San Jose State University

SJSU ScholarWorks

Faculty Research, Scholarly, and Creative Activity

11-1-2023

Prospection of Peptide Inhibitors of Thrombin from Diverse Origins Using a Machine Learning Pipeline

Nivedha Balakrishnan
Alumni

Rahul Katkar
Alumni

Peter V. Pham
San Jose State University, peter.pham@sjsu.edu

Taylor Downey
School of Engineering

Prarthna Kashyap
Alumni

See next page for additional authors

Follow this and additional works at: https://scholarworks.sjsu.edu/faculty_rsca

Recommended Citation

Nivedha Balakrishnan, Rahul Katkar, Peter V. Pham, Taylor Downey, Prarthna Kashyap, David C. Anastasiu, and Anand K. Ramasubramanian. "Prospection of Peptide Inhibitors of Thrombin from Diverse Origins Using a Machine Learning Pipeline" *Bioengineering* (2023). https://doi.org/10.3390/bioengineering10111300

This Article is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Faculty Research, Scholarly, and Creative Activity by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Authors

Nivedha Balakrishnan, Rahul Katkar, Peter V. Pham, Taylor Downey, Prarthna Kashyap, David C. Anastasiu, and Anand K. Ramasubramanian

This article is available at SJSU ScholarWorks: https://scholarworks.sjsu.edu/faculty_rsca/4689

bioengineering

Article

Prospection of Peptide Inhibitors of Thrombin from Diverse Origins Using a Machine Learning Pipeline

Nivedha Balakrishnan [1], Rahul Katkar [1], Peter V. Pham [1], Taylor Downey [2], Prarthna Kashyap [1], David C. Anastasiu [2] and Anand K. Ramasubramanian [1,*]

[1] Department of Chemical and Materials Engineering, San José State University, San Jose, CA 95192, USA; prarthna.kashyap@sjsu.edu (P.K.)
[2] Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA 95053, USA; danastasiu@scu.edu (D.C.A.)
* Correspondence: anand.ramasubramanian@sjsu.edu

Citation: Balakrishnan, N.; Katkar, R.; Pham, P.V.; Downey, T.; Kashyap, P.; Anastasiu, D.C.; Ramasubramanian, A.K. Prospection of Peptide Inhibitors of Thrombin from Diverse Origins Using a Machine Learning Pipeline. *Bioengineering* **2023**, *10*, 1300. https://doi.org/10.3390/bioengineering10111300

Academic Editors: Giovanni Vozzi and Alan Wang

Received: 14 September 2023
Revised: 30 October 2023
Accepted: 4 November 2023
Published: 9 November 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Thrombin is a key enzyme involved in the development and progression of many cardiovascular diseases. Direct thrombin inhibitors (DTIs), with their minimum off-target effects and immediacy of action, have greatly improved the treatment of these diseases. However, the risk of bleeding, pharmacokinetic issues, and thrombotic complications remain major concerns. In an effort to increase the effectiveness of the DTI discovery pipeline, we developed a two-stage machine learning pipeline to identify and rank peptide sequences based on their effective thrombin inhibitory potential. The positive dataset for our model consisted of thrombin inhibitor peptides and their binding affinities (K_I) curated from published literature, and the negative dataset consisted of peptides with no known thrombin inhibitory or related activity. The first stage of the model identified thrombin inhibitory sequences with Matthew's Correlation Coefficient (MCC) of 83.6%. The second stage of the model, which covers an eight-order of magnitude range in K_I values, predicted the binding affinity of new sequences with a log room mean square error (RMSE) of 1.114. These models also revealed physicochemical and structural characteristics that are hidden but unique to thrombin inhibitor peptides. Using the model, we classified more than 10 million peptides from diverse sources and identified unique short peptide sequences (<15 aa) of interest, based on their predicted K_I. Based on the binding energies of the interaction of the peptide with thrombin, we identified a promising set of putative DTI candidates. The prediction pipeline is available on a web server.

Keywords: antithrombotic; anticoagulant; peptide design; classification; regression

1. Introduction

Ever since its discovery in 1872, thrombin has occupied the center stage in the pathophysiology of cardiovascular diseases [1]. Over the years, important roles of thrombin have also been identified in the pathophysiology of a multitude of other diseases including cancer, autoimmune and inflammatory disorders, and most recently in COVID-19 [2–4]. The most prominent function of thrombin is the conversion of plasma fibrinogen to a crosslinked polymeric fibrin network, the structural and functional unit of the blood clot. The inability to generate sufficient thrombin can result in hemorrhage, while unregulated and excessive thrombin generation can lead to thrombosis and tissue damage. Thrombin is uniquely capable of regulating its own production through positive and negative feedback loops involving other enzymes of coagulation and complement activation cascades [5]. Therefore, molecules that precisely tune thrombin activity have always been both fundamental and clinically relevant and are of interest to academia and pharma [6].

Thrombin activity can be modulated either indirectly by altering thrombin generation rates or directly by interfering with thrombin action. Direct thrombin inhibitors (DTIs) are a new class of anticoagulants that bind directly to thrombin and block its interaction with its substrates [7]. The peptide hirudin and its derivatives, including bivalirudin, lepirudin, and desirudin, are the largest class of DTI and have been approved by the US FDA for the treatment of heparin-induced thrombocytopenia, percutaneous coronary intervention, and for prophylaxis against venous thromboembolism. Small peptidomimetic DTIs (dabigatran and argatroban) have also been in clinical use. DTIs are potent antithrombotic agents as they have a higher capacity for the inhibition of fibrin-bound thrombin than indirect inhibitors, such as thrombin, since bound thrombin can actively promote thrombus growth. They also have shorter half-lives in plasma and do not require cofactors for their activity. Despite these advantages, current DTIs are contraindicated in certain situations or limited in their applicability as they are associated with bleeding or thrombotic risks [8]. Therefore, efforts to discover new antithrombotic agents are necessary to meet the capacious demands of adverse cardiovascular events [9].

To this end, we sought machine learning approaches for the discovery of new peptide inhibitors of thrombin. We are motivated to work with peptide inhibitors of thrombin because: (1) the clinical relevance of hirudin, the most well-known and widely utilized DTI, which is a peptide; (2) the analogs and derivatives of hirudin, particularly bivalirudin, with properties that are significantly more desirable than the parent hirudin; (3) the fundamental importance of thrombin-inhibiting peptides in the survival of hematophagous animals, such as leech, snakes, and ticks. In fact, hirudin was discovered from the saliva of the leech *Hirudo medicinalis*; and (4) the wide range of binding affinities spanning nearly 8-orders of magnitude in existing thrombin-inhibiting peptides suggests that peptide inhibitors offer an opportunity for fine-tuning the inhibitory potential.

When carefully built, ML-based models can rapidly screen a vast chemical and biological space, enabling accurate and faster in silico predictions of the biological activity of new molecules. As a cheminformatics model, ML combines chemistry, computer science, and information technology to aid in drug discovery through tasks like virtual screening, library design, and high-throughput screening analysis [10–12]. Machine learning algorithms leverage large chemical datasets for predictive modeling and pattern recognition, including the prediction of the properties and activities of peptides based on their sidechains [13–16]. This integration has accelerated the discovery and design of novel peptides with desired biological activities, opening new avenues for peptide-based drug development. Recently, machine learning models have been developed for the discovery of novel antimicrobial peptides [17], anticancer peptides [18], antibiofilm peptides [19], antihypertensive peptides [20], and membrane-active peptides [21]. Some of the model predictions have already started to show promise in in vitro and in vivo tests [22,23]. Despite the clinical and fundamental importance of antithrombotic peptides, there have not been any attempts to exploit machine learning for the expedient discovery of new thrombin inhibitors with desired activity levels.

In this work, we develop a computational ML pipeline to predict the thrombin inhibitory activity of the peptides from their properties, which are quantitative structure-activity relationship (QSAR) descriptors. This two-stage pipeline consists of classification models to identify, from a very large peptide database, a handful of hits with thrombin inhibitory activity, and regression models that predict the level of activity of the peptide hits. The models processed the multi-dimensional QSAR properties to prioritize those that are key to thrombin inhibition. The identified hits were then ranked based on their binding affinity to thrombin, as determined by the molecular modeling of their interactions. The prediction pipeline is available on the web server: https://thrombin-inhibitor-peptide-predictor.info (accessed on 3 November 2023).

2. Methods

2.1. Dataset Preparation

For our study, we collected datasets from various sources including literature, patents, and protein databases.

- Positive Dataset. We collected only direct thrombin-inhibiting peptides reported as “antithrombotic” from peer-reviewed publications. We collected the sequences of these peptides from UniProt [24], NCBI Protein Database [25], RCSB PDB [26], and PubChem [27]. Next, we obtained the experimentally determined inhibition constants of these peptides against thrombin, also from peer-reviewed publications. After removing the duplicates, we obtained 88 naturally occurring antithrombotic peptides, and inhibition constants of 53 of these peptides (Table S1). Only peptides containing naturally occurring amino acids were chosen.
- Negative Dataset. To prepare the non-antithrombotic negative dataset, we collected peptides from the UniProt and NCBI databases that were not annotated as “anticoagulant”, “antithrombotic”, “hemostasis-impairing”, “antimicrobial”, “anti-inflammatory”, or “thrombin inhibitor”. To minimize bias in the random selection of peptides agnostic to thrombin binding, the ratio of collected negative to positive peptides was 9:1, and we maintained a similar ratio for different sequence lengths within the dataset. We compared the sequences within the negative dataset between the negative and positive datasets for an 80% sequence match and removed the ones above this threshold. Finally, we obtained a negative dataset with 792 sequences.
- Test dataset. To identify thrombin-inhibiting activity in new peptides, we collected a total of 10,743,304 peptides from the UniProt and NCBI protein databases. We searched for the source organism as one of ‘fungi’, ‘bacteria’, ‘snakes’, ‘leeches’, ‘humans’, ‘mice’, ‘eukaryota’, and ‘viruses’, and the results were filtered to a sequence length between 5 and 200 amino acids. Peptides in the sequence range of 5 to 15 amino acids were collected independently of the source. The peptides could be true peptides or random fragments of large proteins.

2.2. Feature Extraction

We extracted features to describe the peptide sequences using global protein sequence descriptors [28]. These features are:

- Global Physico-Chemical Properties (PCP). We used the ‘Biopython ProtParam’ package to extract the global properties of the collected sequences which include sequence length, molecular weight, aromaticity, isoelectric point, and instability. This constitutes a 5-element vector.
- Amino Acid Composition (AAC). The amino acid composition (AAC) is a measure that quantifies the relative abundance of each amino acid in a peptide sequence. These features were extracted using the ‘propy3′ python package [29]. The following equation represents the amino acid composition function:

$$\text{AAC (i)} = \frac{Total\ number\ of\ amino\ acid\ of\ type\ (i)}{Total\ number\ of\ amino\ acids} \times 100. \quad (1)$$

- Composition Transition Distribution (CTD). The CTD descriptor is a 147-element vector that describes different physico-chemical properties of a peptide [30] (Table S2). The physico-chemical properties covered by CTD features are ‘polarity’, ‘polarizability’, ‘charge’, ‘secondary structure’, ‘hydrophobicity’, ‘normalized van der Waals volume’, and ‘solvent accessibility’. The CTD descriptor groups the amino acids into three classes for each physico-chemical property. The composition (C) descriptor describes the global percentage of each class in a peptide sequence, the transition (T) descriptor characterizes the percent frequency of transitions between two classes in a peptide sequence, and the distribution (D) descriptor specifies the distribution patterns of each class in a sequence. These CTD properties were extracted using the ‘propy3 CTD’ package.
- Dipeptide Composition (DPC). The DPC descriptor was extracted using the ‘propy3 AAComposition’ package which returns a 400-element vector containing percent

fractions of dipeptides, i.e., AA, AC, AD, . . ., VY, and VV, in a peptide sequence. The DPC fraction percentage is calculated as follows:

$$\text{DPC (i, j)} = \frac{Total\ number\ of\ dipeptides\ of\ amino\ acid\ of\ type\ (i\ and\ j)}{Total\ number\ of\ possible\ dipeptides\ available} \times 100. \quad (2)$$

Together, 572 peptide features were extracted which were used for training machine learning models using the 880 peptides in the positive and negative datasets.

2.3. Classification Models

We implemented machine learning models for predicting thrombin inhibitory activity, employing various classification algorithms. To evaluate model performance, we randomly split the data into three sets: 60% for training, 20% for validation, and 20% for testing, which ensures an adequate presence of positive samples in all datasets. The training set consisted of 528 samples, the validation set of 176 samples, and the testing set of 176 samples, which served as out-of-sample test data. We employed a support vector machine (SVM) with both linear and radial basis function kernels, logistic regression, random forest, k-nearest neighbors (kNN), and extreme gradient boosting (XGBoost) models for classification. The performance of these models was compared to determine the most suitable one for our final inference. We implemented these models using the widely used scikit-learn package [31], a popular machine learning library in Python.

First, all the baseline models were tuned by choosing the hyperparameters that lead to the best Matthew's Correlation Coefficient (MCC) score on the validation set. We used the RandomizedSearchCV package for hyperparameter tuning and performed 5-fold cross-validation across the joint training and validation data sets. The SVM models were tuned for hyperparameters C and γ; the random forest and XGBoost models for the number of estimators, maximum depth, minimum sample leaf node, and minimum sample split; the logistic regression model was tuned for C; and the kNN model for the number of nearest neighbors. The imbalance in the dataset was accounted for by setting the 'class_weight' parameter of the classifiers to 'balanced' which adjusts the weights according to class frequencies in the dataset. The final models were tested on labeled out-of-sample test sets and their MCC performance was compared with the average MCC score obtained during cross-validation to ensure that the models did not have high generalization errors. Based on these criteria, the best-performing model was chosen and was used to predict thrombin inhibitory activity in peptides collected from protein databases. The performance of the classification models was also estimated using Accuracy and F_1 score (harmonic mean of precision and recall). The three effectiveness measures are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}, \quad (3)$$

$$F_1 = \frac{TP}{TP + \frac{FP+FN}{2}}, \quad (4)$$

$$\text{MCC} = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (5)$$

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

2.4. Clustering

We utilized clustering techniques to group peptides based on similar characteristics. This approach was applied to both the positive peptides in our dataset and the new test peptides collected from protein databases. To ensure uniqueness among the new test peptides, we selected medoids as representatives for each cluster. We employed a hierarchical clustering model to group together similar peptides as determined by the Euclidean distance metric between the feature sets [32]. The cluster proximity metric used was Ward's linkage, which aims to minimize the variance within each cluster when merging clusters. We used the agglomerative clustering algorithm from the scipy.cluster.hierarchy module [33]. For each threshold value given by the clustering algorithm, the number of clusters and silhouette scores were obtained. The optimal number of clusters was determined based on the highest silhouette score, defined as $(b - a)/\max(a,b)$, where 'a' and 'b' represent the average distance between a data point and all other points within the same cluster, and the average distance between the data point and all points in the nearest neighboring cluster, respectively. We computed the Euclidean distance matrix between peptides within each cluster using the scipy.spatial.distance module. Then, we identified medoids, the peptides with the smallest total distance to all other peptides within their cluster, which capture the characteristics of all the peptides in their respective clusters.

2.5. Regression Models

In addition to predicting thrombin inhibitory activity, we also utilized regression models to estimate the inhibition constant (K_I) for the positive peptides. Specifically, we employed three regression models, Support Vector Regressor with radial basis function, Support Vector Regressor with linear kernel, and Lasso regression, which were trained using the sklearn.svm and sklearn.linear_model modules. Out of the 88 positive samples in our dataset, 53 had a K_I value. To evaluate the performance of the regression models, we split the dataset into training, validation, and testing sets following a 60%, 20%, and 20% ratio. As the inhibition constant (K_I) values span a wide range of eight orders of magnitude, we converted them to a logarithmic scale before training the regression models. The best model was selected based on the root mean squared error (RMSE) score, defined as RMSE = sqrt $(1/n * \sum (y - y\_pred)^2)$, where n is the number of data points, y is the actual value of the target variable, and y_pred is the predicted value of the target variable. To improve the model performance, the hyperparameters of these models were tuned through 5-fold cross-validation using GridSearchCV; the support vector models were optimized for C, ε, and γ, while the lasso regression was optimized for α. To prevent overfitting, we adopted a Sequential Forward Selection (SFS) for feature selection with 5-fold cross-validation to select the optimal set of features. This new set of features was used in all the models. Finally, we selected the regression model with the lowest average validation RMSE score over the 5 folds in the 5-fold cross-validation as the best-performing model, which was used to predict the K_I values of new peptides.

2.6. Molecular Docking

For docking peptide sequences with thrombin, we used two web servers, HPEPDOCK [34] and CABS-dock [35]. The PDB file of thrombin (1PPB) and the peptide sequence was entered as inputs to the server utilizing default parameters, and the top-ranked pose of the protein–peptide complex was used for further analysis. The binding strength between the peptide and thrombin was inferred from the docking score estimated by HPEPDOCK. The binding energy was estimated by first obtaining the protein–peptide complex in the PDB format from CABS-dock. Then, the complex structure was entered as input to another server, PRODIGY [36]. Lastly, the determination of the binding sites of the peptide on thrombin was accomplished using PyMol, using a cutoff of 5 Å distance between the interacting atoms.

3. Results

The machine learning pipeline to identify potent inhibitors of thrombin consisted of three phases (Figure 1). First, a classification model was built to predict peptides with thrombin inhibitory activity. Second, a regression model was constructed to estimate the K_I values of peptides with thrombin inhibitory activity. Third, a large set of peptides from databases was tested to identify new peptide candidates that exhibit thrombin-inhibiting activity. The predictions of the machine learning model were confirmed by protein docking studies.

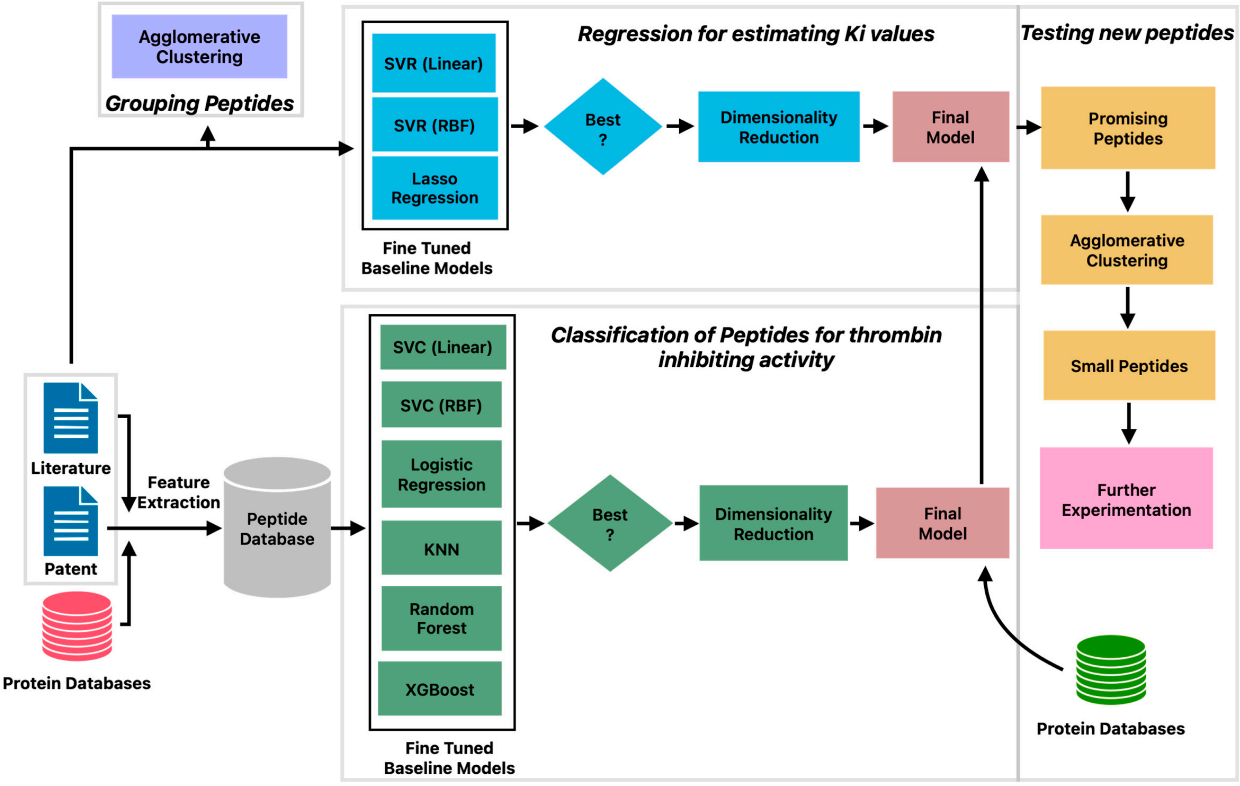

Figure 1. Machine learning pipeline for prediction of peptides with potent thrombin inhibitory activity.

3.1. Characteristics of Thrombin-Inhibiting Peptides

Thrombin inhibitory peptides are direct thrombin inhibitors (DTI) that block the proteolytic action of thrombin on its substrates by interacting with one or more of the three binding sites, namely, the active site (AS), the Anion Binding Exosite I (ABEI), and the Anion Binding Exosite II (ABEII) [37,38] (Figure S1). The AS cleft is hydrophobic and flanked by the exosites on either side. Recent studies have revealed that thrombin activity is also influenced by the binding of the substrates to two additional exosites, namely, the hydrophilic γ-loop and Na^+ binding site. Bivalent DTIs, such as hirudin and bivalirudin, bind to both the active site and exosite I of thrombin, which enhances their inhibitory effect on thrombin activity.

From the published literature, we curated 88 DTI peptides and inhibition constants (K_I) for 53 of these peptides. These sequences, the K_I values, and the reference sources are listed in Table S1, and their 578 features are listed in a supplementary file. The vast majority of the peptides in the positive dataset contained less than 160 amino acids (97.1%), and most of the peptides (82.7%) contained between 3 and 70 amino acids (Figure 2A). As shown in Figure 2B,C, these peptides contain higher percentages of E (13.6%), G (9.1%), D (8.7%), and P (8.7%), and lower percentages of W (0.34%), M (0.7%), H (1.9%), Y (3.6%), Q (3.7%) and I (3.9%). C is over-represented in certain peptides, P19, P44–P49, and P60–P69, suggesting the presence of disulfide linkages (Figure 2B and Table S1). Compared to the peptides in the negative dataset, the thrombin-inhibiting peptides have higher percentages of negatively charged amino acids, E and D and P and G, and lower percentages of positively charged amino acid K, and also the hydrophobic amino acids L, M, V, and A. The highest dipeptide compositions with an average of more than 1.0% in the entire positive dataset are shown in Figure 3A. The composition of dipeptide EE is 2.75%, and of DF, FE, PE, and GD are ~1.7% in the positive dataset, while the corresponding values are 0.13–0.40% in the negative dataset, indicating the importance of these dipeptides on thrombin inhibition. The other dipeptides at significantly higher percentages in the positive dataset are IP, EY, EI, PR, SD, DE, AE, and YL.

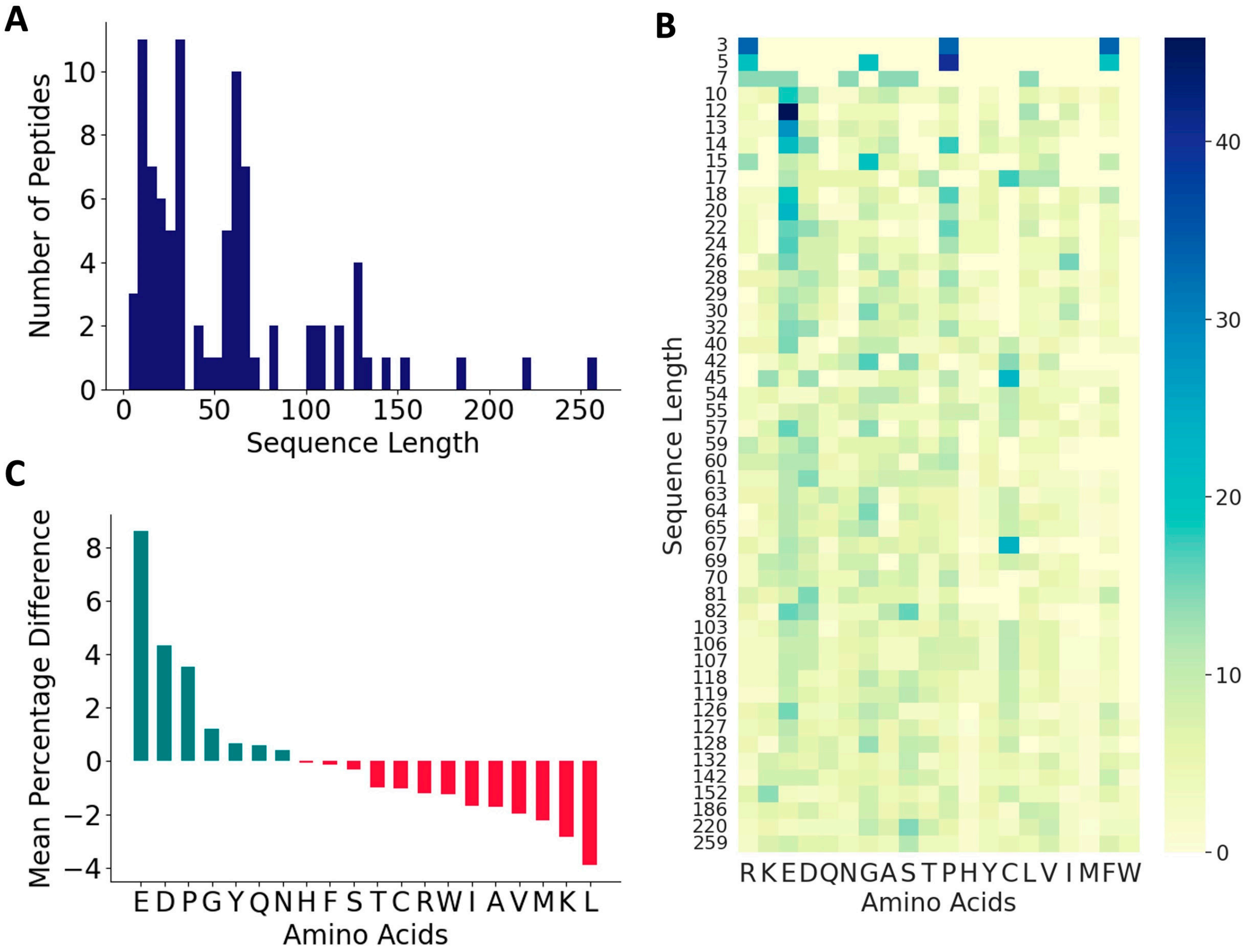

Figure 2. Characteristics of antithrombotic peptides in the positive dataset. (**A**) Distribution of sequence lengths among the 88 peptides in the positive dataset; (**B**) heatmap of various amino acids across peptides of varied sequence lengths in the positive dataset; and (**C**) relative distributions of amino acids.

Next, we analyzed the distributions of isoelectric point, aromaticity, charge, hydropathy index, and secondary structure proclivities between the positive and negative peptides (Figure 3B). The median isoelectric points of positive and negative peptides were 4.24 and 7.92, respectively, indicating that the positive peptides will be negatively charged at a neutral pH. This is also indicated in the charge distributions in the positive and negative peptides. The median aromaticity values of the positive and negative peptides do not differ significantly. The hydropathy index showed that the positive peptides had a higher distribution of hydrophilic and neutral amino acids compared to the negative peptides, while the negative peptides had a higher distribution of hydrophobic amino acids. The thrombin inhibitory peptides are likely to form more coils compared to the negative peptides, which is consistent with the higher fraction of P and G in the positive dataset.

To investigate the commonality between peptides derived from various sources, we first performed a multiple sequence alignment among all positive peptides, but that approach did not reveal any obvious homology between the sequences. Therefore, we used agglomerative clustering to allow for the evolution of peptide clusters based on underlying similarities between the peptides. Agglomerative clustering is a versatile hierarchical clustering algorithm that does not require specifying the number of clusters in advance. The clustering is performed not only using the amino acid sequence information, but also all the 572 features described above. The peptides are represented using a dendrogram, a tree-like structure that is constructed by merging clusters from the bottom up. The height of fusion on the vertical axis indicates the dissimilarity between the two peptides. Based on the silhouette scores of agglomerative clustering, the optimal number of clusters for the positive set was found to be 48 (Figure S2). While 30 clusters contained only one peptide, 9 clusters contained two peptides, and the largest cluster contained eight peptides (Figure S2).

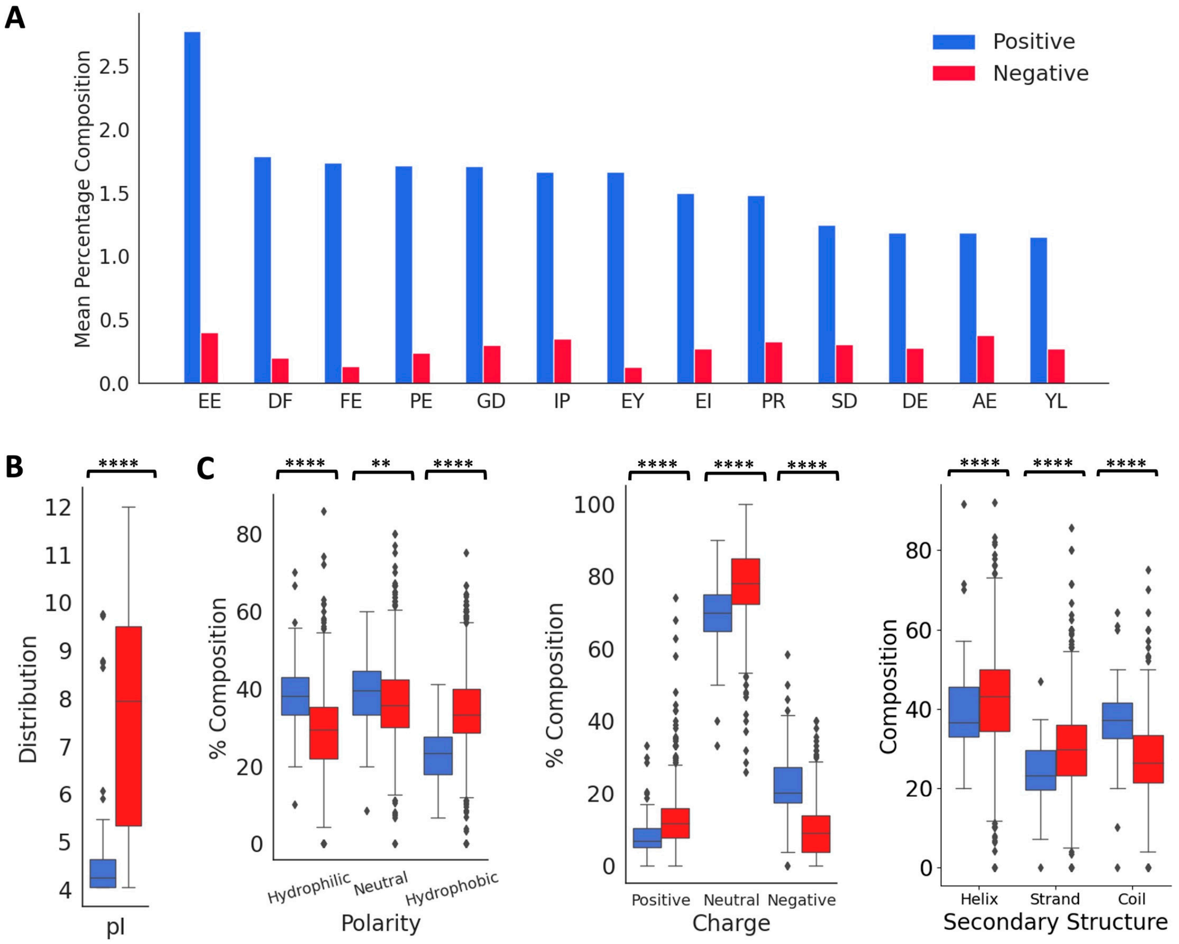

Figure 3. Distribution of key features in the positive and negative datasets. (**A**) Comparison of mean values of dipeptide distributions most represented in the positive dataset. (**B**) Comparison of physico-chemical properties and secondary structure distributions. '**' for $p < 0.01$, and '****' for $p < 0.0001$.

We constructed a dendrogram to visualize the relationships between peptides in different clusters at various levels (Figure 4). At the lowest level, most of the peptides derived from organisms belonging to the same family cluster together, as indicated by a coloring scheme. These families include: (a) genus *Haemaphysalis* (hard-bodied and bush ticks); (b) genus *Ornithodoros* (soft tick); (c) genus *Anopheles* (mosquito); (d) genus Bothropos (pit viper snake); (e) genera *Hirudo* and *Hirudinaria* (leech); (f) genus *Dipetalogaster* (kissing bug); (g) genus *Crassostrea* (oysters); and (h) genus *Amblyomma* (ixodid tick). Not surprisingly, synthetic peptides clustered together because they were designed from the same template. Further, the peptides belonging to the same cluster inhibited thrombin through similar mechanisms. For instance, clusters (a) and (d) contained peptides that were reported to bind the active site and the exosite II, and to exosite I and exosite II of thrombin, respectively. Most other peptides inhibit thrombin by binding to both active site and exosite I. In contrast, at the highest level, the dendrogram reveals two large clusters: the larger peptides from various natural sources clustered together as one group, while shorter and synthetic peptides clustered together as another group, indicated as top and bottom in Figure 4.

Name	**Source**	**Sequence Length**	
Haemathrin-2	*Haemaphysalis bispinosa*	59	P50
Madanin-2	*Haemaphysalis longicornis*	61	P54
Haemathrin-1	*Haemaphysalis bispinosa*	59	P51
Unmodified Madanin-1	*Haemaphysalis longicornis*	60	P53
Anophelin	*Anopheles albimanus*	30	P35
Anophelin	*Anopheles albimanus*	61	P56
Anophelin	*Anopheles albimanus*	61	P55
Anophelin	*Anopheles albimanus*	61	P57
Anophelin	*Anopheles albimanus*	61	P58
	Bothrops insularis	132	P83
	Bothrops insularis	127	P80
Bothrojaracin	*Bothrops jararaca*	259	P88
Hirudin Ia	*Hirudo medicinalis*	17	P19
	Hirudinaria manillensis	42	P45
Hirudin HM2	*Hirudinaria manillensis*	64	P61
Hirullin P18	*Hirudinaria manillensis*	61	P59
	Hirudinaria manillensis	64	P62
Desirudin	Recombinant Hirudin	65	P63
Hirudin II	*Hirudo medicinalis*	65	P64
Recombinant Hirudin P18	*Hirudinaria manillensis*	65	P65
Lepirudin	LT-desulfato hirudin	65	P66
Hirudin P6	*Hirudinaria manillensis*	63	P60
Hirudin-HN	*Hirudo nipponia*	69	P69
Rhodinin	*Rhodnius prolixus*	103	P74
Dipetalin I	*Dipetalogaster maximus*	55	P48
Infestin I	*Triatoma infestans*	103	P73
Dipetalin I+II	*Dipetalogaster maximus*	106	P75
Dipetalogastin	*Dipetalogaster maximus*	107	P76
Chimadinin	*Haemaphysalis longicornis*	70	P70
Hyalomin-1	*Hyalomma marigatum rufipes*	59	P52
cE5	*Anopheles gambiae*	82	P72
TTI	*Glossina morsitans morsitans*	32	P41
Thrombostatin	*Haematobia irritans*	81	P71
Monobin	*Argas monolakensis*	128	P82
Savignin	*Ornithodoros savignyi*	118	P77
Ornithodorin	*Ornithodoros moubata*	119	P78
PEBP	Rat brain	186	P86
Amblin	*Amblyomma hebraeum*	152	P85
Antibody	Human Fab phage library	220	P87
Bicolin	*Vespa bicolor Fabricius*	54	P47
AaTI	*Aedes aegypti*	65	P67
Hemalin	*Haemaphysalis longicornis*	126	P79
Boophilin	*Boophilus microplus*	127	P81
CrSPI-1D1	*Carcinoscorpius rotundicauda*	40	P44
Granulin-like	*Hirudo nipponia*	45	P46
Threomin	*Theromyzon tessulatum*	67	P68
Haemadin	*Haemadipsa sylvestris*	57	P49
Triabin	*Triatoma pallidipennis*	142	P84
	Synthetic	5	P2
	Tenebrio molitor	10	P5
Peptide 26	*Mytilus edulis mussels*	13	P12
FbpA	Human fibrinogen	15	P18
	Isolated	14	P15
FbpB	Human fibrinogen	15	P17
	Synthetic	3	P1
	Tenebrio molitor larvae	10	P4
P-2-CG	*Crassostrea gigas*	12	P10
	Milk peptide	10	P7
P-3-CG	*Crassostrea gigas*	7	P3
	Isolated	10	P6
S12H	*Amblyomma variegatum*	29	P33
S12A	*Amblyomma variegatum*	30	P39
QT26	*Amblyomma variegatum*	26	P29
beta-avathrin	*Amblyomma variegatum*	30	P40
K10R	*Amblyomma variegatum*	30	P36
Avathrin	*Amblyomma variegatum*	30	P37
IS20	*Amblyomma variegatum*	20	P22
G2D,Q5D	*Amblyomma variegatum*	30	P38

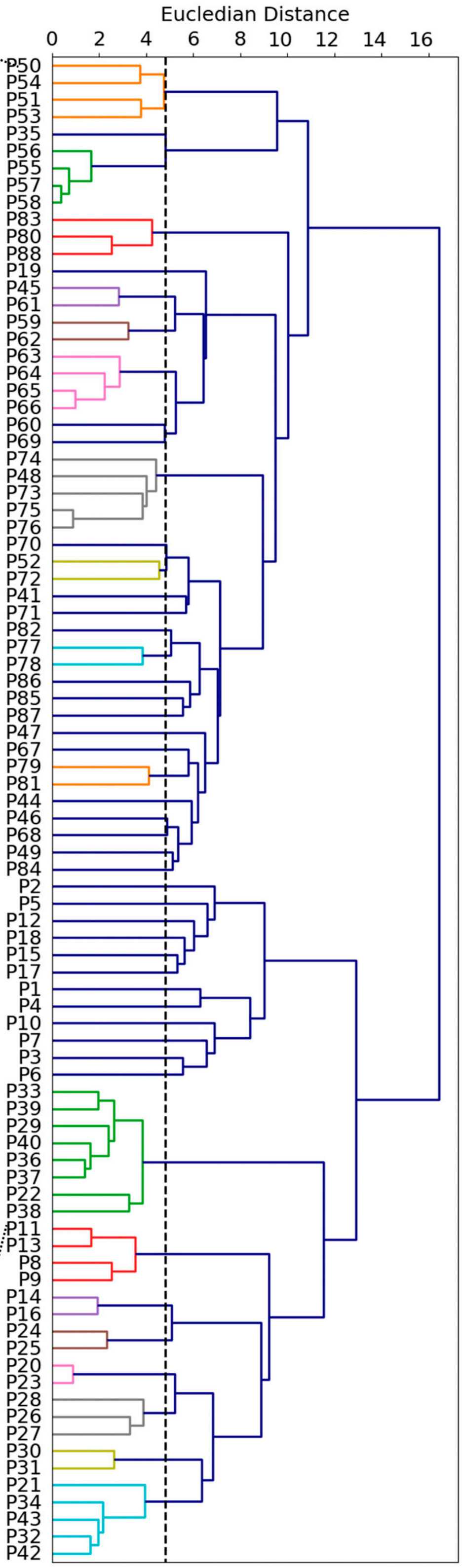

Figure 4. Clustering of positive peptides. Dendrogram visualization of positive peptides reveals that these peptides group in two large clusters. The natural, known peptides are indicated. The peptides at the bottom are synthetic peptides and are not assigned a specific name.

3.2. Development of Machine Learning Models for Thrombin Inhibition

The clustering analysis presented above reveals large diversity in the patterns of amino acid sequences in thrombin-inhibiting peptides sourced from different organisms. This suggests that a more sophisticated approach is needed to explore the relationship between these patterns and the inhibitory potential of the peptides, which may guide toward an intelligent drug design. Toward this end, we constructed a machine learning pipeline (Figure 1) consisting of six steps: curation of positive and negative datasets, development of classification model, prediction of new antithrombotic peptides, identification of unique peptides by clustering, development of regression model, and prediction of inhibition constants.

We trained several classification models using the support vector classifier with linear kernel (SVC-L) and with radial bias function kernel (SVC-R), random forest (RF), k-nearest neighbor (KNN), logistic regression (LR), and the XGBoost (XGB) algorithms. We employed 5-fold cross-validation to test the performance of these models, and estimated accuracy, F_1 scores, and MCC. We tuned the hyperparameters of each model to obtain the best MCC scores (Table S3). We focus on MCC scores for two reasons: First, MCC is widely adopted in peptide classification due to its ability to handle imbalanced datasets and provide a balanced evaluation metric. Second, MCC considers both true positives and true negatives, making it a more comprehensive and reliable performance measure compared to the F_1 score. To evaluate the variance, the model performance was computed for ten different combinations of training, validation, and test sets. As shown in Figure 5A, the MCC scores of all the models for the training set were higher (96.6% to 100%) than those of the validation set (73.5% to 79.7%) and the test set (71.1% to 77.6%). Among the baseline models, SVC RBF achieved relatively good performance with a validation MCC of 79.7% and a test score of 77.5%. XGBoost also showed promising results with a validation MCC of 78.5% and a test score of 77.6%. It is important to note that these scores are the mean values obtained from 10 different sets. It is also worth noting that the standard deviation of the test scores is relatively high compared to the validation scores, indicating some variability in the model's performance on unseen data. This suggests that the models may not generalize as well to new and unseen samples.

To improve the validation and test MCC scores, we sought to implement feature reduction techniques. A pairwise correlation analysis revealed the existence of a strong correlation ($<|0.8|$) between 77 features and 54 pairs (Figure S3). Therefore, we performed feature engineering by applying the Recursive Feature Elimination (RFE) algorithm with tuned SVC-L, LR, RF, and XGB models. Sequential Forward Selection (SFS) was performed for the SVC-R and KNN models. The performance of the feature-engineered models is summarized in Figure 5B, and the optimal hyperparameters for these models are listed in Table S4. After applying RFE/SFS, significant improvements were observed in the validation and test MCC scores for the SVC-L, SVC-R, and LR models. The SVC-L model showed the best performance with a validation MCC score of 83.6% and a test MCC score of 81.1%, followed by the LR model, which showed a validation MCC score of 83.4% and a test MCC score of 82.1%. On the other hand, the KNN model showed a decrease in performance with a validation MCC score of 69.8% and a test MCC score of 64.2%, accompanied by higher standard deviations for both sets. The optimal number of features that gave the best performance for each model is given in Table 1. Based on these results, the SVC-L model with RFE-reduced features (120 features) was selected as the best-performing model.

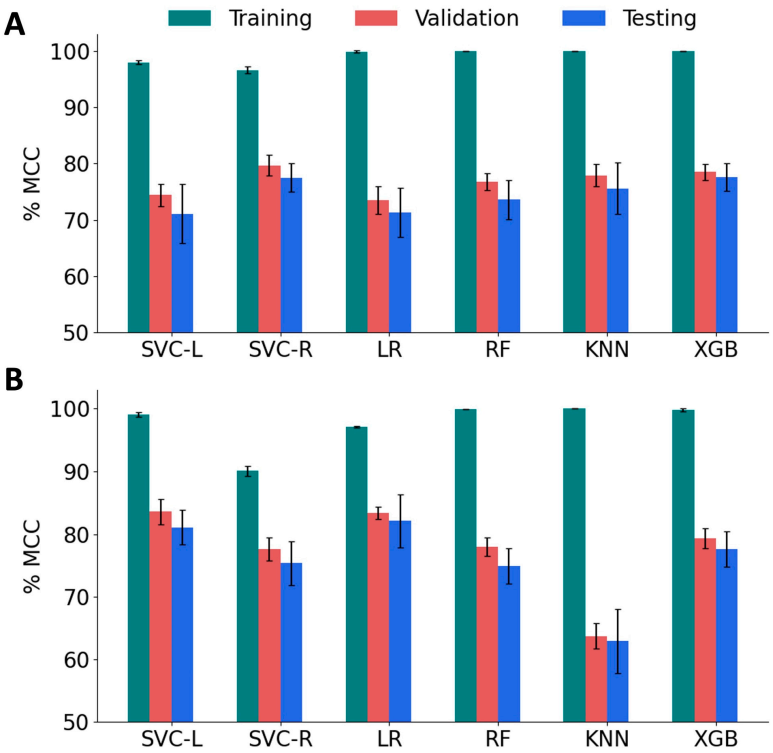

Figure 5. Performance of machine learning models on various sets of features. (**A**) The Matthew's Correlation Coefficient (MCC) score for baseline classification models, namely, SVM linear kernel (SVC-L), SVM RBF kernel (SVC-R), logistic regression (LR), random forest (RF), and k-nearest neighbors (KNN) and XGBoost (XGB). (**B**) MCC of reduced-feature classification models.

Table 1. Optimal number of features selected from RFE/SFS.

Model	Feature Reduction Method	Number of Features
SVC Linear	RFE	120
SVC RBF	SFS	314
Logistic Regression	RFE	54
Random Forest	RFE	508
KNN	SFS	257
XGBoost	RFE	32

3.3. Prediction of Antithrombotic Efficacy of Peptide Hits

Next, we developed a feature-based regression model to predict the inhibition constants (K_I) of thrombin inhibitory peptides. The original training dataset consisted of 53 peptides with K_I ranging from femtomolar to millimolar. This dataset was filtered to remove ineffective outliers with values larger than 40 µM, which reduced the dataset to 49 peptides. The inhibition constants still spanned eight orders of magnitude, and therefore, we logarithmically scaled the values expressed in nanomolar to the range -5 to +5. We first tested three models, namely, SVM with linear kernel (SVR-L), SVM with RBF kernel (SVR-R), and Lasso regression (Lasso).

Among the baseline regression models, the SVR with RBF kernel demonstrated the best performance with a 5-fold validation RMSE of 1.847 and a test RMSE of 1.541. We applied Sequential Forward Selection (SFS) on all models to improve the model performance.

The SFS algorithms improved the performance of all the models, as seen by a decrease in RMSE (Table 2). Based on the performance of the model on the validation dataset, we chose the SVR with RBF kernel with 125 features as the final regression model. As shown in Figure 6, the predicted K_I values correlate well with ground truth K_I (R^2 = 0.93) over the entire range of concentrations, indicating the robust performance of the model. The predicted values of K_I for the positive dataset, including peptides without known K_I values, are presented in Table S5.

Table 2. Comparison of performance of regression models before and after SFS optimization.

Model	Stage	Log Training RMSE	Log Validation RMSE	Log Test RMSE
SVR with Linear Kernel	Baseline	0.715	1.951	1.47
	SFS with 51 features	0.778	1.149	1.221
SVR with RBF Kernel	Baseline	0.279	1.847	1.541
	SFS with 125 features	0.2	1.114	1.06
Lasso Regression	Baseline	1.14	1.887	1.802
	SFS with 28 features	1.388	1.728	1.107

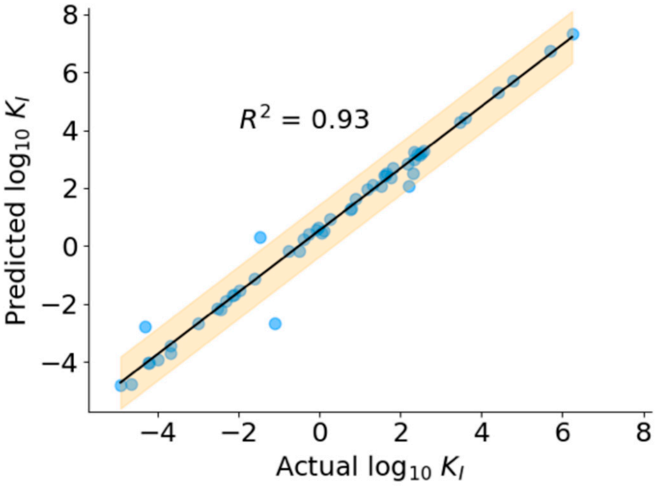

Figure 6. Performance of regression model.

3.4. Prediction of Antithrombotic Activity in Test Peptides

We used the feature-reduced SVM model with the linear kernel to classify 10,743,304 new peptides as those with and without thrombin inhibitory activity. These test peptides represented various sources including fungi (36%), humans (23.3%), eukaryotes (10.8%), bacteria (8.6%), snakes (0.98%), viruses (6.4%), leeches (0.16%), mice (0.36%), and others (13.4%), as shown in the Figure 7A. We extracted the features of these test peptides, as mentioned in the Methods section, and applied the classification model. The model classified 50,325 peptides out of the total test set as positive peptides, or a first pass hit rate of 0.46%. To eliminate false positives, we ranked these positively classified peptides based on their distance from the SVM hyperplane, determined by the decision function of the SVM classifier. A similar ranking of the training set showed that a cutoff value of 0.5 is a good separator to distinguish any erroneously classified false positives (Figure S4). Therefore, using a threshold of 0.5, we found that 15,645 peptides may be considered as true positives with a second pass hit rate of 0.089%. To determine the antithrombotic efficacy of selected peptides, we used the final regression model to predict the K_I values of the 15,645 peptides (Figure 7B). As mentioned earlier, low K_I values indicate a higher affinity for the target and are preferred in terms of thrombin-inhibiting activity. Therefore, to further refine our peptide selection to those that may be truly efficacious, we specifically considered those with $\log_{10}(K_I)$ <0 (i.e., 1 nM or less) and a decision function ≤ 2, which resulted in a total of 308 peptides, comprising both short and long sequences (Figure 7C).

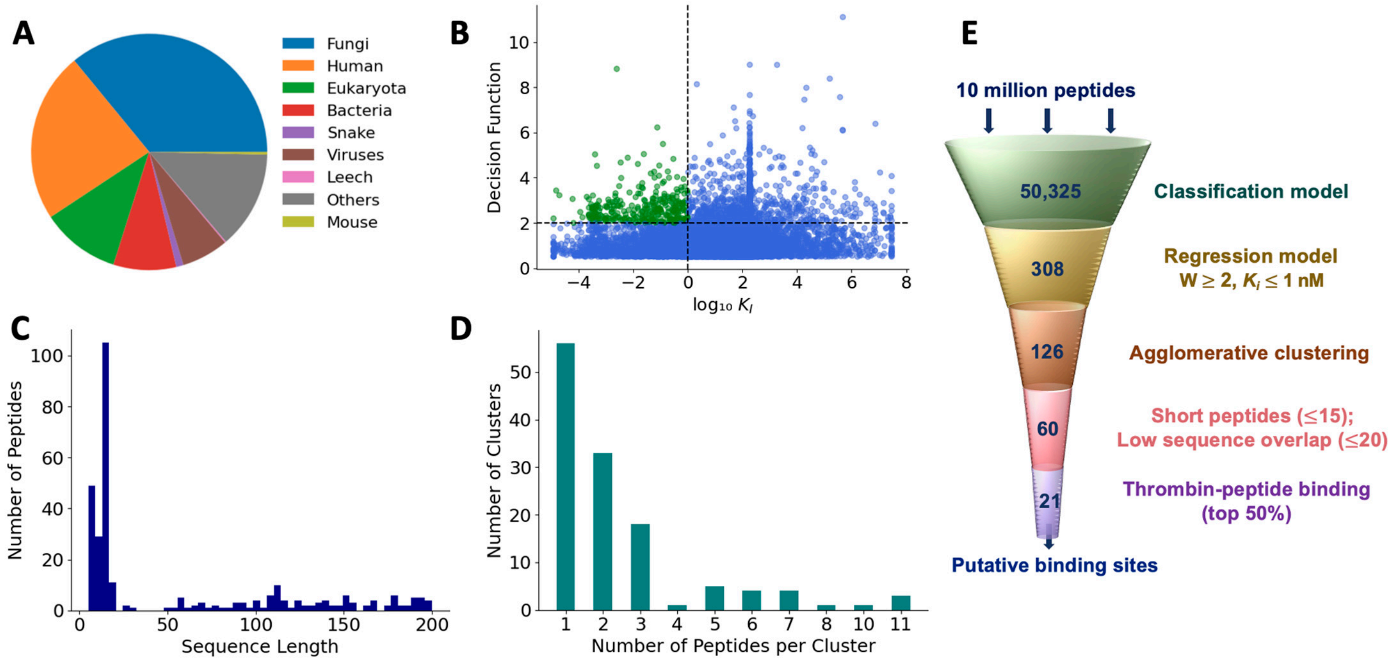

Figure 7. Prediction of antithrombotic peptide hits. (**A**) Distribution of sources of test peptides; (**B**) most efficacious peptides were chosen based on higher decision function and lower K_I; (**C**) sequence length of hits; (**D**) hierarchical clustering of the first-pass hits; and (**E**) filtering pipeline to obtain efficacious peptide hits.

3.5. Clustering of Hits to Identify Unique Peptides

To reduce the set of peptide hits to only unique sequences, we pared down those with overlapping sequences using an agglomerative clustering algorithm with 120 features that were employed in the final classification model. Based on the optimal silhouette score (Figure S5), we found that the 308 peptides were grouped into 126 clusters. The number of clusters showed an exponential distribution with a long tail. While the 56 clusters contained only 1 peptide, 3 clusters contained as many as 11 peptides (Figure 7D). Since each cluster is uniquely represented by a medoid, we obtained 126 medoids. Further, since we are interested only in short peptides for potential translational benefits, we selected medoids with a sequence length of less than 15. After eliminating peptides with >80% sequence similarity, we obtained 59 peptides as the most promising hits from the machine learning-based screening algorithm (Table S6).

3.6. Ranking of Top Hits Based on Binding Scores

To determine the quality of the hits obtained using the machine learning model, we used HPEPDOCK and CABS-dock to dock the 59 peptide sequences with thrombin. HPEPDOCK implements a hierarchical docking protocol with fast conformational sampling of peptide conformations followed by ensemble docking, while CABS-dock uses a replica exchange Monte Carlo algorithm. The docking scores for top-ranked poses from HPEPDOCK are a measure of their binding affinity. Since CABS-dock does not provide such a score, we used the web server PRODIGY to obtain the binding energy for the top-ranked poses generated by CABS-dock. As the most likely candidates for effective thrombin inhibition, we chose 21 peptides that were in the top 50 percentile of the binding/docking scores by both methods (Figure 8). This corresponds to an HPEPDOCK docking score between

−224.97 kcal/mol and −166.31 kcal/mol, and CABS-dock derived the binding energy of −11 kcal/mol and −7.5 kcal/mol. The binding affinity computed from the CABS-dock model for these peptides was predicted to be between 9 nM and 1500 nM.

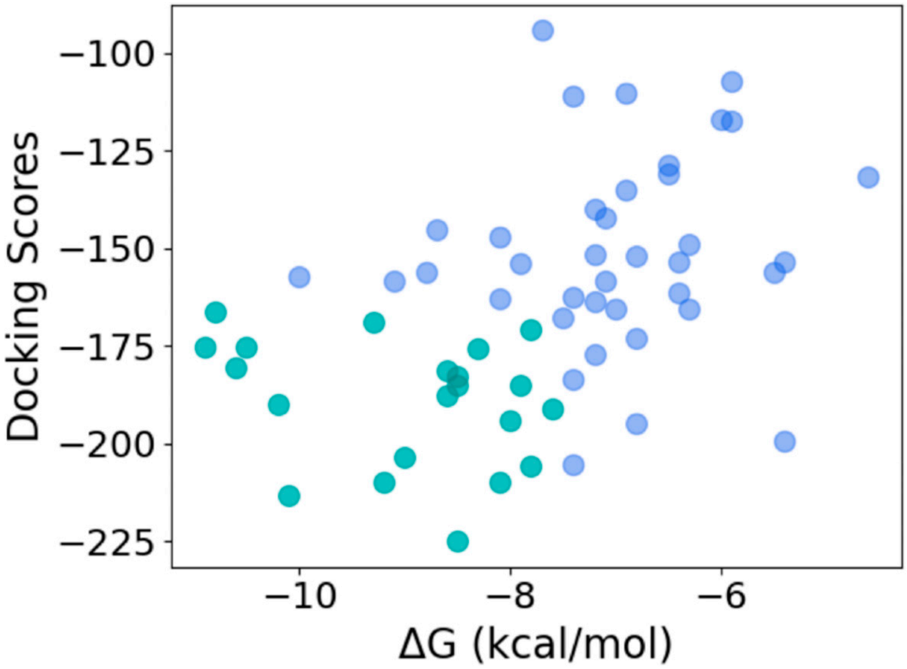

Figure 8. Selection of most efficacious hits at the intersection of the top 50 percentile of docking scores from HPEPDOCK and binding energy obtained from CABS-dock-derived structures (top 50 and bottom 50 percentiles shown in green and blue, respectively).

The 21 peptides were 11 to 14 amino acids long (Table 3). The most and least commonly occurring amino acids in the 21 peptides were E, G, D, P, S, H, M, and W, which were also similarly represented in the positive set (Figure S6). Except for ‘FE’, the most occurring dipeptides in the hits were different from the positive set. The charge and polarity distributions were similar to the positive set. Although the average pI of the hits (4.20) was comparable to that of the positive set (4.24), a few peptides (T46, T49, and T55) have a pI of more than 8.5. Next, we ran a BLASTp search on these hits and, based on the percentage sequence identity, we selected the top-ranked matching protein and the source organism. While 50% of the hits matched protein sequences from bacteria, the rest were derived from plants, humans, primates, viruses, fishes, birds, and reptiles. Further, the hit peptides were fragments of enzymes, structural proteins, ribosomal proteins, and synthetic constructs, indicating the diversity of this set.

Table 3. Binding characteristics of peptide hits.

Peptide	Sequence	Source	K_D (nM)	Docking Scores	Binding Residues	Binding Sites
T49	QGNRKTTKEGSNDL	*Homo sapiens* (cytokine-dependent hematopoietic cell linker isoform X1)	9	−175.365	S20, D21, A22, E23, I24, G25, M26, P28, K70, H71, E80, D116, Y117, I118, Y134, K135, R137, V158, N159, E185, K202, S203, P204, N205, R206, W207	Exosite 1

Table 3. *Cont.*

Peptide	Sequence	Source	K_D (nM)	Docking Scores	Binding Residues	Binding Sites
T34	EYEEVEASPEKET	*Meleagris gallopavo* (tubulin beta-1 chain)	12	−166.309	I47, S48, W51, K87, Y89, I90, H91, P92, R93, L105, K107, K109, K110, P111, V112, C122, L123, R126, E127, F232, K236, K240	Exosite 2
T45	SGEGSFQPSQQNPQ	*Triticum aestivum* (gliadin peptide)	16	−180.56	F34, R35, K36, S37, P38, Q38A, E40, R67, K70, H71, R73, T74, R75, Y76, E77, R77a, N78, I79, W141, N143, L144, Q151, P152, S153	Active site and Exosite 1
T56	ARATAETDATANRG	*Mycobacterium tuberculosis* (prophage protein)	20	−175.107	E23, I24, K36, P38, K70, H71, S72, R73, T74, R75, E77, E80, S153, V154	Exosite 1
T52	EPTTEDLYFQSDND	M13 helper phage (pIII)	31	−189.841	N98, D100, R101, T147, R173, R175, T177, E217, R232	Active site and Exosite 2
T33	IYRFEPSKFIGE	*Nymphaea colorata* (unnamed protein)	37	−213.129	E39, L40, R93, E98, N143, S171, R173, I174, R175, I176, E192, E217, A221D	Active site
T39	ACENEDFEGIPGEA	*Homo sapiens* (hirugen, synthetic construct)	150	−168.785	S20, D21, A22, E23, I24, G25, M26, P28, W29, I68, G69, K70, I79, E80, K81, A113, F114, S115, D116, K135, G149a, K149b, V157, V158, N159, E184a, G186c, K186c, K202, S203, P204, R204a	Exosite 1
T57	FEFEFEPGGGRGDS	*Spirochaetales bacterium* (SpoIIE family protein phosphatase)	170	−209.705	K36, S37, P38, Q39, E40, W60a, S72, R73, T74, R75, Y76, R97, E98, N99, N143, L144, K145, W147, T148, Q151, S153, C191, D221a	Active site and Exosite 1
T54	RYEVRAELPGVDPD	*Mycobacterium tuberculosis* (erythromycin esterase)	240	−203.537	E23, M32, R35, K70, H71, R73, T74, R75, Y76, V154, Q156	Exosite 1
T27	VQIYEEARKFS	*Potamochoerus porcus* (DEAD-box protein 3)	480	−187.722	H91, P92, R93, Y94, L99, D100, R101, D125, I176, T177, N179, H230, V231, F232, R233, L234, W237, I238, I242, D243	Exosite 2
T44	GNTRTAESGDEDFF	*Eubacteriales bacterium* (transglycosylase domain-containing protein)	530	−181.246	R35, P38, Q39, E40, R67, K70, H71, S72, R73, T74, Y76, E77, R78, N79, E80, G142, N143, L144, Q151, P152, S153, V154, E192	Active site and Exosite 1

Table 3. *Cont.*

Peptide	Sequence	Source	K_D (nM)	Docking Scores	Binding Residues	Binding Sites
T55	NRLVQNPPKKFSGE	*Burkholderia* sp. Bp9140 (hypothetical protein)	610	−224.973	S20, D21, A22, E23, Q39, V67, H71, S72, T74, Y76, S116, Y117, K135, W141, A149V, S153, V154, L155, V157, N158, E185, R187, K202, S203, P204, R206	Exosite 1
T31	AEYETVQNSFNQ	*Cellvibrio fibrivorans* (cellulase family glycosyl hydrolase)	630	−185.041	P92, R93, W96, N98, L99, D100, R101, D102, I103, R126, A129A, B129S, Q131, E164, R175, I176, T177, D178, N179, H230, F232, R233, K236, Q244	Exosite 2
T46	SSGSVGESSSKGPR	*Pan pansicus* (cytokeratin-10)	630	−182.793	E23, I24, G25, F34, R35, S37, P38, Q39, E40, L41, D60W, K70, H71, S72, R73, N79, E98, N99, D116, N143, L144, K145, P152, S153, L155, Q156, E192, W215, G216, E217	Active site and Exosite 1
T40	VQGSDQSDSANVQR	*Hoeflea* sp. (UDP N-acetylmuramate L-alanine ligase)	770	−175.557	I23, F34, R35, K36, S37, P38, Q39, E40, L42, R73, W140, N143, L144, E146, C149V, P152, S153, E192, E216, G218, C219, D220, R221	Active site
T41	NDDEDPKSHRDPSN	FGF-4 synthetic construct	1200	−209.886	I24, G25, Q30, K70, R78, N79, I80, E80, K81, I82, K107, L108, K109, K110, P111, V112, F114, Y117, I118, H119	Exosite 1
T32	GEKPDEFESGSP	*Poecilia Mexicana* (ribosomal protein S7)	1300	−194.088	R101, R126, T128, A129A, S130, L132, Q133, E164, R165, K169, D178, N179, M180, S203, P204, F205, H230, R233, K236	Exosite 2
T42	RGNNDIGSGFNDDP	*Cellulomonas soli* (glycosyl transferase)	1600	−185.149	N95, W96, E97, N98, L99, D100, V163, P166, K169, D170, S171, T172, I174, R175, I176, Y184a, K184b, E185, E217, R221b, D222, G223, K224	Exosite 2
T48	GIGPKFQHSGGEPP	*Mycobacterium tuberculosis* (prophage protein)	1800	−205.784	I90, H91, R93, Y94, N95, N99, L100, D100, R173, I174, R175, I176, F227, V241, I242, F245, E246	Exosite 2

Table 3. *Cont.*

Peptide	Sequence	Source	K_D (nM)	Docking Scores	Binding Residues	Binding Sites
T29	MEEGPSDPGSRS	*Mogibacterium* sp. (haloacid dehalogenase-like hydrolase)	1800	−170.831	I24, G25, Q30, K70, R78, N79, I80, E80, K81, I82, K107, L108, K109, K110, P111, V112, F114, Y117, I118, H119	Exosite 2
T43	HGEGTFTSDLSKQM	*Heloderma suspectum* (exendin 4 venom)	2500	−190.877	E23, I24, G25, M26, E39, K70, H71, E77, R77a, N78, I79, D116, I118, H119, N143, S153, V154, Q156	Exosite 1

To obtain the binding sites of the peptides on thrombin, we analyzed the top 10 poses of the 21 thrombin–peptide complexes obtained from CABS-dock using PyMol. We first benchmarked our approach by finding the binding sites on thrombin for positive peptides, hirugen, and avathrin. Consistent with the published reports, our model predicted that these two peptides bind to exosite I [39] and to active site + exosite I [40], respectively (Figure S7).

The docking analysis revealed that the seven peptides bind only to exosite I or to exosite II, two peptides bind only to the active site, and six peptides bind to both the active site and exosite I or exosite II, thus indicating different mechanisms of inhibition of the hits (Table 3). Figure 9 shows the complexes of representative hits with thrombin along with their binding sites.

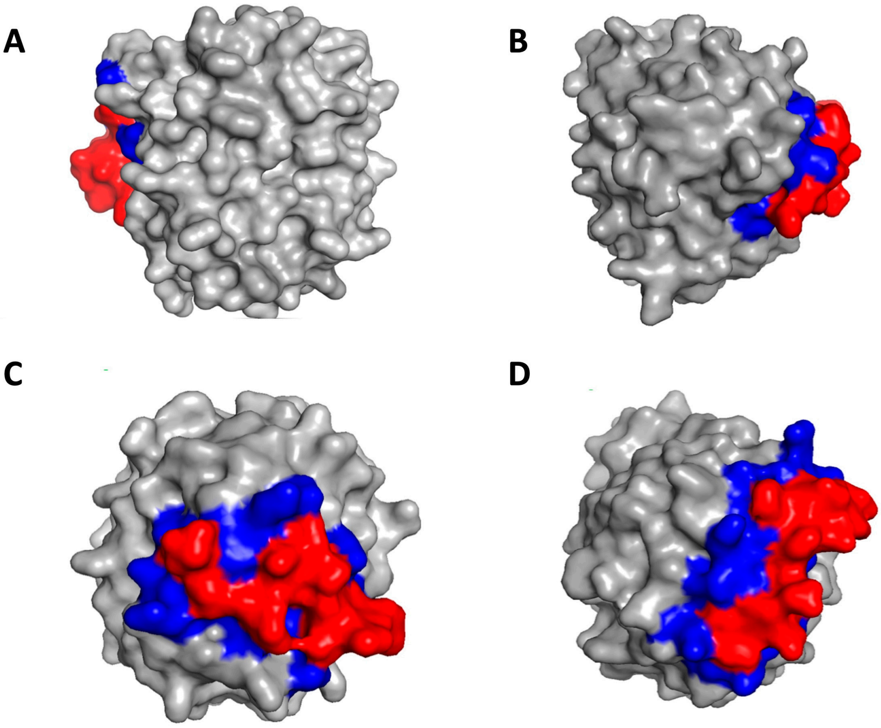

Figure 9. Identification of binding sites on thrombin by peptide hits. (**A**) Peptide T29 binds to exosite II; (**B**) Peptide T43 binds to exosite I; (**C**) Peptide T40 binds to active site; (**D**) Peptide T45 binds to exosite I and active site. The red color denotes the peptide and the blue color denotes the binding residues on thrombin.

Next, inspired by the design of bivalent inhibitors such as bivalirudin, we sought to use the peptide hits to design new bivalent inhibitors. To exemplify this approach, we combined peptides that interact with different binding sites on thrombin with varying levels of strength; peptide hits T40 and T33 (both active site binders) were combined with T39, T55 (both exosite I binders), T29, or T27 (both exosite II binders), either at N- or at C-terminal. The bivalent peptides were docked with thrombin and the binding energy was computed. As shown in Figure S8, combining T33 with T39, T55, T29, or T27 resulted in an incremental change in binding energy compared to a single peptide binding to thrombin, and so did combining T40 with T39, T55, or T29. In contrast, as shown in Figure 10A, combining T40 with T27 resulted in a significant change in binding energy for the bivalent peptide compared to either T40 or T27 alone, and the change depends on the N-C concatenation. The T40–T27 concatenation resulted in a significant increase in binding energy, while the T27–T40 concatenation resulted in a significant decrease in binding energy. The docking analysis showed that the favorable binding of the T27–T40 bivalent peptide could be due to a higher binding interaction of this bivalent peptide around the exosite II of thrombin. This 'wrap-around' site is possible because of the flexible structure of the T27–T40 peptide, comprising two short helices connected by a flexible loop (Figure 10B). On the other hand, the T40–T27 is a single long helix, the rigidity of which reduces the ability of the bivalent peptide to bind thrombin. For the bivalent peptides T27–T40 and T40–T27, the K_D predicted by PRODIGY are 0.034 nM and 11 nM, and the K_I predicted by our regression model are 0.34 nM and 917 nM, respectively. These values not only confirm the differences in the efficacies of the peptides, but also the qualitative agreement between the mechanism-agnostic ML model and the structure-dependent docking models.

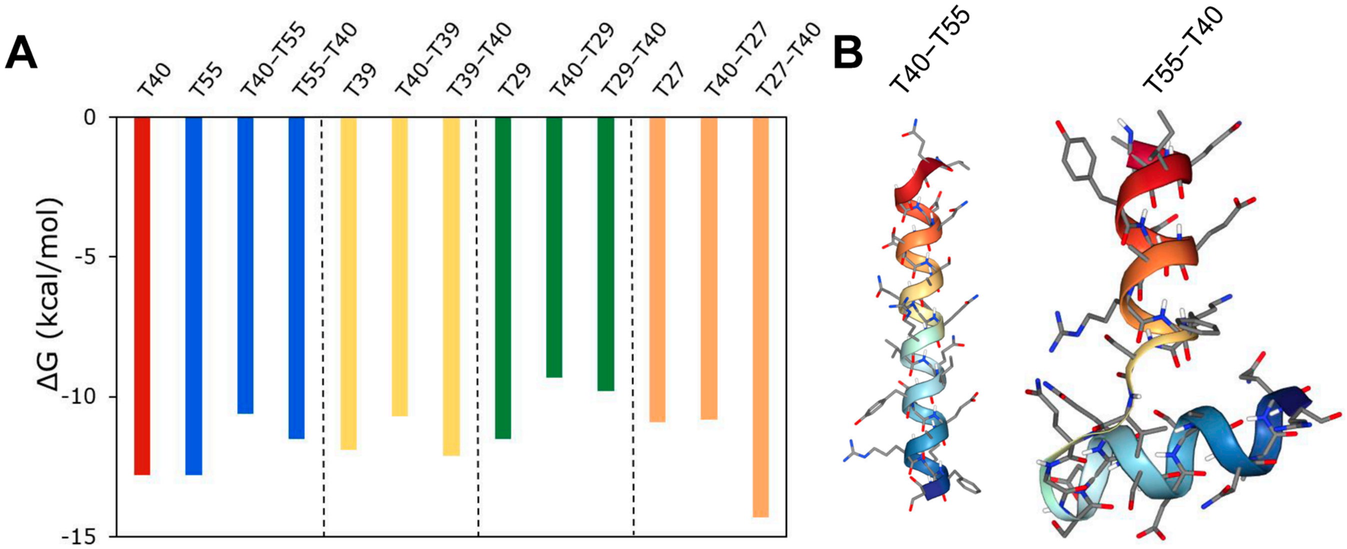

Figure 10. Design of a bivalent peptide. (**A**) Free energy change in binding of thrombin to bivalent peptide composed of T40 peptide combined with either T55, T39, or T27 peptides in either N-C or C-N concatenation. (**B**) 3D conformation of T40–T27 and T27–T40 peptides.

4. Discussion

In this work, using a machine learning model pipeline, we have identified features that characterize thrombin-inhibiting activity in peptide sequences and established relationships between these features and their potential for thrombin inhibition. We have discovered, from diverse sources, new peptides with varying levels of antithrombotic activity and varying degrees of sequence homology with known peptide sequences.

This is the first time, to our knowledge, a comprehensive evaluation of known thrombin-inhibiting peptides has been performed. We collected all the available naturally occurring thrombin-inhibiting peptides along with their inhibition constants starting with classical hirudin and other peptides published since 1976 [41]. Our sequence alignment analysis shows that the known, naturally occurring thrombin inhibitor sequences are indeed structurally diverse, with inhibition constants ranging over eight orders of magnitude. This analysis also demonstrated that the varying degrees of effectiveness of antithrombotic peptides depends on the strength of interactions with allosteric or active site interactions or both. Although most known antithrombin inhibitors seem to have originated from hematophagous organisms, this class includes an estimated 15,000 species of arthropods and a large number of leeches and hookworms, and their blood-feeding behavior has evolved independently over six times, suggesting that there are likely to be many more undiscovered peptides [42].

ML-based methods provide much deeper insights into structure-activity relationships than sequence alignment methods. The ML algorithms search higher dimensional spaces for non-physical pattern matching while sequence alignment methods are centered on properties of individual amino acids. Features extracted using ML models corroborated with previous results based on sequence alignment analysis about the characteristics of peptides that are considered important for thrombin binding, namely the presence of negatively charged amino acids (D, E), lower isoelectric point, and hydrophobicity. Interestingly, the most common dipeptide in the positive set 'EE' occurred only once in 3 out of the 21 hits, and D/E-containing dipeptides were present in only 8 out of 21 hits, suggesting that thrombin inhibitory activity may also be determined by features defined by the primary sequence. Performing a regression analysis provided additional insights into amino acids and sub-sequences that determine the affinity of the peptide for thrombin. To obtain further insights, we compared the distribution of 120 features in the 21 hits and the 88 positive peptides that were selected by the final classification model (Table 1, SVC-Linear with RFE). In Table S7, the features are listed along with their relative importance in the model (i.e., weights), and the *p*-value of the distribution of the features between the hit and the positive sets. We found that 87 out of 120 features of the hit peptides were distributed similarly to those of the positive peptides ($p > 0.05$). This observation suggests that while the similarity between the two sets in some of the features is obvious (such as isoelectric point, % of M, and FE), some others are subtler (such as the charge transitions 'ChargeT13′ and 'ChargeT23′). Still, others may be unique to hits as these features or dissimilar to those of the positive set (such as EY and PE). This information may be used to design novel peptides that are distinct from the known positive peptides.

This is also the first time, to our knowledge, machine learning models have been applied to the discovery and evaluation of novel thrombin-inhibiting peptide sequences. Previous approaches have focused on the evaluation of small molecule inhibitors based on molecular structures derived from crystallographic information and computationally heavy protein-structure predictions [43–47]. Although structure-based models tend to be more accurate, they also require enormously more, often mechanistic, information compared to ML-based models, which follow an agnostic approach for rapid, high-throughput screening of enormous datasets [48]. The efficiency of machine learning models made it easy to search a vast biological space, resulting in a low hit rate of ~1500 hits per million peptides. Upon screening more than 10 million peptides from various peptide databases for thrombin inhibition activity, our model identified sequences that included both peptides with previously known bioactivity unrelated to thrombin and peptides that do not have any known bioactivity. We identified thrombin-inhibiting activity in antimicrobial, antiviral, antifungal, and anti-inflammatory peptides. Therefore, these peptides may be used for drug repurposing for their anticoagulant properties or may serve to probe the cross-acting role of thrombin in infection or inflammation.

As with most first-of-its-kind studies, this work is not without limitations: (1) We were able to find only 88 unique thrombin-inhibiting peptides, and only 53 with inhibition constants. Further, these inhibition constants were curated from the data generated in different labs with unavoidable variations in the assay conditions. Given that structurally complex peptides of up to 100 amino acids can be reliably and rapidly synthesized on a large scale, we propose that high-throughput experimentation will generate larger databases and improve the model predictions. (2) The hits were tested for sensitivity to thrombin but not for specificity. Therefore, some of the hits may be active against other serine proteases involved in clotting dynamics, including FXa or plasmin, and further analyses for specificity are essential. (3) Both the ML model and the docking model do not classify the type of inhibition (such as reversible/irreversible), although this may be included in the model if additional information is available. (4) The mechanisms of inhibition and interaction sites on thrombin for the peptide were based on top poses that were computed by the model in the absence of any ions or pH changes, which may be critical in a physiological milieu. (5) Experimental validation of these putative hits using in vitro enzyme inhibition assays and in vivo animal models should be performed to confirm the predictive power of the model [49].

Direct Thrombin Inhibitors (DTIs) have a pharmacological advantage over indirect thrombin inhibitors because of their ability to bind both circulating and fibrin-bound thrombin, better efficacy, predictable pharmacokinetics, and fewer off-target effects. Despite these advantages, their usage has been limited to certain indications because of issues with the lack of specific antidotes, bleeding, and clot destabilization. The peptide sequences identified in this work open up the possibility of discovering new DTIs with tunable affinities. Further, the discovery of thrombin inhibitory potential in peptides with known bioactivity, such as antimicrobial, anticancer, and anti-inflammatory, opens up the possibility of drug repurposing for co-morbidity due to thrombotic complications. In vitro inhibition assays of the peptide hits will provide lead candidates with sufficient specificity and sensitivity, which may be advanced to testing in animal models of arterial or venous thrombosis. Lastly, the classification–regression staged model pipeline developed in this work may be readily applied for the discovery of peptides targeting other coagulation proteases such as FXa, FXIa, or plasmin, provided peptide inhibitors and their K_I values are available. Our approach can reduce the turnaround time in drug discovery and provide better quality hits.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bioengineering10111300/s1; Figure S1: Thrombin geography (1PPB) in standard orientation.; Figure S2: Agglomerative clustering of peptides in the positive dataset.; Figure S3: Highly correlated (>80%) features in the positive dataset.; Figure S4: Decision function of peptides in the positive and negative datasets.; Figure S5: Agglomerative clustering of peptide hits.; Figure S6: Characteristics of peptide hits.; Figure S7: Confirmation of binding sites of two positive peptides on thrombin using CABS DOCK.; Figure S8: Design of a bivalent peptide.; Table S1: Positive peptides.; Table S2: Composition-Transition-Distribution (CTD) descriptor groups.; Table S3: Hyperparameters of classification models obtained using RandomSearchCV.; Table S4: Hyperparameters of RFE/SFS models obtained using RandomSearchCV.; Table S5: Prediction of inhibition constants for peptides in positive dataset.; Table S6: Short peptide hits from the machine learning model pipeline.; Table S7: Features selected by the SVM-Linear with RFE model, and their relative importance. References [50–103] are cited in Table S1 in Supplementary Materials

Author Contributions: Research design: N.B., R.K., D.C.A., and A.K.R.; Data curation: N.B. and A.K.R.; Model development: N.B., P.V.P., R.K., T.D., D.C.A., and A.K.R.; Docking analysis: N.B., P.K., and A.K.R.; Data analysis: N.B., P.V.P., D.C.A., and A.K.R.; Manuscript preparation: N.B., D.C.A., and A.K.R.; Funding acquisition: A.K.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the College of Engineering, San José State University.

Data Availability Statement: The data presented in this study are available in the Supplementary Materials Tables.

Acknowledgments: Assistance with establishing the webserver from Prasad Jayakumar is gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Marcum, J.A. Defending the priority of "remarkable researches": The discovery of fibrin ferment. *Hist. Philos. Life Sci.* **1998**, *20*, 51–76. [PubMed]
2. Remiker, A.S.; Palumbo, J.S. Mechanisms coupling thrombin to metastasis and tumorigenesis. *Thromb. Res.* **2018**, *164*, S29–S33. [CrossRef]
3. Aliter, K.F.; Al-Horani, R.A. Thrombin Inhibition by Argatroban: Potential Therapeutic Benefits in COVID-19. *Cardiovasc. Drugs Ther.* **2021**, *35*, 195–203. [CrossRef] [PubMed]
4. Lane, D.A.; Philippou, H.; Huntington, J.A. Directing thrombin. *Blood* **2005**, *106*, 2605–2612. [CrossRef] [PubMed]
5. Mann, K.G. Thrombin formation. *Chest* **2003**, *124* (Suppl. S3), 4S–10S. [CrossRef]
6. Gustafsson, D.; Bylund, R.; Antonsson, T.; Nilsson, I.; Nyström, J.E.; Eriksson, U.; Bredberg, U.; Teger-Nilsson, A.C. A new oral anticoagulant: The 50-year challenge. *Nat. Rev. Drug Discov.* **2004**, *3*, 649–659. [CrossRef]
7. Di Nisio, M.; Middeldorp, S.; Büller, H.R. Direct Thrombin Inhibitors. *N. Engl. J. Med.* **2005**, *353*, 1028–1040. [CrossRef] [PubMed]
8. Chan, N.; Sobieraj-Teague, M.; Eikelboom, J.W. Direct oral anticoagulants: Evidence and unresolved issues. *Lancet* **2020**, *396*, 1767–1776. [CrossRef]
9. Montinari, M.R.; Minelli, S. From ancient leech to direct thrombin inhibitors and beyond: New from old. *Biomed. Pharmacother.* **2022**, *149*, 112878. [CrossRef]
10. Soares, T.A.; Nunes-Alves, A.; Mazzolari, A.; Ruggiu, F.; Wei, G.-W.; Merz, K. The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) studies propelled by the surge of machine learning methods. *J. Chem. Inf. Model.* **2022**, *62*, 5317–5320. [CrossRef] [PubMed]
11. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477. [CrossRef] [PubMed]
12. Bian, Y.; Xie, X.-Q. Generative chemistry: Drug discovery with deep learning generative models. *J. Mol. Model.* **2021**, *27*, 71. [CrossRef] [PubMed]
13. Ye, J.; Li, A.; Zheng, H.; Yang, B.; Lu, Y. Machine learning advances in predicting peptide/protein-protein interactions based on sequence information for lead peptides discovery. *Adv. Biol.* **2023**, *7*, 2200232. [CrossRef] [PubMed]
14. Syrlybaeva, R.; Strauch, E.M. Deep learning of protein sequence design of protein–protein interactions. *Bioinformatics* **2023**, *39*, btac733. [CrossRef] [PubMed]
15. Chandra, A.; Tünnermann, L.; Löfstedt, T.; Gratz, R. Transformer-based deep learning for predicting protein properties in the life sciences. *Elife* **2023**, *12*, e82819. [CrossRef]
16. Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390. [CrossRef]
17. Xiao, X.; Wang, P.; Lin, W.-Z.; Jia, J.-H.; Chou, K.-C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **2013**, *436*, 168–177. [CrossRef]
18. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121–77136. [CrossRef]
19. Bose, B.; Downey, T.; Ramasubramanian, A.K.; Anastasiu, D.C. Identification of distinct characteristics of antibiofilm peptides and prospection of diverse sources for efficacious sequences. *Front. Microbiol.* **2022**, *12*, 783284. [CrossRef]
20. Kumar, R.; Chaudhary, K.; Singh Chauhan, J.; Nagpal, G.; Kumar, R.; Sharma, M.; Raghava, G.P. An *in-silico* platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep.* **2015**, *5*, 12512. [CrossRef]
21. Lee, E.Y.; Fulan, B.M.; Wong, G.C.L.; Ferguson, A.L. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13588–13593. [CrossRef] [PubMed]
22. Lakshmaiah Narayana, J.; Mishra, B.; Lushnikova, T.; Wu, Q.; Chhonker, Y.S.; Zhang, Y.; Zarena, D.; Salnikov, E.S.; Dang, X.; Wang, F.; et al. Two distinct amphipathic peptide antibiotics with systemic efficacy. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 19446–19454. [CrossRef] [PubMed]
23. Das, P.; Sercu, T.; Wadhawan, K.; Padhi, I.; Gehrmann, S.; Cipcigan, F.; Chenthamarakshan, V.; Strobelt, H.; dos Santos, C.; Chen, P.-Y.; et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **2021**, *5*, 613–623. [CrossRef]
24. Bateman, A.; Martin, M.J.; Orchard, S.; Magrane, M. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
25. Agarwala, R.; Barret, T.; Beck, J.; Benson, D.A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2018**, *46*, D8–D13. [CrossRef]
26. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
27. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395. [CrossRef] [PubMed]
28. Chapman, B.; Chang, J. Biopython: Python Tools for Computation Biology. 2000. Available online: http://www.bris.ac.uk/Depts/Chemistry/MOTM/ (accessed on 3 November 2023).

29. Xiao, N.; Cao, D.S.; Zhu, M.F.; Xu, Q.S. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. In *Bioinformatics*; Oxford University Press: Oxford, UK, 2015; pp. 1857–1859. [CrossRef]
30. Govindan, G.; Nair, A.S. Composition, Transition and Distribution (CTD)—A dynamic feature for predictions based on hierarchical structure of cellular sorting. In Proceedings of the 2011 Annual IEEE India Conference, Hyderabad, India, 16–18 December 2011; pp. 1–6. [CrossRef]
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Duchesnay, É. Scikit-learn: Machine learning in Python. *J. Machine Learn. Res.* **2011**, *12*, 2825–2830.
32. Randriamihamison, N.; Vialaneix, N.; Neuvial, P. Applicability and interpretability of ward's hierarchical agglomerative clustering with or without contiguity constraints. *J. Classif.* **2021**, *38*, 363–389. [CrossRef]
33. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
34. Zhou, P.; Jin, B.; Li, H.; Huang, S.Y. HPEPDOCK: A web server for blind peptide-protein docking based on a hierarchical algorithm. *Nucleic Acids Res.* **2018**, *46*, W443–W450. [CrossRef] [PubMed]
35. Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res.* **2015**, *43*, W419–W424. [CrossRef] [PubMed]
36. Xue, L.C.; Rodrigues, J.P.; Kastritis, P.L.; Bonvin, A.M.; Vangone, A. PRODIGY: A web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* **2016**, *32*, 3676–3678. [CrossRef] [PubMed]
37. Huntington, J.A. Molecular recognition mechanisms of thrombin. *J. Thromb. Haemost.* **2005**, *3*, 1861–1872. [CrossRef] [PubMed]
38. Di Cera, E. Thrombin. *Mol. Asp. Med.* **2008**, *29*, 203–254. [CrossRef] [PubMed]
39. Krishnaswamy, S. Exosite-driven substrate specificity and function in coagulation. *J. Thromb. Haemost.* **2005**, *3*, 54–67. [CrossRef]
40. Iyer, J.K.; Koh, C.Y.; Kazimirova, M.; Roller, L.; Jobichen, C.; Swaminathan, K.; Mizuguchi, J.; Iwanaga, S.; Nuttall, P.A.; Chan, M.Y.; et al. Avathrin: A novel thrombin inhibitor derived from a multicopy precursor in the salivary glands of the ixodid tick, *Amblyomma variegatum*. *FASEB J.* **2017**, *31*, 2981–2995. [CrossRef] [PubMed]
41. Peeters, H. *Protides of the Biological Fluids*; Elsevier: Amsterdam, The Netherlands, 1975. [CrossRef]
42. Ribeiro, J.M. Blood-feeding arthropods: Live syringes or invertebrate pharmacologists? *Infect Agents Dis.* **1995**, *4*, 143–152. [PubMed]
43. Myles, T.; Church, F.C.; Whinna, H.C.; Monard, D.; Stone, S.R. Role of thrombin anion-binding exosite-I in the formation of thrombin-serpin complexes. *J. Biol. Chem.* **1998**, *273*, 31203–31208. [CrossRef]
44. Mans, B.J.; Louw, A.I.; Neitz, A.W.H. Amino acid sequence and structure modeling of savignin, a thrombin inhibitor from the tick, *Ornithodoros savignyi*. *Insect Biochem. Mol. Biol.* **2002**, *32*, 821–828. [CrossRef] [PubMed]
45. Howard, N.; Abell, C.; Blakemore, W.; Chessari, G.; Congreve, M.; Howard, S.; Jhoti, H.; Murray, C.W.; Seavers, L.C.; van Montfort, R.L. Application of fragment screening and fragment linking to the discovery of novel thrombin inhibitors. *J. Med. Chem.* **2006**, *49*, 1346–1355. [CrossRef] [PubMed]
46. Jacobson, M.; Sali, A. Comparative protein structure modeling and its applications to drug discovery. *Annu. Rep. Med. Chem.* **2004**, *39*, 259–276. [CrossRef]
47. Böhm, H.-J.; Stahl, M. Structure-based library design: Molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* **2000**, *4*, 283–286. [CrossRef] [PubMed]
48. Giguère, S.; Laviolette, F.; Marchand, M.; Tremblay, D.; Moineau, S.; Liang, X.; Biron, É.; Corbeil, J. Machine learning assisted design of highly active peptides for drug discovery. *PLoS Comp. Biol.* **2015**, *11*, e1004074. [CrossRef] [PubMed]
49. Koh, C.Y.; Shih, N.; Yip, C.Y.C.; Li, A.W.L.; Chen, W.; Amran, F.S.; Leong, E.J.E.; Iyer, J.K.; Croft, G.; Mazlan, M.I.B.; et al. Efficacy and safety of next-genertion tick transcriptome-derived direct thrombin inhibitors. *Nat. Commun.* **2021**, *12*, 6912. [CrossRef] [PubMed]
50. Kelly, A.B.; Maraganore, J.M.; Bourdon, P.; Hanson, S.R.; Harker, L.A. Antithrombotic effects of synthetic peptides targeting various functional domains of thrombin. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 6040–6044. [CrossRef]
51. Hasan, A.A.; Warnock, M.; Nieman, M.; Srikanth, S.; Mahdi, F.; Krishnan, R.; Tulinsky, A.; Schmaier, A.H. Mechanisms of Arg-Pro-Pro-Gly-Phe inhibition of thrombin. *Amer. J. Physiol. Heart. Circ. Physiol.* **2003**, *285*, H183–H193. [CrossRef]
52. Cheng, S.; Tu, M.; Liu, H.; An, Y.; Du, M.; Zhu, B. A novel heptapeptide derived from *Crassostrea gigas* shows anticoagulant activity by targeting for thrombin active domain. *Food Chem.* **2021**, *334*, 127507. [CrossRef]
53. Chen, F.; Jiang, H.; Lu, Y.; Chen, W.; Huang, G. Identification and in silico analysis of antithrombotic peptides from the enzymatic hydrolysates of *Tenebrio molitor* larvae. *Eur. Food Res. Technol.* **2019**, *245*, 2687–2695. [CrossRef]
54. Kazimtrova, M.; Kini, R.M.; Koh, C.Y. Thrombin Inhibitor. U.S. Patent 9217027, 2016.
55. Liu, H.; Tu, M.; Cheng, S.; Xu, Z.; Xu, X.; Du, M. Anticoagulant decapeptide interacts with thrombin at the active site and exosite-I. *J. Agric. Food Chem.* **2020**, *68*, 176–184. [CrossRef]
56. Cheng, S.; Wu, D.; Liu, H.; Xu, X.; Zhu, B.; Du, M. A novel anticoagulant peptide discovered from *Crassostrea gigas* by combining bioinformatics with the enzymolysis strategy: Inhibitory kinetics and mechanisms. *Food Funct.* **2021**, *12*, 10136–10146. [CrossRef] [PubMed]
57. Naski, M.C.; Fenton, J.W.; Maraganore, J.M.; Olson, S.T.; Shafer, J.A. The COOH-terminal domain of hirudin. An exosite-directed competitive inhibitor of the action of alpha-thrombin on fibrinogen. *J. Biol. Chem.* **1990**, *265*, 13484–13489. [CrossRef]

58. Feng, L.; Tu, M.; Qiao, M.; Fan, F.; Chen, H.; Song, W.; Du, M. Thrombin inhibitory peptides derived from *Mytilus edulis* proteins: Identification, molecular docking and in silico prediction of toxicity. *Eur. Food Res. Technol.* **2018**, *244*, 207–217. [CrossRef]
59. Mosesson, M.W.; Meh, D.A. Thrombin Inhibitor. U.S. Patent 5985833, 2000.
60. Stubbs, M.T.; Oschkinat, H.; Mayr, I.; Huber, R.; Angliker, H.; Stone, S.R.; Bode, W. The interaction of thrombin with fibrinogen. A structural basis for its specificity. *Eur. J. Biochem.* **1992**, *206*, 187–195. [CrossRef] [PubMed]
61. Scharf, M.; Engels, J.; Tripier, D. Primary structures of new iso-hirudins. *FEBS Lett.* **1989**, *255*, 105–110. [CrossRef] [PubMed]
62. Maraganore, J.M.; Bourdon, P.; Jablonski, J.; Ramachandran, K.L.; Fenton, J.W. Design and characterization of hirulogs: A novel class of bivalent peptide inhibitors of thrombin. *Biochemistry* **1990**, *29*, 7095–7101. [CrossRef]
63. Ni, F.; Tolkatchev, D.; Natapova, A.; Koutychenko, A. Peptide Inhibitors of Thrombin as Potent Anticoagulants. U.S. Patent US7456152B2, 2008.
64. Figueiredo, A.C.; de Sanctis, D.; Gutiérrez-Gallego, R.; Cereija, T.B.; Macedo-Ribeiro, S.; Fuentes-Prior, P.; Pereira, P.J. Unique thrombin inhibition mechanism by anophelin, an anticoagulant from the malaria vector. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E3649–E3658. [CrossRef]
65. Cappello, M.; Li, S.; Chen, X.; Li, C.B.; Harrison, L.; Narashimhan, S.; Beard, C.B.; Aksoy, S. Tsetse thrombin inhibitor: Bloodmeal-induced expression of an anticoagulant in salivary glands and gut tissue of *Glossina morsitans morsitans*. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14290–14295. [CrossRef]
66. Koh, C.Y.; Kazimirova, M.; Trimnell, A.; Takac, P.; Labuda, M.; Nuttall, P.A.; Kini, R.M. Variegin, a novel fast and tight binding thrombin inhibitor from the tropical bont tick. *J. Biol. Chem.* **2007**, *282*, 29101–29113. [CrossRef]
67. Giri, P.K.; Tang, X.; Thangamani, S.; Shenoy, R.T.; Ding, J.L.; Swaminathan, K.; Sivaraman, J. Modifying the substrate specificity of *Carcinoscorpius rotundicauda* serine protease inhibitor domain 1 to target thrombin. *PLoS ONE* **2010**, *5*, e15258. [CrossRef]
68. Sarmientos, P.; Poet, P.D.T.D.; Nitti, G.; Scacheri, E. Antithrombin Polypeptides. U.S. Patent US5439820A, 1995.
69. Hong, S.J.; Kang, K.W. Purification of granulin-like polypeptide from the blood-sucking leech, *Hirudo nipponia*. *Protein Expr. Purif.* **1999**, *16*, 340–346. [CrossRef] [PubMed]
70. Yang, X.; Wang, Y.; Lu, Z.; Zhai, L.; Jiang, J.; Liu, J.; Yu, H. A novel serine protease inhibitor from the venom of *Vespa bicolor Fabricius*. *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.* **2009**, *153*, 116–120. [CrossRef] [PubMed]
71. Schlott, B.; Wöhnert, J.; Icke, C.; Hartmann, M.; Ramachandran, R.; Gührs, K.H.; Glusa, E.; Flemming, J.; Görlach, M.; Grosse, F.; et al. Interaction of Kazal-type inhibitor domains with serine proteinases: Biochemical and structural studies. *J. Mol. Biol.* **2002**, *318*, 533–546. [CrossRef] [PubMed]
72. Strube, K.H.; Kröger, B.; Bialojan, S.; Otte, M.; Dodt, J. Isolation, sequence analysis, and cloning of haemadin. An anticoagulant peptide from the Indian leech. *J. Biol. Chem.* **1993**, *268*, 8590–8595. [CrossRef] [PubMed]
73. Brahma, R.K.; Blanchet, G.; Kaur, S.; Kini, R.M.; Doley, R. Expression and characterization of haemathrins, madanin-like thrombin inhibitors, isolated from the salivary gland of tick *Haemaphysalis bispinosa* (Acari: Ixodidae). *Thromb. Res.* **2017**, *152*, 20–29. [CrossRef]
74. Clayton, D.; Kulkarni, S.S.; Sayers, J.; Dowman, L.J.; Ripoll-Rozada, J.; Pereira, P.J.; Payne, R.J. Chemical synthesis of a haemathrin sulfoprotein library reveals enhanced thrombin inhibition following tyrosine sulfation. *RSC Chem. Biol.* **2020**, *1*, 379–384. [CrossRef]
75. Jablonka, W.; Kotsyfakis, M.; Mizurini, D.M.; Monteiro, R.Q.; Lukszo, J.; Drake, S.K.; Ribeiro, J.M.; Andersen, J.F. Identification and mechanistic analysis of a novel tick-derived inhibitor of thrombin. *PLoS ONE* **2015**, *10*, e0133991. [CrossRef]
76. Thompson, R.E.; Liu, X.; Ripoll-Rozada, J.; Alonso-García, N.; Parker, B.L.; Pereira, P.J.B.; Payne, R.J. Tyrosine sulfation modulates activity of tick-derived thrombin inhibitors. *Nature Chem.* **2017**, *9*, 909–917. [CrossRef] [PubMed]
77. Iwanaga, S.; Okada, M.; Isawa, H.; Morita, A.; Yuda, M.; Chinzei, Y. Identification and characterization of novel salivary thrombin inhibitors from the ixodidae tick, *Haemaphysalis longicornis*. *Eur. J. Biochem.* **2003**, *270*, 1926–1934. [CrossRef] [PubMed]
78. Krstenansky, J.L.; Owen, T.J.; Yates, M.T.; Mao, S.J.T. The C-terminal binding domain of hirullin P18. *FEBS Lett.* **1990**, *269*, 425–429. [CrossRef] [PubMed]
79. Steiner, V.; Knecht, R.; Börnsen, K.O.; Gassmann, E.; Stone, S.R.; Raschdorf, F.; Schlaeppi, J.M.; Maschler, R. Primary structure and function of novel O-glycosylated hirudins from the leech *Hirudinaria manillensis*. *Biochemistry* **1992**, *31*, 2294–2298. [CrossRef] [PubMed]
80. Scacheri, E.; Nitti, G.; Valsasina, B.; Orsini, G.; Visco, C.; Ferrera, M.; Sawyer, R.T.; Sarmientos, P. Novel hirudin variants from the leech *Hirudinaria manillensis*. Amino acid sequence, cDNA cloning and genomic organization. *Eur. J. Biochem.* **1993**, *214*, 295–304. [CrossRef]
81. Rydel, T.J.; Ravichandran, K.G.; Tulinsky, A.; Bode, W.; Huber, R.; Roitsch, C.; Fenton, J.W., 2nd. The structure of a complex of recombinant hirudin and human alpha-thrombin. *Science* **1990**, *249*, 277–280. [CrossRef] [PubMed]
82. Stone, S.R.; Hofsteenge, J. Kinetics of the inhibition of thrombin by hirudin. *Biochemistry* **1986**, *25*, 4622–4628. [CrossRef] [PubMed]
83. Warkentin, T.E. Bivalent direct thrombin inhibitors: Hirudin and bivalirudin. *Best Pract. Res. Clin. Haematol.* **2004**, *17*, 105–125. [CrossRef] [PubMed]
84. Watanabe, R.M.O.; Tanaka-Azevedo, A.M.; Araujo, M.S.; Juliano, M.A.; Tanaka, A.S. Characterization of thrombin inhibitory mechanism of rAaTI, a Kazal-type inhibitor from *Aedes aegypti* with anticoagulant activity. *Biochimie* **2011**, *93*, 618–623. [CrossRef] [PubMed]

85. Salzet, M.; Chopin, V.; Baert, J.; Matias, I.; Malecha, J. Theromin, a novel leech thrombin inhibitor. *J. Biol. Chem.* **2000**, *275*, 30774–30780. [CrossRef] [PubMed]
86. Cheng, B.; Liu, F.; Guo, Q.; Lu, Y.; Shi, H.; Ding, A.; Xu, C. Identification and characterization of hirudin-HN, a new thrombin inhibitor, from the salivary glands of *Hirudo nipponia*. *PeerJ* **2019**, *7*, e7716. [CrossRef] [PubMed]
87. Nakajima, C.; Imamura, S.; Konnai, S.; Yamada, S.; Nishikado, H.; Ohashi, K.; Onuma, M. A novel gene encoding a thrombin inhibitory protein in a cDNA library from *Haemaphysalis longicornis* salivary gland. *J. Vet. Med. Sci.* **2006**, *68*, 447–452. [CrossRef]
88. Zhang, D.; Cupp, M.S.; Cupp, E.W. Thrombostasin: Purification, molecular cloning and expression of a novel anti-thrombin protein from horn fly saliva. *Insect Biochem. Mol. Biol.* **2002**, *32*, 321–330. [CrossRef]
89. Pirone, L.; Ripoll-Rozada, J.; Leone, M.; Ronca, R.; Lombardo, F.; Fiorentino, G.; Andersen, J.F.; Pereira, P.J.; Arcà, B.; Pedone, E. Functional analyses yield detailed insight into the mechanism of thrombin inhibition by the antihemostatic salivary protein CE5 from *Anopheles gambiae*. *J. Biol. Chem.* **2017**, *292*, 12632–12642. [CrossRef] [PubMed]
90. Campos, I.T.; Amino, R.; Sampaio, C.A.; Auerswald, E.A.; Friedrich, T.; Lemaire, H.G.; Schenkman, S.; Tanaka, A.S. Infestin, a thrombin inhibitor presents in *Triatoma infestans* midgut, a Chagas' disease vector: Gene cloning, expression and characterization of the inhibitor. *Insect Biochem. Mol. Biol.* **2002**, *32*, 991–997. [CrossRef]
91. Friedrich, T.; Kröger, B.; Bialojan, S.; Lemaire, H.G.; Höffken, H.W.; Reuschenbach, P.; Otte, M.; Dodt, J. A Kazal-type inhibitor with thrombin specificity from *Rhodnius prolixus*. *J. Biol. Chem.* **1993**, *268*, 16216–16222. [CrossRef] [PubMed]
92. Mende, K.; Petoukhova, O.; Koulitchkova, V.; Schaub, G.A.; Lange, U.; Kaufmann, R.; Nowak, G. Dipetalogastin, a potent thrombin inhibitor from the blood-sucking insect *Dipetalogaster maximus* cDNA cloning, expression and characterization. *Eur. J. Biochem.* **1999**, *266*, 583–590. [CrossRef] [PubMed]
93. Nienaber, J.; Gaspar, A.R.M.; Neitz, A.W.H. Savignin, a potent thrombin inhibitor isolated from the salivary glands of the tick *Ornithodoros savignyi* (Acari: Argasidae). *Exp. Parasitol.* **1999**, *93*, 82–91. [CrossRef]
94. van de Locht, A.; Stubbs, M.T.; Bode, W.; Friedrich, T.; Bollschweiler, C.; Höffken, W.; Huber, R. The ornithodorin-thrombin crystal structure, a key to the TAP enigma? *EMBO J.* **1996**, *15*, 6011–6017. [CrossRef] [PubMed]
95. Liao, M.; Zhou, J.; Gong, H.; Boldbaatar, D.; Shirafuji, R.; Battur, B.; Nishikawa, Y.; Fujisaki, K. Hemalin, a thrombin inhibitor isolated from a midgut cDNA library from the hard tick *Haemaphysalis longicornis*. *J. Insect Physiol.* **2009**, *55*, 164–173. [CrossRef] [PubMed]
96. Oliveira-Carvalho, A.L.; Guimarães, P.R.; Abreu, P.A.; Dutra, D.L.S.; Junqueira-de-Azevedo, I.L.M.; Rodrigues, C.R.; Ho, P.L.; Castro, H.C.; Zingali, R.B. Identification and characterization of a new member of snake venom thrombin inhibitors from *Bothrops insularis* using a proteomic approach. *Toxicon* **2008**, *51*, 659–671. [CrossRef]
97. Macedo-Ribeiro, S.; Almeida, C.; Calisto, B.M.; Friedrich, T.; Mentele, R.; Stürzebecher, J.; Fuentes-Prior, P.; Barbosa Pereira, P.J. Isolation, cloning and structural characterization of Boophilin, a multifunctional kunitz-type proteinase inhibitor from the cattle tick. *PLoS ONE* **2008**, *3*, e1624. [CrossRef]
98. Mans, B.J.; Andersen, J.F.; Schwan, T.G.; Ribeiro, J.M.C. Characterization of anti-hemostatic factors in the argasid, Argas monolakensis: Implications for the evolution of blood-feeding in the soft tick family. *Insect Biochem. Mol. Biol.* **2008**, *38*, 22–41. [CrossRef]
99. Noeske-Jungblut, C.; Haendler, B.; Donner, P.; Alagon, A.; Possani, L.; Schleuning, W.-D. Triabin, a highly potent exosite inhibitor of thrombin. *J. Biol. Chem.* **1995**, *270*, 28629–28634. [CrossRef] [PubMed]
100. Lai, R.; Takeuchi, H.; Jonczy, J.; Rees, H.H.; Turner, P.C. A thrombin inhibitor from the ixodid tick, *Amblyomma hebraeum*. *Gene* **2004**, *342*, 243–249. [CrossRef] [PubMed]
101. Hengst, U.; Albrecht, H.; Hess, D.; Monard, D. The Phosphatidylethanolamine-binding protein is the prototype of a novel family of serine protease inhibitors. *J. Biol. Chem.* **2001**, *276*, 535–540. [CrossRef] [PubMed]
102. Wu, Y.; Eigenbrot, C.; Liang, W.-C.; Stawicki, S.; Shia, S.; Fan, B.; Ganesan, R.; Lipari, M.T.; Kirchhofer, D. Structural insight into distinct mechanisms of protease inhibition by antibodies. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19784–19789. [CrossRef] [PubMed]
103. Arocas, V.; Castro, H.C.; Zingali, R.B.; Guillin, M.C.; Jandrot-Perrus, M.; Bon, C.; Wisner, A. Molecular cloning and expression of bothrojaracin, a potent thrombin inhibitor from snake venom. *Eur. J. Biochem.* **1997**, *248*, 550–557. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.