PREDICTION OF ANTITHROMBOTIC PEPTIDES FROM VARIOUS SOURCES USING MACHINE LEARNING MODELS

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

[1]Department of Chemical and Materials Engineering, San José State University, San José, CA 95192, USA
[2]Department of Computer Science, Santa Clara University, Santa Clara, CA 95053, USA

SJSU SAN JOSÉ STATE UNIVERSITY

Background

- Thrombosis is a major cause of morbidity and mortality
- Thrombin is the key enzyme in mediating clot formation by converting plasma fibrinogen to crosslinked polymeric fibrin network
- Direct thrombin inhibitors (DTI) have higher capacity for the direct inhibition of fibrin-bound thrombin
- Limitations of DTI are associated with bleeding or thrombotic risks in certain situations
- Therefore, discovery of new antithrombotic agents are necessary

Objective

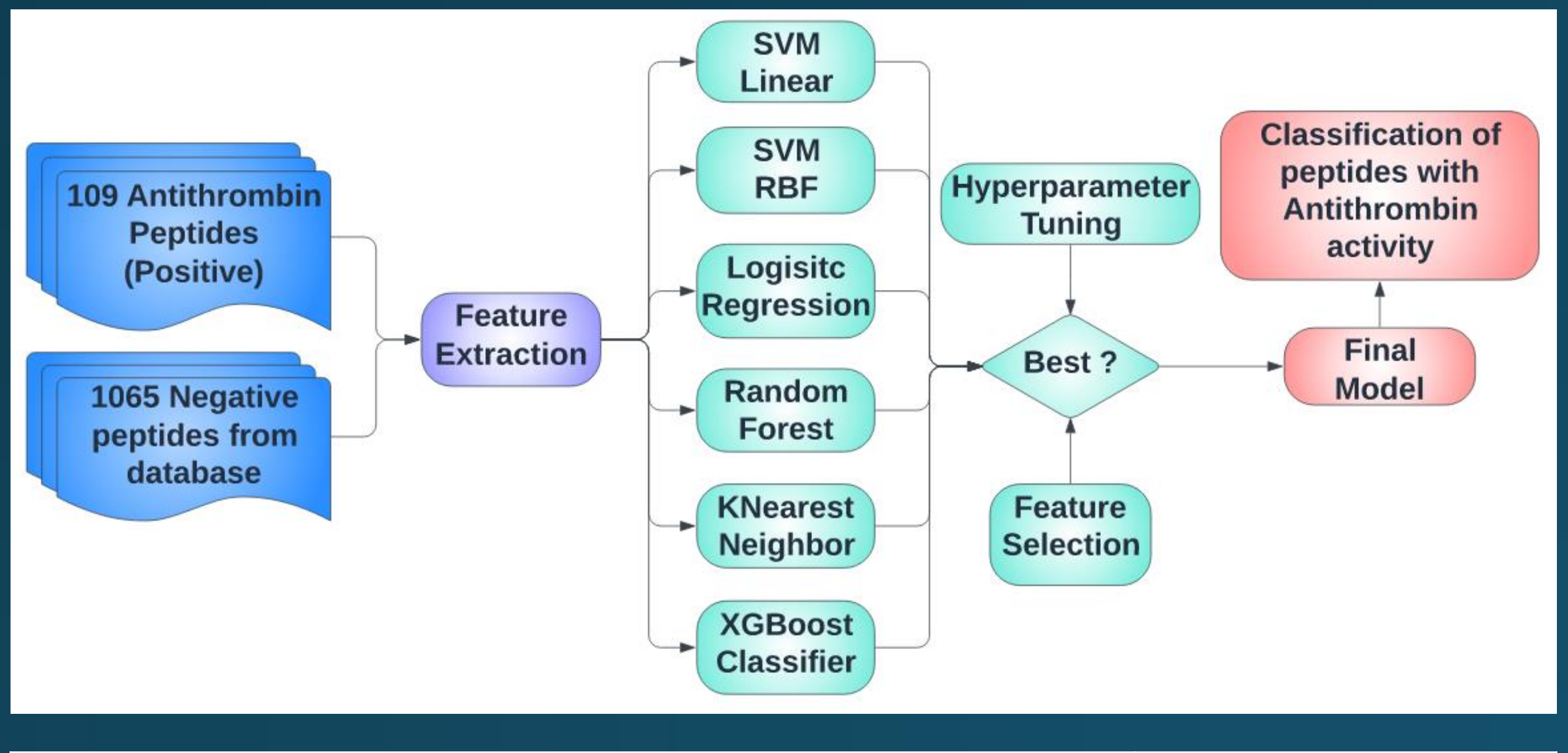
- To develop a two-staged machine learning (ML) model to predict new antithrombotic peptides
- ML models can rapidly screen a vast chemical and biological space, and enable accurate and faster *in silico* prediction of new molecules
- Classification models identify peptides with thrombin inhibitory activity
- Regression models rank peptides based on their effective thrombin inhibitory potential

Prediction of peptides with antithrombin activity

Data Preparation

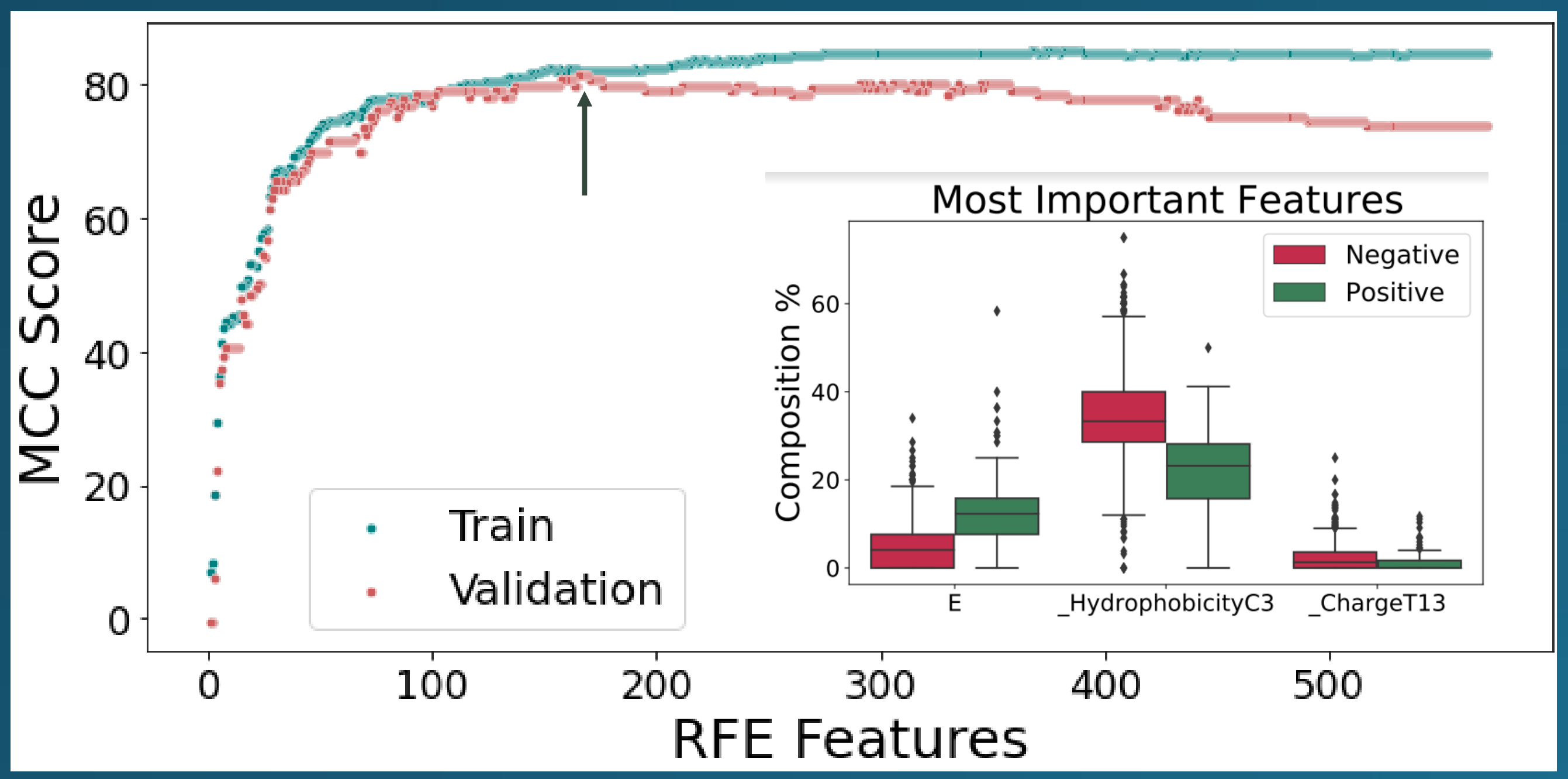
- Positive antithrombotic peptide dataset curated from peer reviewed publications
- Negative peptide dataset collected from NCBI and Uniprot protein database.

Classification Pipeline

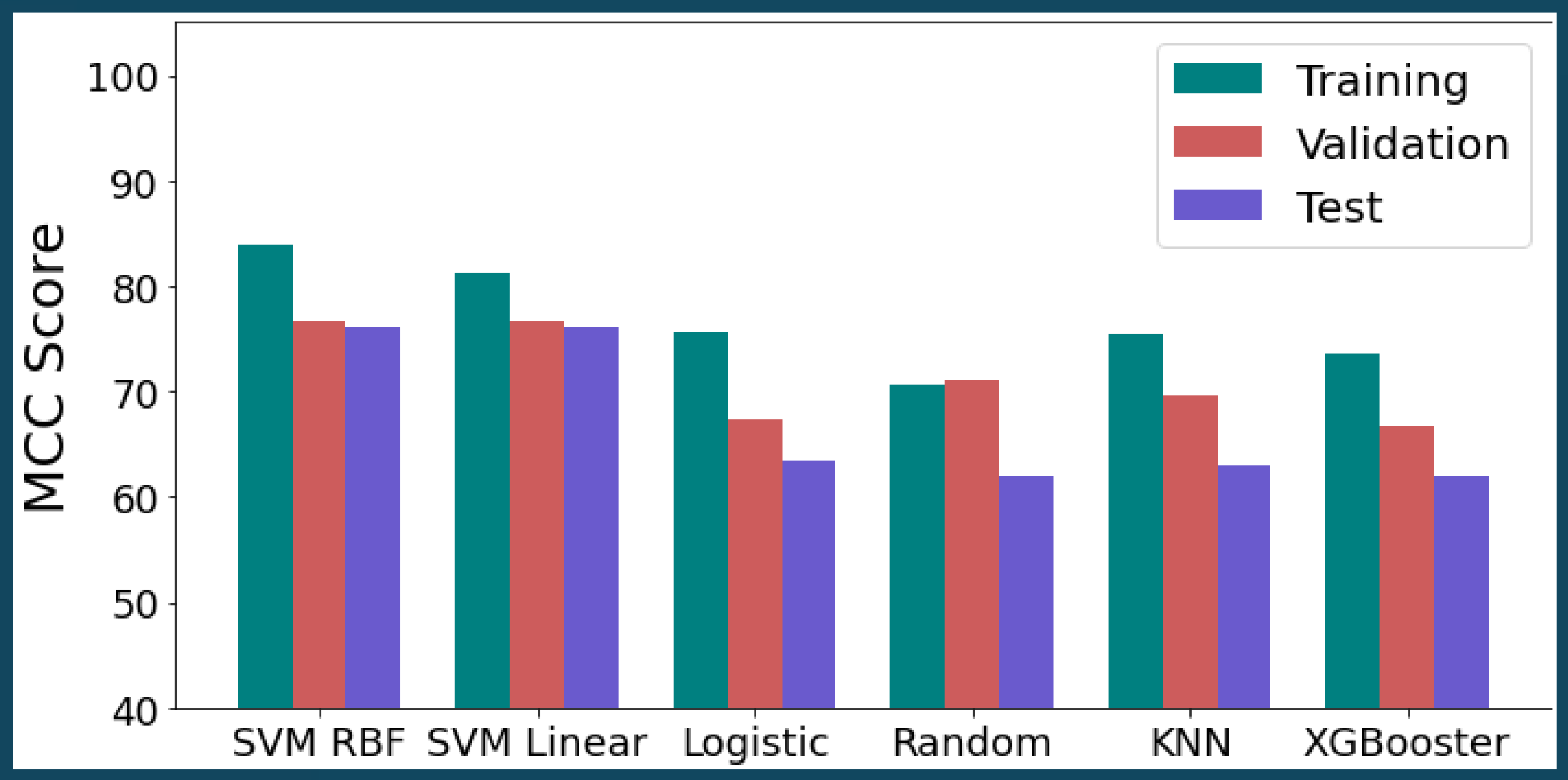


Extracted Features	No. of Features
Global Physico-chemical properties (GPC)	5
Amino Acid Composition (AAC)	20
Dipeptide Composition (DPC)	400
Composition Transition and Distribution (CTD)	147

165 Optimal features selected using Recursive Feature Elimination (RFE)

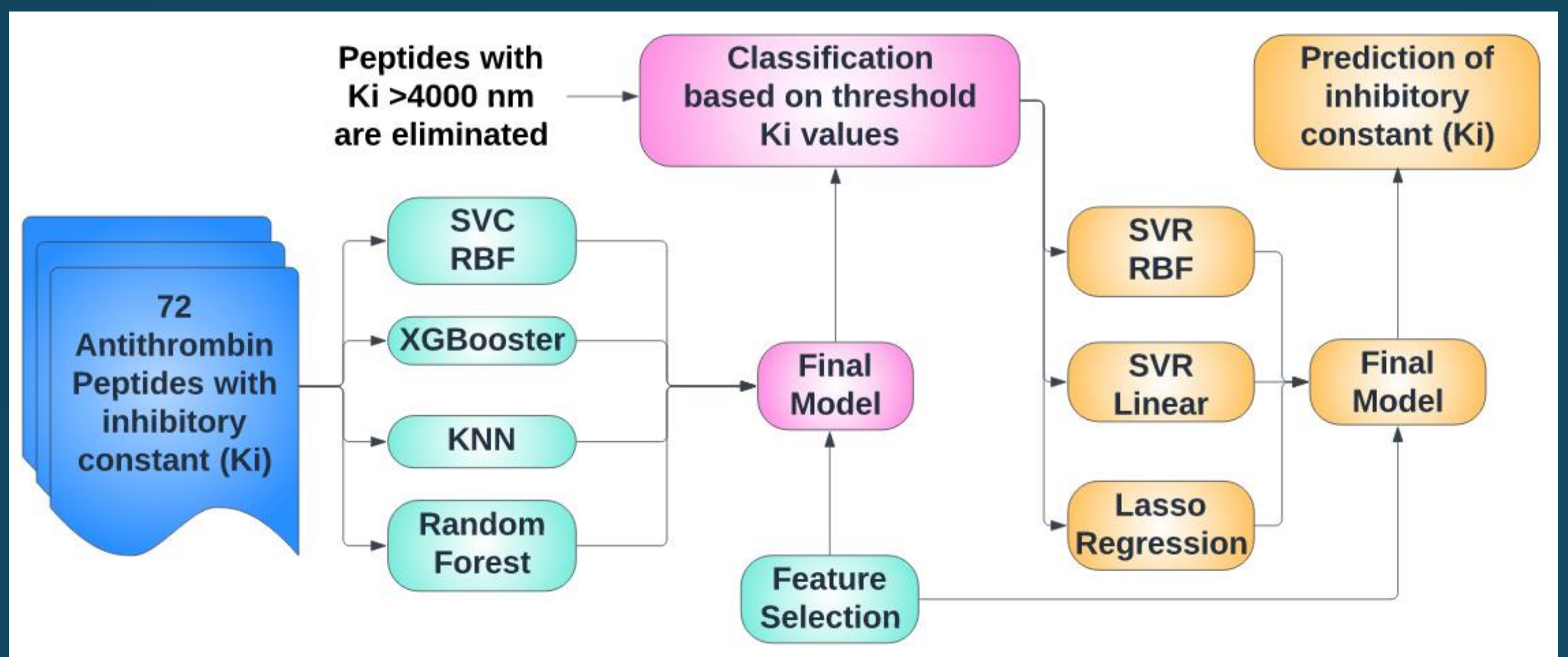


SVM RBF and SVM Linear as best performing models

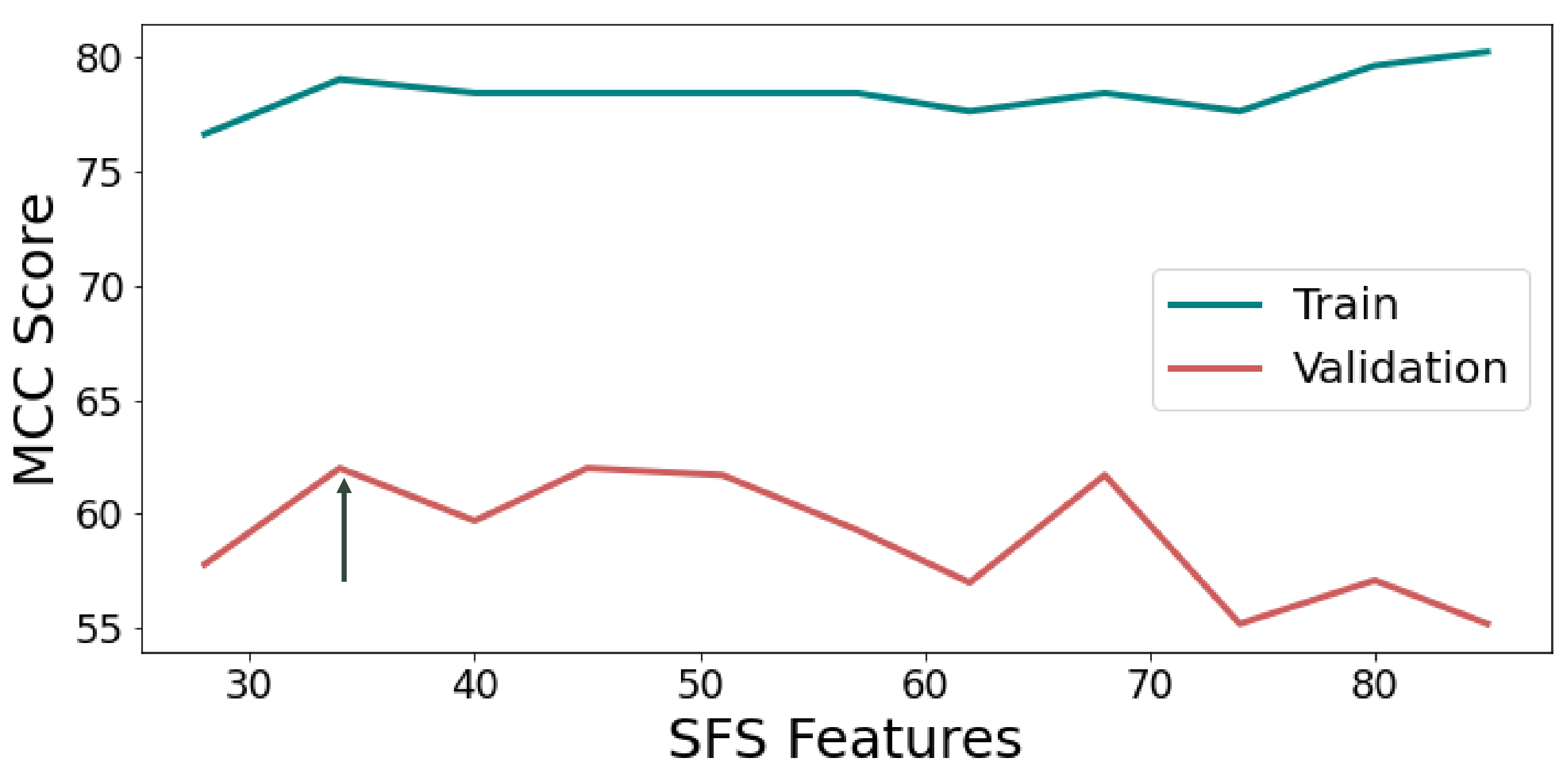


Prediction of antithrombotic efficacy

Regression Pipeline


Peptides with Ki >4000 nm are eliminated
Classification based on threshold Ki values
Prediction of inhibitory constant (Ki)
72 Antithrombin Peptides with inhibitory constant (Ki)
SVC RBF
XGBooster
KNN
Random Forest
Final Model
Feature Selection
SVR RBF
SVR Linear
Lasso Regression
Final Model

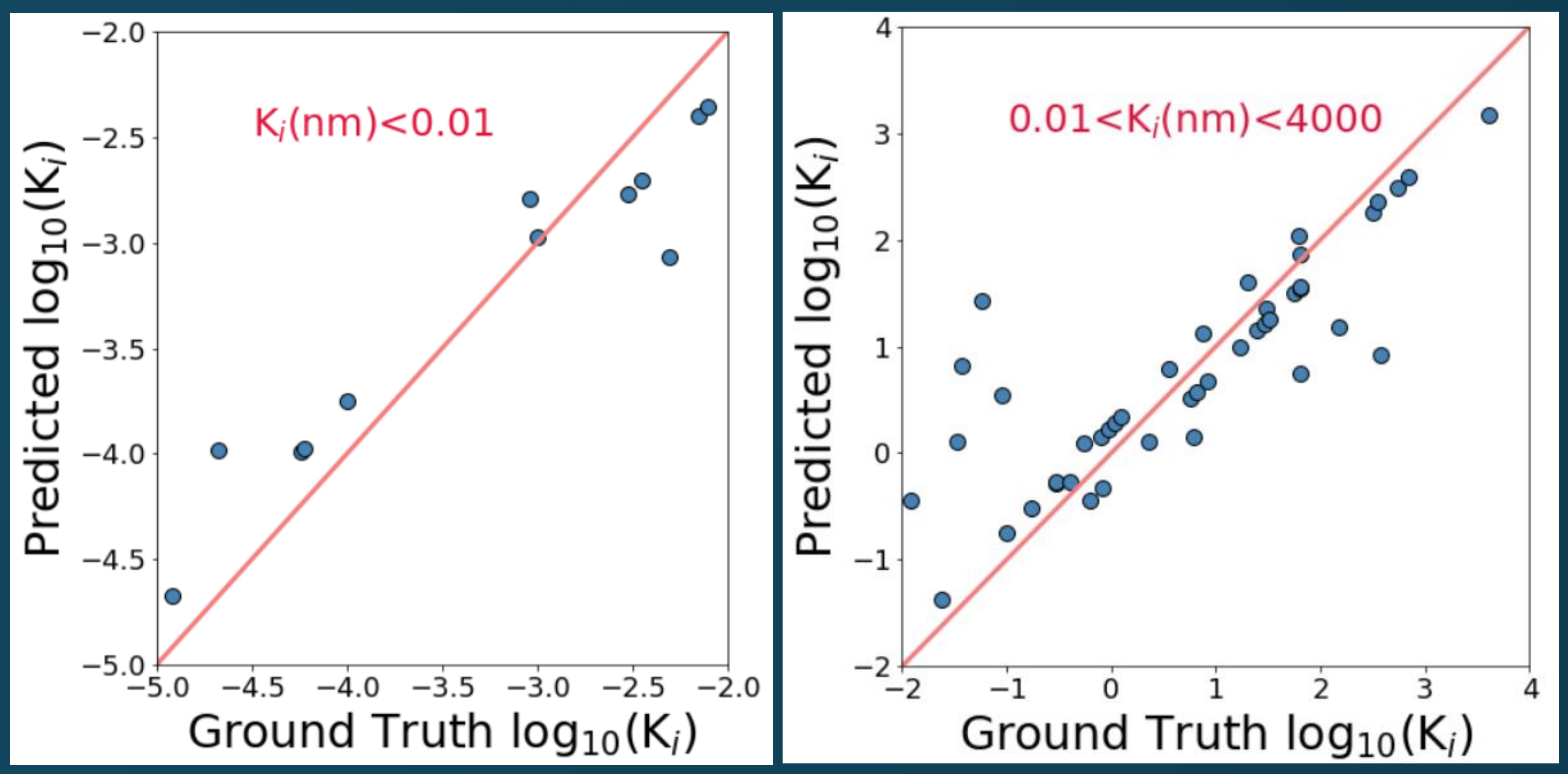
34 Optimal features selected using Sequential Forward Selection (SFS)



SVM RBF and SVR RBF are best performing models

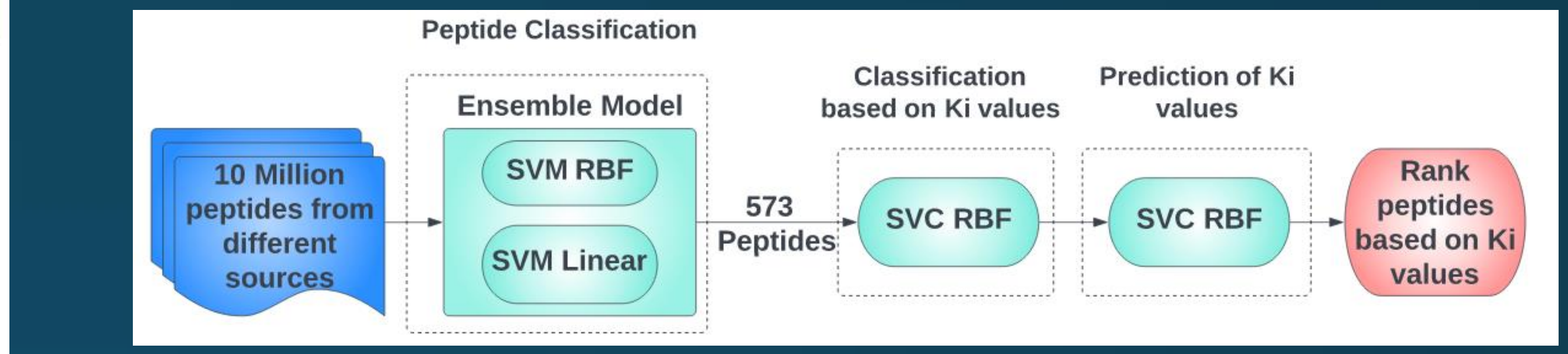
Classification Model	Regression Model	Validation RMSE	Validation MCC
SVM RBF	SVR RBF	1.64	0.62
SVM RBF	SVR Linear	1.64	0.62
SVM RBF	Lasso Regression	1.73	0.62
XGBoost Classifier	SVR RBF	1.82	0.56
XGBoost Classifier	SVR Linear	1.94	0.56
XGBoost Classifier	Lasso Regression	1.91	0.56
Random Forest	SVR RBF	1.81	0.44
Random Forest	SVR Linear	1.92	0.50
Random Forest	Lasso Regression	1.80	0.45

Performance of Regression Model

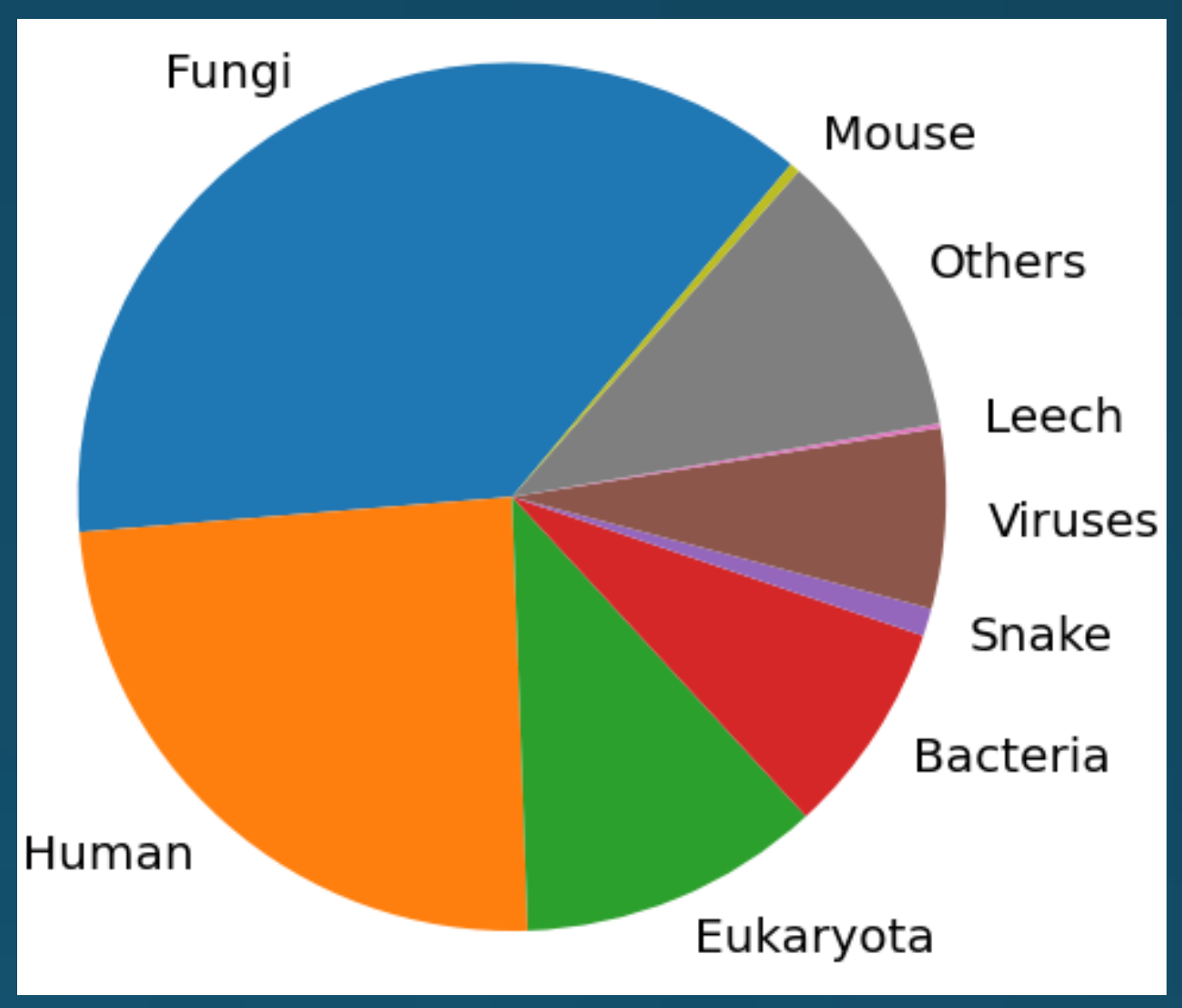


Identification of new peptides with antithrombin activity

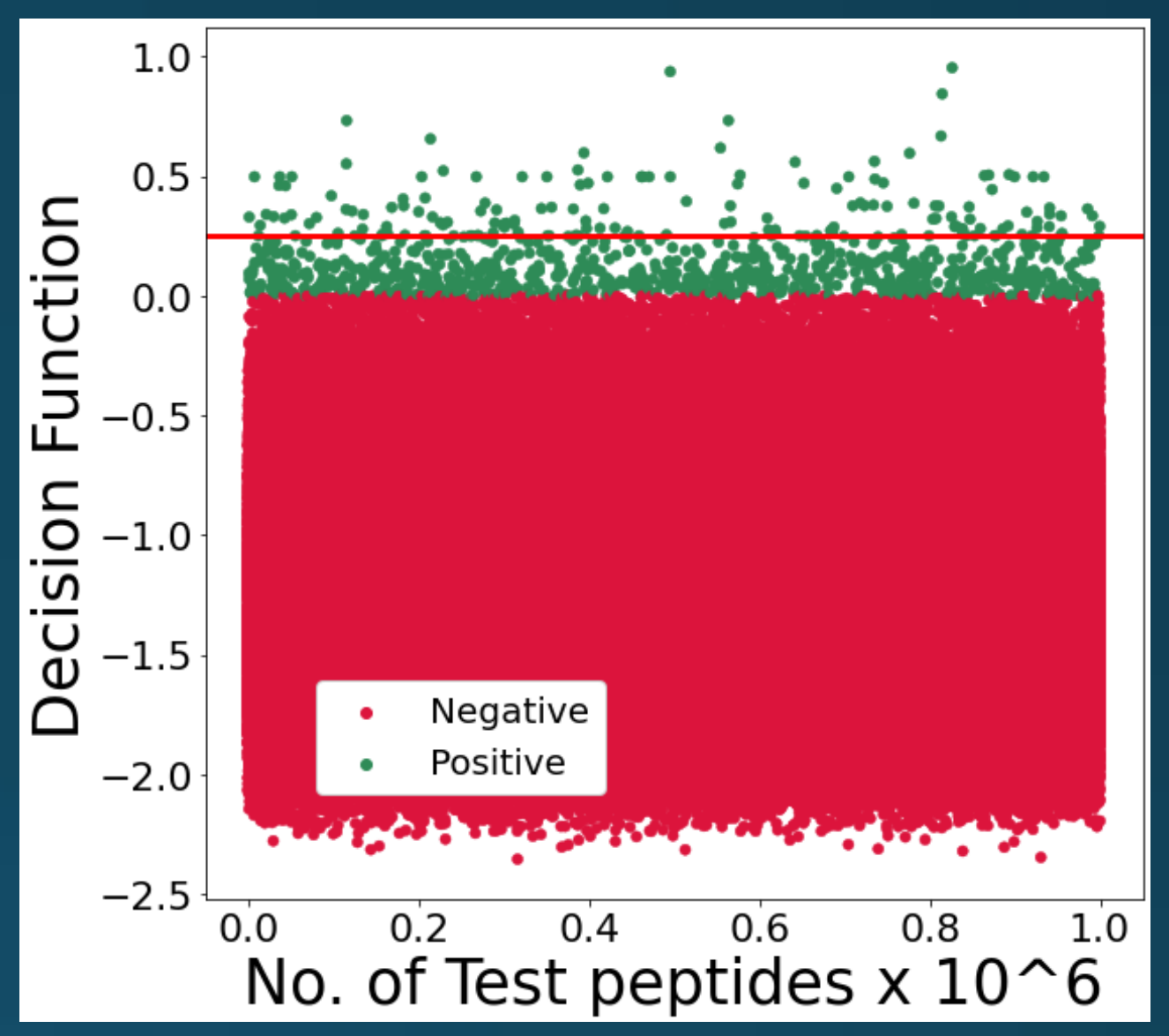
Pipeline to predict new antithrombin peptides


Peptide Classification
Ensemble Model
10 Million peptides from different sources
SVM RBF
SVM Linear
573 Peptides
Classification based on Ki values
SVC RBF
Prediction of Ki values
SVC RBF
Rank peptides based on Ki values

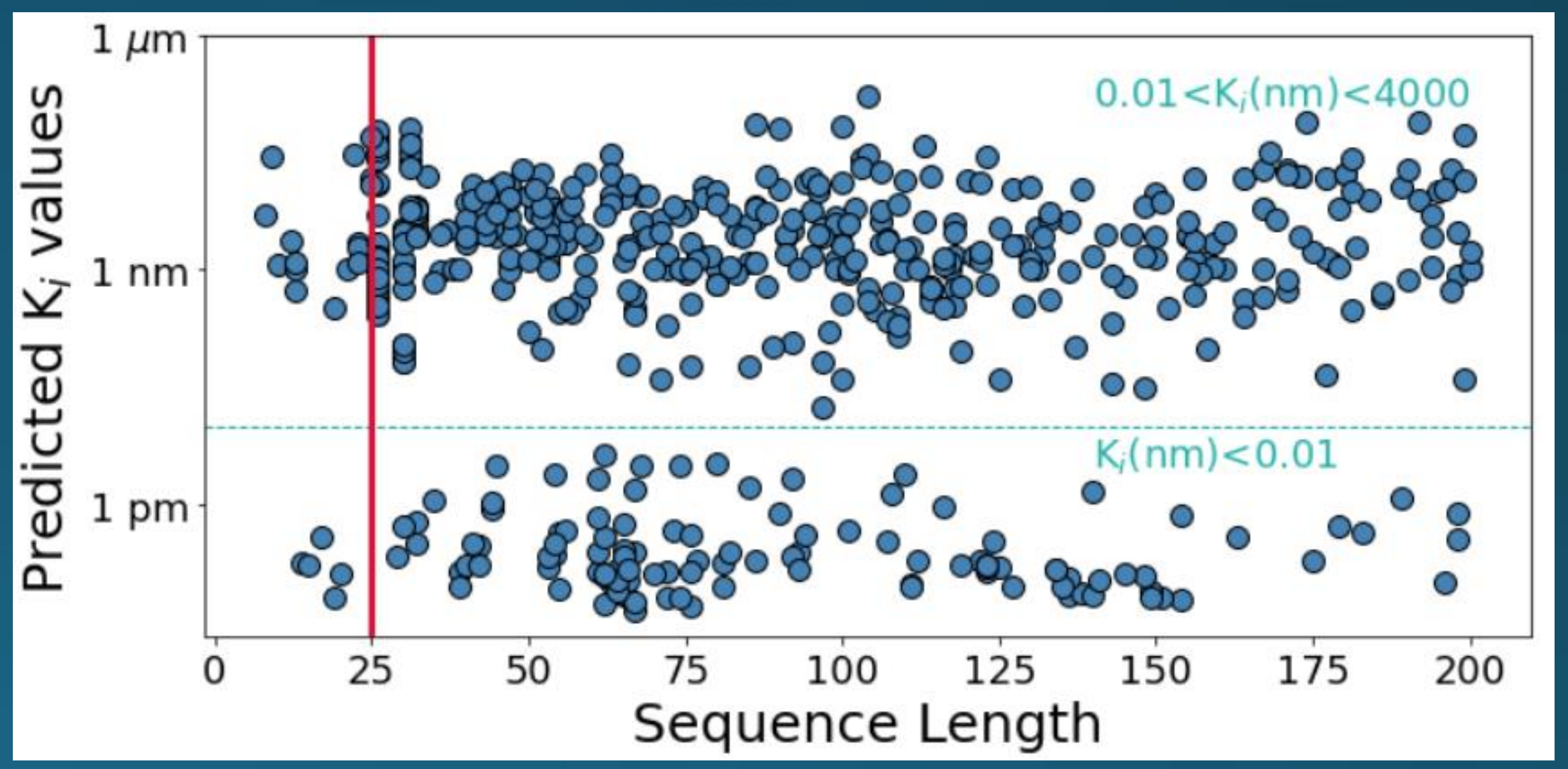
Source Distribution of the test peptides



Prediction of test peptides using decision function



18 new peptides are selected with seq length < 25



Summary

- SVC with RBF model and SVC with Linear model performed best based on 10-fold CV MCC of 72.08% and 73.53% respectively.
- Ensemble model of SVC Linear and SVC RBF is selected as the final classification model.
- For Regression model, SVR with Linear kernel gave best results with RMSE score of 1.75.
- Out of 10 million peptides, 581 peptides are distilled from the classification model.
- Best 18 peptides are chosen based on Ki values and Sequence length
- These peptides are being further analyzed for experimentation