87	0.73	0.73
	HYBRID	0.89	0.89	0.11	0.88	0.86	0.89	0.89	0.77	0.77
ALTERNATE DATASET										
ML MODEL	**Dataset**	**Sensitivity**	**Specificity**	**FPR**	**F1**	**Precision**	**Accuracy**	**AUC**	**MCC**	**kappa**
RF	**TRAIN**	0.81	0.95	0.05	0.71	0.62	0.94	0.88	0.68	0.67
	VALIDATION	0.82	0.93	0.07	0.66	0.55	0.92	0.87	0.63	0.61
	HYBRID	0.85	0.94	0.06	0.69	0.58	0.93	0.89	0.67	0.65
DT	**TRAIN**	0.61	0.96	0.04	0.63	0.65	0.94	0.79	0.60	0.60

	VALIDATION	0.66	0.96	0.04	0.65	0.64	0.93	0.81	0.61	0.61
	HYBRID	0.76	0.97	0.03	0.72	0.69	0.95	0.86	0.70	0.70
KNN	TRAIN	0.76	0.95	0.05	0.68	0.61	0.93	0.86	0.65	0.64
	VALIDATION	0.72	0.94	0.06	0.63	0.55	0.92	0.83	0.59	0.58
	HYBRID	0.82	0.94	0.06	0.66	0.56	0.92	0.88	0.64	0.62
SVM	TRAIN	0.78	0.97	0.03	0.75	0.72	0.95	0.87	0.72	0.72
	VALIDATION	0.82	0.96	0.04	0.73	0.67	0.95	0.89	0.71	0.70
	HYBRID	0.86	0.95	0.05	0.74	0.65	0.95	0.91	0.72	0.71
XGB	TRAIN	0.92	0.93	0.07	0.71	0.58	0.93	0.93	0.70	0.68
	VALIDATION	0.92	0.91	0.09	0.65	0.51	0.91	0.91	0.64	0.61
	HYBRID	0.93	0.92	0.08	0.68	0.54	0.92	0.92	0.67	0.64

Table 8: Performance of ML models developed with amino di-peptide composition features

MAIN DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.92	0.85	0.15	0.89	0.87	0.88	0.89	0.77	0.77
	VALIDATION	0.95	0.86	0.14	0.90	0.85	0.90	0.91	0.81	0.80
	HYBRID	0.97	0.87	0.13	0.91	0.86	0.92	0.92	0.84	0.83
DT	TRAIN	0.87	0.75	0.25	0.82	0.79	0.81	0.81	0.63	0.62
	VALIDATION	0.89	0.86	0.14	0.86	0.84	0.87	0.88	0.75	0.75
	HYBRID	0.95	0.87	0.13	0.90	0.86	0.91	0.91	0.82	0.82
KNN	TRAIN	0.89	0.67	0.33	0.81	0.74	0.78	0.78	0.58	0.56
	VALIDATION	0.97	0.67	0.33	0.82	0.70	0.80	0.82	0.65	0.62
	HYBRID	0.98	0.70	0.30	0.83	0.72	0.83	0.84	0.69	0.66
SVM	TRAIN	0.95	0.85	0.15	0.91	0.87	0.90	0.90	0.80	0.80
	VALIDATION	0.95	0.86	0.14	0.90	0.85	0.90	0.91	0.81	0.80
	HYBRID	0.97	0.87	0.13	0.91	0.86	0.92	0.92	0.84	0.83
XGB	TRAIN	0.91	0.81	0.19	0.87	0.83	0.86	0.86	0.72	0.72
	VALIDATION	0.91	0.82	0.18	0.85	0.81	0.86	0.86	0.73	0.72
	HYBRID	0.97	0.84	0.16	0.89	0.83	0.90	0.90	0.80	0.79
ALTERNATE DATASET										
ML MODEL	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
RF	TRAIN	0.88	0.92	0.08	0.65	0.52	0.91	0.90	0.63	0.60
	VALIDATION	0.89	0.92	0.08	0.66	0.52	0.92	0.90	0.64	0.61
	HYBRID	0.92	0.93	0.07	0.69	0.55	0.92	0.92	0.68	0.65
DT	TRAIN	0.61	0.98	0.02	0.67	0.74	0.95	0.80	0.64	0.64
	VALIDATION	0.56	0.98	0.02	0.63	0.71	0.94	0.77	0.60	0.60
	HYBRID	0.73	0.97	0.03	0.73	0.72	0.95	0.85	0.70	0.70
KNN	TRAIN	0.75	0.91	0.09	0.57	0.47	0.89	0.83	0.54	0.52
	VALIDATION	0.70	0.93	0.07	0.59	0.51	0.91	0.82	0.55	0.54
	HYBRID	0.77	0.93	0.07	0.62	0.51	0.91	0.85	0.59	0.57
SVM	TRAIN	0.73	1.00	0.00	0.82	0.95	0.97	0.86	0.81	0.81
	VALIDATION	0.68	0.99	0.01	0.77	0.89	0.96	0.83	0.76	0.75
	HYBRID	0.80	0.98	0.02	0.81	0.81	0.97	0.89	0.79	0.79
XGB	TRAIN	0.92	0.92	0.08	0.68	0.54	0.92	0.93	0.67	0.64
	VALIDATION	0.93	0.92	0.08	0.69	0.55	0.92	0.93	0.68	0.65
	HYBRID	0.93	0.93	0.07	0.70	0.56	0.93	0.93	0.69	0.66

Table 9: Performance of ML models developed with tri-peptide composition features

Model	Dataset	Sensitivity	Specificity	FPR	F1	Precision	Accuracy	AUC	MCC	kappa
DPC based RF model scores on alternate dataset	TRAIN	0.91	0.92	0.08	0.92	0.92	0.91	0.91	0.83	0.83
	VALIDATION	0.89	0.90	0.10	0.88	0.88	0.90	0.89	0.79	0.79
	HYBRID	0.92	0.92	0.08	0.91	0.91	0.92	0.92	0.84	0.84
AAC based XGB model scores on main dataset	TRAIN	0.89	0.94	0.06	0.71	0.59	0.93	0.91	0.69	0.67
	VALIDATION	0.90	0.92	0.08	0.67	0.54	0.92	0.91	0.66	0.63
	HYBRID	0.94	0.93	0.07	0.71	0.56	0.93	0.94	0.70	0.67

Table 10: Performance of Best Models

Design and implementation of a web server

For the purpose of predicting therapeutic proteins/peptides, ThpPred (https://webs.iiitd.edu.in/raghava/thppred/) has been created. Our top four models, Model-1 (AAC-based XGB approach), Model-2 (AAC-based XGB + motif approach), Model-3 (DPC-based RF approach), and Model-4 (DPC-based RF + motif approach), have all been put into practice. For the purpose of predicting therapeutic proteins/peptides, models 1 and 2 are trained on the alternate dataset, whereas models 3 and 4 are trained on the main dataset. The web server integrates the key components, including (i) prediction, (ii) motif scan, (iii) therapeutic protein design, and (iv) download. The 'prediction module' enables users to submit both single and multiple protein sequences in FASTA format. Effectively separating therapeutic from non-therapeutic proteins/peptides is possible with this module. Utilizing only motifs, found in therapeutic proteins/peptides, the "motif scan module" may be used to separate therapeutic proteins/peptides from non-therapeutic proteins/peptides. Additionally, it maps or scans the patterns in the user-provided query protein sequence. The 'Therapeutic protein design module' helps the user by creating therapeutic proteins/peptides with a single point mutation in their query sequence. With a responsive HTML template and support for several operating systems, the web server is developed. We have created a ThpPred standalone Python package, which can be used through the web server's "Download" module, to let users predict therapeutic proteins/peptides at the genome scale.

Chapter 5: DISCUSSION

The focus of this work is on the creation of ThpPred, a machine learning model for predicting therapeutic proteins/peptides that may be used to new drug discoveries. In contrast to earlier models that concentrated on proteins/peptides with therapeutic properties, the new model was trained on protein sequences of medications that have previously been produced. It is emphasised how important proteins/peptides are to biological processes and how effective they may be as therapeutic agents. The bulk of biological activities in organisms are carried out by proteins/peptides, which have also been connected to the start of some illnesses. The use of therapeutic proteins/peptides and peptides in the treatment of diseases including diabetes, cancer, and neurological problems has shown to be a successful strategy. The desire to speed up medication development and cut down on costs makes therapeutic protein prediction crucial.

Virtual screening and machine learning are two computational techniques that are essential for expediting the identification of therapeutic proteins/peptides. Machine learning-based predictors have proven to be capable of swiftly and efficiently identifying therapeutic proteins/peptides by analysing massive volumes of data.

The ThpPred model was created in an effort to enhance the selection of treatments and medications by concentrating on proteins/peptides that have already been utilised to create pharmacological molecules. The model intends to give improved predictions of therapeutic proteins/peptides for upcoming drug discoveries by using protein sequences of existing medications. To hasten the process of creating better drugs, therapeutic protein prediction is essential. The therapeutic potential of isolated proteins/peptides must thus be established immediately. ThpPred, a tool that combines motif search with RF and XGB models based on amino acid composition and dipeptide composition to predict if a protein is therapeutic or not, has been created. Machine learning alone cannot explain the reason for therapeutic activity of a protein. So, we applied the Motif based approach combined with ML.

This work also sheds information on issues with protein medication development, such as poor in vivo stability and membrane impermeability. Through methods like rational design and phage display, which enable the quick creation of effective and selective lead proteins/peptides, research efforts have been concentrated on overcoming these constraints.

The ThpPred model was developed using a variety of feature selection techniques, including SVC-L1, variance threshold, and correlation-coefficient elimination. According to the findings, the SVC-L1 and variance threshold approaches performed better than correlation-coefficient elimination.

Additionally, Principal Component Analysis (PCA) was utilised, with positive outcomes in terms of AUC ratings. The calculation of amino acid composition (AAC), dipeptide composition (DPC), and tripeptide composition (TPC) characteristics for creating machine learning models is also mentioned in this paper. The prediction scores for therapeutic proteins/peptides were enhanced by the models built on AAC, DPC, and motif incorporation. As seen in Table 10, adding motif labels to ML predicted labels raises the AUC score. As shown in Table 10, we have obtained the best performance with balanced sensitivity and specificity and more accuracy. This implies that if we remove negative motifs from sequences or add positive motifs to in negative sequences, we may be able to convert them from non-therapeutic to therapeutic proteins/peptides.

A web server was created to make the ThpPred model is available, allowing users to contribute protein sequences and predict therapeutic proteins/peptides. The server also has modules for designing therapeutic proteins/peptides with particular mutations as well as a module for detecting themes connected to therapeutic proteins/peptides. We have created a thorough platform in the current work so that users may categorize therapeutic and non-therapeutic proteins/peptides. We think that people researching protein or peptide therapeutics will find our findings to be helpful. To help the scientific community and promote wider use of the recommended prediction technique, we provided a free web server and a standalone package of ThpPred. The website now has the model that consistently predicts both therapeutic and non-therapeutic proteins/peptides. It is envisaged that the researchers will make considerable use of this prediction approach to develop more effective and precise protein-based therapies for treating a range of ailments.

In a nutshell the creation of the ThpPred machine learning model is a crucial step towards the prediction of therapeutic proteins/peptides for upcoming drug discoveries. The model attempts to improve the selection of therapies and pharmaceuticals by using the protein sequences of well-established drugs, which will ultimately result in the creation of more effective treatments and better patient outcomes. Combining the potential of protein-based drugs with improvements in computational techniques has the potential to revolutionise personalised medicine by addressing unmet medical needs.

Chapter 6: LIMITATIONS AND FUTURE SCOPE

It was discovered during data compilation that sequences of quite a few drugs could not be retrieved from any source. More sequence data would have helped us to train the model better. Even though 356 positive sequences, yielded good results, a larger sample size would have improved the model even more. As, more sequences data of therapeutic proteins/peptides will increase in the near future, we shall train the model again to make the predictions more robust and accurate.

References

[1] D. S. Dimitrov, "Therapeutic proteins," *Methods Mol Biol,* vol. 899, pp. 1-26, 2012.

[2] C. Johnson-Leger, C. A. Power, G. Shomade, J. P. Shaw, and A. E. Proudfoot, "Protein therapeutics--lessons learned and a view of the future," *Expert Opin Biol Ther,* vol. 6, no. 1, pp. 1-7, Jan 2006.

[3] V. Fathi Vavsari and S. J. J. o. t. I. C. S. Balalaie, "An overview on the two recent decades' study of peptides synthesis and biological activities in Iran," pp. 1-21, 2022.

[4] K. N. Day, *Peptide Affinity Ligands for Next-Generation Downstream Bioprocessing*. North Carolina State University, 2020.

[5] Y. P. Zhang and Q. Zou, "PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning," *Bioinformatics,* vol. 36, no. 13, pp. 3982-3987, Jul 1 2020.

[6] P. K. Pramanik, M. N. Alam, D. Roy Chowdhury, and T. Chakraborti, "Drug Resistance in Protozoan Parasites: An Incessant Wrestle for Survival," *J Glob Antimicrob Resist,* vol. 18, pp. 1-11, Sep 2019.

[7] S. K. Auld and M. C. Tinsley, "The evolutionary ecology of complex lifecycle parasites: linking phenomena with mechanisms," *Heredity (Edinb),* vol. 114, no. 2, pp. 125-32, Feb 2015.

[8] L. Wang *et al.*, "Therapeutic peptides: Current applications and future directions," vol. 7, no. 1, p. 48, 2022.

[9] K. Fosgerau and T. J. D. d. t. Hoffmann, "Peptide therapeutics: current status and future directions," vol. 20, no. 1, pp. 122-128, 2015.

[10] X. Chen, Z. L. Ji, and Y. Z. J. N. a. r. Chen, "TTD: therapeutic target database," vol. 30, no. 1, pp. 412-415, 2002.

[11] K. Yan, H. Lv, J. Wen, Y. Guo, and B. J. C. B. Liu, "TP-MV: therapeutic peptides prediction by multi-view learning," vol. 17, no. 2, pp. 174-183, 2022.

[12] M. Attique, M. S. Farooq, A. Khelifi, and A. J. I. A. Abid, "Prediction of therapeutic peptides using machine learning: computational models, datasets, and feature encodings," vol. 8, pp. 148570-148594, 2020.

[13] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test," *J Theor Biol,* vol. 341, pp. 34-40, Jan 21 2014.

[14] F. M. Li and X. Q. Wang, "Identifying anticancer peptides by using improved hybrid compositions," *Sci Rep,* vol. 6, p. 33910, Sep 27 2016.

[15] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artif Intell Med,* vol. 79, pp. 62-70, Jun 2017.

[16] C. Wu, R. Gao, Y. Zhang, and Y. J. B. b. De Marinis, "PTPD: predicting therapeutic peptides by deep learning and word2vec," vol. 20, no. 1, pp. 1-8, 2019.

[17] L. Wei, C. Zhou, R. Su, and Q. Zou, "PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning," *Bioinformatics,* vol. 35, no. 21, pp. 4272-4280, Nov 1 2019.

[18] K. Yan, H. Lv, J. Wen, Y. Guo, Y. Xu, and B. Liu, "PreTP-Stack: Prediction of Therapeutic Peptides Based on the Stacked Ensemble Learing," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 20, no. 2, pp. 1337-1344, Mar-Apr 2023.

[19] Y. Guo, K. Yan, H. Lv, and B. Liu, "PreTP-EL: prediction of therapeutic peptides based on ensemble learning," *Brief Bioinform,* vol. 22, no. 6, Nov 5 2021.

[20] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res,* vol. 46, no. D1, pp. D1074-D1082, Jan 4 2018.

[21] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics,* vol. 22, no. 13, pp. 1658-9, Jul 1 2006.

[22] A. Pande *et al.*, "Pfeature: A Tool for Computing Wide Range of Protein Features and Building Prediction Models," *J Comput Biol,* vol. 30, no. 2, pp. 204-222, Feb 2023.

[23] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," vol. 12, pp. 2825-2830, 2011.

[24] J. J. M. L. M. Brownlee, "How to choose a feature selection method for machine learning," vol. 10, 2019.

[25] X. Bouthillier *et al.*, "Accounting for variance in machine learning benchmarks," vol. 3, pp. 747-769, 2021.

[26] F. Kherif and A. Latypova, "Principal component analysis," in *Machine Learning*: Elsevier, 2020, pp. 209-225.

[27] C. Vens, M. N. Rosso, and E. G. Danchin, "Identifying discriminative classification-based motifs in biological sequences," *Bioinformatics,* vol. 27, no. 9, pp. 1231-8, May 1 2011.

[28] P. Geurts, D. Ernst, and L. J. M. l. Wehenkel, "Extremely randomized trees," vol. 63, pp. 3-42, 2006.

[29] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. J. J. o. C. A. J. o. t. C. S. Brown, "An introduction to decision tree modeling," vol. 18, no. 6, pp. 275-285, 2004.

[30] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "k-Nearest Neighbor Classification," in *Data Mining in Agriculture*New York, NY: Springer New York, 2009, pp. 83-106.

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.

[32] C. Zhang, X. Shao, and D. J. P. C. S. Li, "Knowledge-based support vector classification based on C-SVC," vol. 17, pp. 1083-1090, 2013.

[33] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, and G. P. J. B. i. B. Raghava, "AlgPred 2.0: an improved method for predicting allergenic proteins
and mapping of IgE epitopes," vol. 22, no. 4, p. bbaa294, 2021.

[34] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, and G. P. J. B. i. b. Raghava, "AntiCP 2.0: an updated model for predicting anticancer peptides," vol. 22, no. 3, p. bbaa153, 2021.

[35] S. Saha and G. P. S. J. N. a. r. Raghava, "AlgPred: prediction of allergenic proteins and mapping of IgE epitopes," vol. 34, no. suppl_2, pp. W202-W209, 2006.

[36] N. Sharma, S. Patiyal, A. Dhall, N. L. Devi, G. P. J. C. i. B. Raghava, and Medicine, "ChAlPred: A web server for prediction of allergenicity of chemical compounds," vol. 136, p. 104746, 2021.